Advanced Triple-Regression Algorithm for Customized Long-Term Forecasting

Cite this research:
Xu,X.,Bai,Z.,Jia,J.,Liang,Y.(2021) A Multi-Variate Triple-Regression Forecasting Algorithm for Long-Term Customized Allergy Season Prediction. SSRAML SageScience, 4(1), 59–65.

Xiaoyu Wu,
Boston University, Boston MA USA, Email: xywu@bu.edu
Zeyu Bai,
University of California Los Angeles, Los Angeles CA USA, Email: zeyubai21@engineering.ucla.edu
Jianguo Jia,
Deakin University, Victoria Australia, Email: j.jianguo@gmail.com
Youzhi Liang
Massachusetts Institute of Technology, Cambridge MA USA, Email: youzhil@mit.edu

Article history:
Received:
November/12/2020
Accepted:
March/08/2021

Abstract

This study introduces an innovative multi-variate triple-regression algorithm designed to forecast personalized airborne-pollen allergy seasons over extended periods. Our approach begins with preprocessing, where we integrate historical pollen data with inferential signals from various covariates, including meteorological information. The algorithm comprises three sequential regression stages: the first stage employs a regression model to predict the allergy season's start and end dates using a feature matrix derived from 12 time series covariates with a rolling window technique. The second stage predicts the uncertainty of these dates based on the feature matrix and Stage 1 outcomes. Finally, the third stage uses a weighted linear regression model, leveraging results from the previous stages, significantly enhancing forecasting accuracy and reducing uncertainty. This algorithm allows for individual customization based on varying pollen-triggered allergy sensitivities. Our backtesting results show a mean absolute error of 4.7 days, underscoring the potential of this method for both general and long-term allergy season predictions.

INTRODUCTION

Airborne-pollen allergy is prevalent, affecting up to 40% of the total population worldwide [1,2]. The long-term customized forecasting of pollen allergy provides individuals with guidance for travel planning and medication planning [3]–[5]. For pharmaceutical companies, the demand of medication for pollen allergy treatment, in addition to the sales and operations planning for supply chain management, further necessities the forecasting of the start/end date of the allergy season [6]–[8].

In Fig. 1, the concentration of pollen across the years from 2004 to 2008 is shown. We observe a significant change of the start date and end date for varying years with no explicit trend. For example, the allergy seasons in 2005 and in 2007 exhibit almost no overlapping; the length of the allergy season in 2004 roughly doubles that in any other year.

For a more rigorous definition of the allergy season tailored for each patient, we introduce the concentration threshold δ_C which is the minimum customized concentration requirement of pollen for a typical day in the allergy season, and the number threshold δ_N which is the minimum number of typical days within a week (7 consecutive days) in the allergy season. The start date of allergy season is defined as the day, during the following week of which the number of typical days (pollen concentration > δ_C) is at least δ_N. Both δ_C and δ_N can be customized according to different allergy sensitivity level of individuals. For example, if $\delta_C = 120$ and $\delta_N = 4$, the standard deviation of the start date, the end date and the length of allergy season are 19.9 days, 41.4 days and 47.2 days, respectively, from the year 2003 to year 2019.

Univariate forecasting methodologies, such as Exponential Smoothing, ARIMA and T-BATS, can achieve expected performance on time series data with a strong signal of level, trend and seasonality [9]–[11]. In addition to the classic univariate models, the use of network and neural network architectures, such as convolutional neural network (CNN) and recurrent neural network (RNN) furthers the scope of application [12,13]. Ordinary/Partial differential equations used widely in the field of physics and robotics shed light on the machine learning models, further driving the univariate forecasting investigation recently [14]–[18].

However, univariate forecasting methodologies exhibit underperformance in the context of cyclic intermittency, in particular for predicting the start date and the end date of the allergy season [19]. In the scenario of airborne-pollen allergy season prediction, the concentration of pollen, primarily produced by plants, is closely dependent on local environmental conditions like the weather and geography, necessitating the integration of the meteorological information into the forecasting methodology [20,21]. Univariate models, such as Holt-Winters exponential smoothing, ARIMA and Facebook Prophet model, are unable to integrate other time-varying covariates, in particular the weather information such as the precipitation, temperature and wind [22,23].

In this paper, we propose a multi-variate triple-regression algorithm to predict the airborne-pollen allergy season in the long term, i.e. the start date and end date of the season. The triggering concentration of the pollen – the definition of the allergy season can be customized as aforementioned. The proposed algorithm leverages the inferential signal from other covariates to make long-term accurate predictions, and uses a novel three-stage modelling approach to improve forecasting performance. In particular, we take into consideration the other 11 covariates including the history of temperature, wind and precipitation in addition to the historical data of pollen concentrations. The prediction results from early stage(s) are used in later regressions to further improve the accuracy and reduce the uncertainty of prediction.

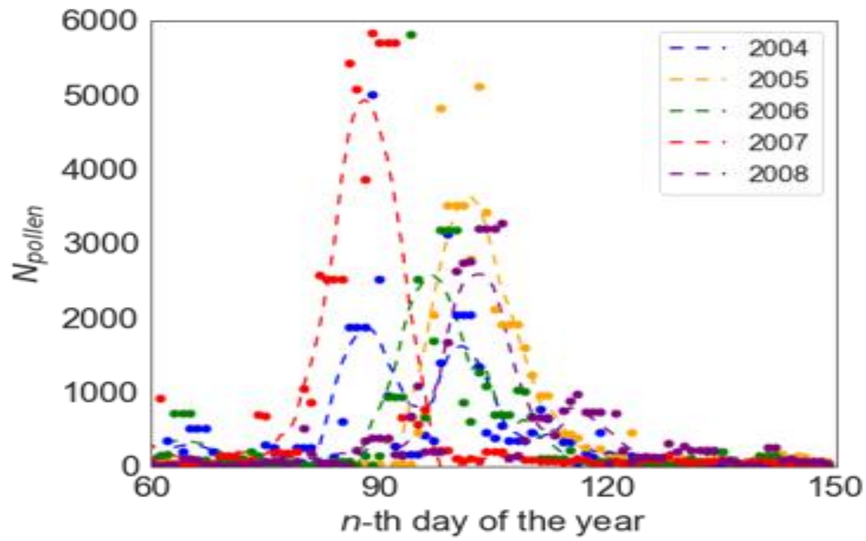

Fig. 1: The concentration of pollen as a function of the n-th day of the year within a 120-day range, i.e. the count of pollen in a cubic meter of air, measured per day for the year from 2004 to 2008. The dashed lines indicate smoothed trend using the Savitzky–Golay filter, merely serving for the visualization.

ALGORITHM

The proposed algorithm encompasses a three-stage regression for the start/end date of the allergy season prediction, together with a pre-processing for the feature extraction. The data pipeline and the regression algorithm are outlined in Fig. 2. In the pre-processing, a total of N_f (=30) time series selected features are extracted by applying a 14-day rolling window to each of the n (=12) original time series vectors x_i, including the pollen concentration history, temperature history, wind history, participation history, etc.

The feature matrix corresponding to x_i is denoted as $\mathbb{X}^{(i)}_{N\times N_f}$, and the ensemble feature matrix is $\mathbb{X}^{(1:n)}_{N\times N_f\times n}$. The ensemble feature matrix is then fed into the three-stage regression. In Stage 1, a Gradient Boosting model to predict the start/end date is trained on a training set S1 which is based on the feature matrix $\mathbb{X}^{(1:n)}_{N\times N_f\times n}$. The ground truth for the start/end date of allergy season is determined following the definition after we select appropriate pollen concentration threshold δ_C and number threshold δ_N. The vector of prediction is given by $\hat{y} = f_y(\mathbb{X}^{(1:n)}_{N\times N_f\times n})|\mathbb{S}_1$.

In Stage 2, we select another training set S2 based on the feature matrix $\mathbb{X}^{(1:n)}_{N\times N_f\times n}$ and the predicted start/end date $\hat{y}$ from Stage 1 to train another Gradient Boosting model to predict the uncertainty in $\hat{y}$. The vector of uncertainty is given by $\hat{u} = f_u(\hat{\boldsymbol{y}}, \mathbb{X}^{(1:n)}_{N\times N_f\times n}))|\mathbb{S}_2$. In Stage 3, we perform a weighted linear regression on $\hat{y}$ based on the linear constraint on start/end date predictions at consecutive days ahead of the allergy season. The weights are assigned according to the predicted uncertainty $\hat{u}$. Thus, the final predicted start/end date of the allergy season is obtained by $\hat{y}^* = f_{WL}(\hat{y})|W = \hat{u}$.

Although one can opt to terminate the algorithm at Stage 1 when the predictions are already made, we emphasize that the three-stage regression can guarantee its uncertainty to be smaller than that using only one-stage regression. The proof is given for a simplified problem as follows.

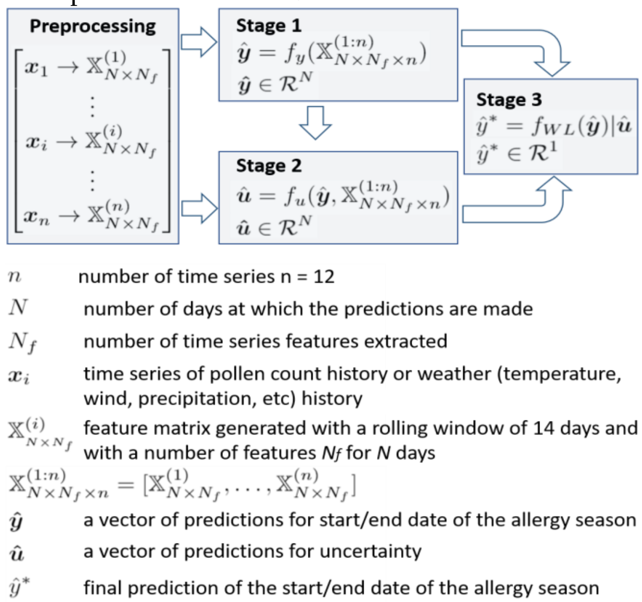

Fig. 2: Data pipeline of the proposed triple-regression algorithm with a nomenclature. The feature matrix from pre-processing is the input for Stage 1; the outcome from pre-processing and Stage 1 is the input for Stage 2; the regression results from Stage 1 and Stage 2 are used in Stage 3.

We may assume that each prediction in the first-stage regression $\hat{y}_i$ follows a Gaussian distribution:

$$\hat{y}_i \in \mathcal{N}(\mu_i, \sigma_i^2), \tag{1}$$

where the variance σ_i is assumed to be constant σ_0, and μ_i is theoretically constrained by a linear relationship given the predictions are made at consecutive days z_i. To set up the problem of a multi-linear regression in general, we may consider p independent variables $x_1,...,x_n$ and one dependent variable y. Suppose we have $n(n > p)$ observations,

$$\hat{y}_i = \beta_0 + \sum \beta_i z_{ij} + \epsilon_i,\ i = 1, ..., n, \tag{2}$$

where β_i are the coefficients of the i-th dependent variable x_i. Our goal is to minimize the sum of the weighted squared residuals (errors) $_i$. Thus, the cost function is:

$$\sum_{i=1}^{n} w_i \epsilon_i^2 = \sum_{i=1}^{n} w_i \left(\hat{y}_i - \beta_0 - \sum_{j=1}^{p} \beta_j z_{ij} \right)^2 \tag{3}$$

where the weights w_i is provided by the inverse of the uncertainty $\hat{u}^2_i|_{(2)}$ from the regression results in Stage 2, $w_i = 1/\sigma_i^2|_{(2)} = 1/\hat{u}^2_i|_{(2)}$. No analytical solution can be obtained for a set of random weights. Without loss of generality, we can focus on a simplified scenario with uniform weights, and the linear regression results can be expressed as:

$$\hat{\boldsymbol{Y}} = \mathbb{E}[\boldsymbol{Y}] + \frac{\mathbf{Cov}(\boldsymbol{Z}, \boldsymbol{Y})}{\mathbf{Var}(\boldsymbol{Z})}(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}]), \tag{4}$$

where E[Y] is the expectation of random variable Y, **Cov**(Z,Y) is the covariance of random variable Z and Y, and **Var**(Z) is the variance of random variable Z.

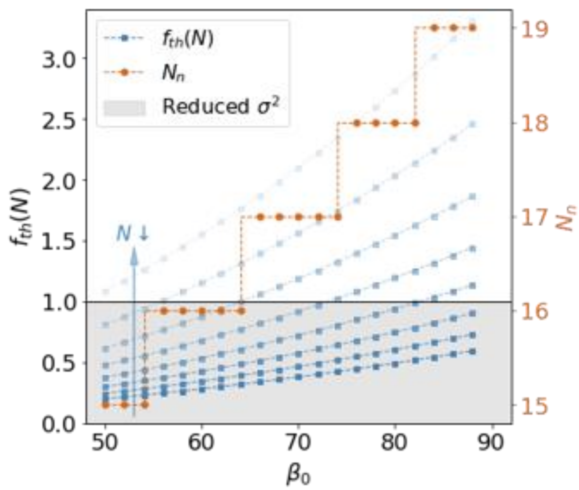

Fig. 3: Threshold function $f_{th}(N)$, and minimum number of days N_n to reduce uncertainty, as functions of the regression coefficient β_0 with varying number of days N used for prediction. The shaded area indicates the regime where the variance using a three-stage regression is reduced compared with the one-stage regression counterpart. Note that β_1 is assumed to be 1 under the condition of ideal scenario.

In the context of predicting the start/end date of the allergy season, the only independent variable in the weighted linear regression is the date z_i at which the prediction is made. Thus, we only have two non-zero coefficients, β_0 and β_1, to be learned from the regression. The variance of the predicted coefficient can be approximated by [24]:

$$\sigma^2(\hat{\beta}_0) = \sigma'^2 \left(\frac{1}{n} + \frac{\overline{z}^2}{\sum_0^N (z_i - \overline{z})^2} \right)$$

$$\sigma^2(\hat{\beta}_1) = \sigma'^2 \frac{1}{\sum_0^N (z_i - \overline{z})^2},$$

,

It can be shown that σ^{02} is related to the uncertainty from the Gradient Boosting regression in Stage 1 through the following formula:

$$\sigma'^2 = \frac{\sigma_0^2}{N}. \tag{6}$$

The final prediction of the start/date date $\hat{y}^*$ is given by the z-intercept of the line from the weighted linear regression:

$$\hat{y}^* = -\frac{\beta_0}{\beta_1} \tag{7}$$

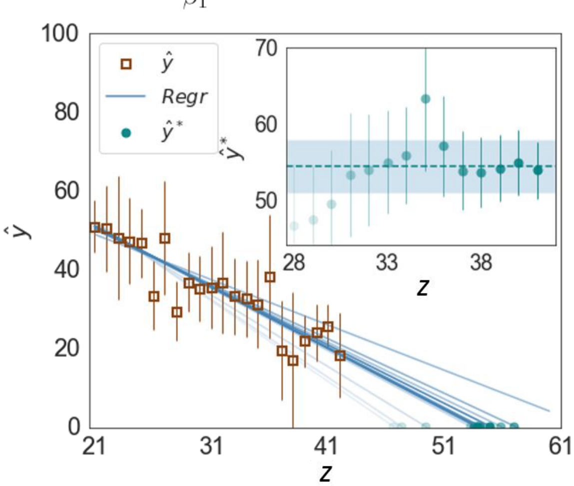

Fig. 4: Start date $\hat{y}$ predicted in Stage 1, as a function of the z-th day of the year 2005, where the error bar indicates the standard deviation predicted in Stage 2. Straight lines are the weighted linear regression results with

different number of predictions considered. (Inset) Final prediction of the start date $\hat{y}^*$ predicted in Stage 3, as a function of the z-th day of the year. Shaded area indicates the band of standard deviation.

Thus, the standard deviation of $\hat{y}^*$ is approximated by

$$\sigma(\hat{y}^*) = \hat{y}^* \sqrt{\frac{\sigma'^2(\hat{\beta}_0)}{\hat{\beta}_0^2} + \frac{\sigma'^2(\hat{\beta}_1)}{\hat{\beta}_1^2}} = -\frac{\hat{\beta}_0}{\hat{\beta}_1} \sqrt{\frac{\sigma'^2}{\hat{\beta}_0^2}\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i - \overline{x})^2}\right) + \frac{\sigma'^2}{\hat{\beta}_1^2}\frac{1}{\sum(x_i - \overline{x})^2}} \quad (8)$$

Utilizing the third-stage regression, we aim for a reduced variance, i.e. $\sigma^2(\hat{y}^*) < \sigma_0^2$. In other words, the uncertainty should be reduced from the one-stage regression result. It is obvious that a minimum number of days N_n where the predictions are made in Stage 1 is required to guarantee the reduced uncertainty. We can calculate N_n by first defining the threshold function:

$$f_{th}(N_n) = \frac{1}{N_n \hat{\beta}_1^2}\left(\frac{1}{N_n} + \frac{\overline{x}^2}{\sum_0^{N_n}(x_i - \overline{x})^2} + \frac{\hat{\beta}_0^2/\hat{\beta}_1^2}{\sum_0^{N_n}(x_i - \overline{x})^2}\right), \quad (9)$$

then setting $f_{th}(N_n)$ to 1, indicating that the uncertainty does not change after the weighted linear regression.

RESULTS AND DISCUSSION

To guarantee a reduced uncertainty in the three-stage regression compared with the one-stage regression, we obtain the minimum number of days, N_n, used for making predictions. In Fig. 3, we plot the threshold function $f_{th}(N)$ as a function of the coefficient β_0 with varying N. The shaded area in Fig. 3 indicates the regime where the three-stage regression has a reduced uncertainty.

Therefore, the corresponding minimum number of days N_n for a specific β_0 value can be obtained from the threshold function which satisfies $f_{th}(N) < 1$ and has the smallest N among all threshold functions. As β_0 increases, the value of threshold function increases, reflecting an increased minimum number of days N_n is required.

We apply the data pipeline and three-stage regression algorithm to the dataset described in Section II to predict the start date of the allergy season. The ground truth of the start date of the allergy season is set by the thresholds, $\delta_C = 120$ and $\delta_N = 4$, i.e., for seven consecutive days after the start day, the number of days when the pollen concentrations is greater than δ_C is at least δ_N.

In Fig. 4, we show the mean $\hat{y}$ predicted in Stage 1, and the corresponding error bar, which is the standard deviation $\sigma(\hat{y})$ predicted in Stage 2 as functions of the z-th day of the year 2005. The models to predict $\hat{y}$ and $\sigma(\hat{y})$ are trained using the dataset from year 2003 and 2004 respectively. The blue lines in decreasing transparency represent weighed linear regression results with increasing number of predictions $\hat{y}$ used, where the inverse of the uncertainty $\sigma(\hat{y})$ serves as the weights. The green circles locates at the z-intercept represents the final prediction made in Stage 3. It is manifested in the Fig. 4 (Inset) that the prediction converges to Day 54 as the number of days accounted increases, while the actual start date is Day 51 for year 2005. We also performed backtesting for year 2006, 2007 and 2008, and a mean absolute error of 4.7 days was achieved using the triple-regression algorithm.

The intercept of y-axis, coefficient β_0, indicates the approximate prediction of the start/end date.

CONCLUSIONS

The airborne-pollen allergy season exhibits significant variations in terms of the start/end date and the length of the allergy season. Univariate models fail to extract its seasonality or trend and fail to integrate other covariates such as the temperature and precipitation.

In our proposed triple-regression algorithm, (a) the pollen allergy season can be customized based on each individual's allergy sensitivity level, (b) the predictions are obtained based on the historical data of pollen concentration together with other 11 covariates, (c) most importantly, the uncertainty of the prediction in Stage 3 can be reduced given that the minimum number of predictions obtained from Stage 1 is satisfied. The final prediction in Stage 3 converges to a mean value with the increasing number of predictions obtained from Stage 1. The weighted linear regression further improve the accuracy by integrating the uncertainty predicted in Stage 2. In our backtesting, a mean absolute error of 4.7 days was achieved using the triple-regression forecasting algorithm. We conclude that this algorithm could be useful in both generic and long-term forecasting problems.

REFERENCES

G. D’Amato, L. Cecchi, S. Bonini, C. Nunes, I. Annesi-Maesano,

H. Behrendt, G. Liccardi, T. Popov, and P. Van Cauwenberge, “Allergenic pollen and pollen allergy in europe,” *Allergy*, vol. 62, no. 9, pp. 976–990, 2007.

J. Sanchez-Mesa,´ C. Galan,´ J. Mart´ınez-Heras, and C. Hervas-Mart´ ´ınez, “The use of a neural network to forecast daily grass pollen concentration in a mediterranean region: the southern part of the iberian peninsula,” *Clinical & Experimental Allergy*, vol. 32, no. 11, pp. 1606–1612, 2002.

M. Prank, D. S. Chapman, J. M. Bullock, J. Belmonte, U. Berger, A. Dahl, S. Jager, I. Kovtunenko, D. Magyar, S. Niemel¨ a,¨ *et al.*, “An operational model for forecasting ragweed pollen release and dispersion in europe,” *Agricultural and forest meteorology*, vol. 182, pp. 43–53, 2013.

C. Arizmendi, J. Sanchez, N. Ramos, and G. Ramos, “Time series predictions with neural nets: application to airborne pollen forecasting,” *International journal of biometeorology*, vol. 37, no. 3, pp. 139–144, 1993.

A. Ranzi, P. Lauriola, V. Marletto, and F. Zinoni, “Forecasting airborne pollen concentrations: Development of local models,” *Aerobiologia*, vol. 19, no. 1, pp. 39–45, 2003.

N. Singh, S. J. Olasky, K. S. Cluff, and W. F. Welch Jr, “Supply chain demand forecasting and planning,” July 18 2006. US Patent 7,080,026.

S. Jaipuria and S. Mahapatra, “An improved demand forecasting method to reduce bullwhip effect in supply chains,” *Expert Systems with Applications*, vol. 41, no. 5, pp. 2395–2408, 2014.

Y. Barlas and B. Gunduz, “Demand forecasting and sharing strategies to reduce fluctuations and the bullwhip effect in supply chains,” *Journal of the Operational Research Society*, vol. 62, no. 3, pp. 458–473, 2011.

E. S. Gardner Jr, “Exponential smoothing: The state of the art,” *Journal of forecasting*, vol. 4, no. 1, pp. 1–28, 1985.

S. C. Hillmer and G. C. Tiao, “An arima-model-based approach to seasonal adjustment,” *Journal of the American Statistical Association*, vol. 77, no. 377, pp. 63–70, 1982.

H. Hassani, E. S. Silva, N. Antonakakis, G. Filis, and R. Gupta, “Forecasting accuracy evaluation of tourist arrivals,” *Annals of Tourism Research*, vol. 63, pp. 112–127, 2017.

Z. Bai, R. Yang, and Y. Liang, “Mental task classification using electroencephalogram signal,” *arXiv preprint arXiv:1910.03023*, 2019.

A. Chavez, D. Koutentakis, Y. Liang, S. Tripathy, and J. Yun, “Identify statistical similarities and differences between the deadliest cancer types through gene expression,” *arXiv preprint arXiv:1903.07847*, 2019.

Y. Chen, B. Yang, Q. Meng, Y. Zhao, and A. Abraham, “Time-series forecasting using a system of ordinary differential equations,” *Information Sciences*, vol. 181, no. 1, pp. 106–114, 2011.

B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, “N-beats: Neural basis expansion analysis for interpretable time series forecasting,” *arXiv preprint arXiv:1905.10437*, 2019.

Y. Liang, A. E. Hosoi, M. F. Demers, K. D. Iagnemma, J. R. Alvarado, R. A. Zane, and M. Evzelman, “Solid state pump using electro-rheological fluid,” June 4 2019. US Patent 10,309,386.

R. Fish, Y. Liang, K. Saleeby, J. Spirnak, M. Sun, and X. Zhang, “Dynamic characterization of arrows through stochastic perturbation,” *arXiv preprint arXiv:1909.08186*, 2019.

Y. Liang, J. R. Alvarado, K. D. Iagnemma, and A. Hosoi, “Dynamic sealing using magnetorheological fluids,” *Physical Review Applied*, vol. 10, no. 6, p. 064049, 2018.

J. H. Stock and M. W. Watson, “A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series,” tech. rep., National Bureau of Economic Research, 1998.

K. K. R. Samal, K. S. Babu, S. K. Das, and A. Acharaya, “Time series based air pollution forecasting using sarima and prophet model,” in *Proceedings of the 2019 International Conference on Information Technology and Computer Communications*, pp. 80–85, 2019.

R. Meese and J. Geweke, “A comparison of autoregressive univariate forecasting procedures for macroeconomic time series,” *Journal of Business & Economic Statistics*, vol. 2, no. 3, pp. 191–200, 1984.

R. Fildes and E. J. Lusk, “The choice of a forecasting model,” *Omega*, vol. 12, no. 5, pp. 427–435, 1984.

A. M. De Livera, R. J. Hyndman, and R. D. Snyder, “Forecasting time series with complex seasonal patterns using exponential smoothing,” *Journal of the American statistical association*, vol. 106, no. 496, pp. 1513–1527, 2011.

G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.