Article:
Zhao, X, Barber, S, Taylor, CC et al. (2022) Spatio-temporal forecasting using wavelet transform-based decision trees with application to air quality and covid-19 forecasting. Journal of Applied Statistics, 50 (9). pp. 2036-2054. ISSN: 0266-4763

https://doi.org/10.1080/02664763.2022.2064976

Spatio-temporal forecasting using wavelet transform-based decision trees with application to air quality and covid-19 forecasting

Xin Zhao[1,2], Stuart Barber[2], Charles C Taylor[2], Xiaokai Nie[3], Wenqian Shen[1]

1. School of Mathematics, Southeast University, Nanjing, 211189, P.R. China.
2. School of Mathematics, University of Leeds, Leeds LS2 9JT, U.K.
3. School of Automation, Southeast University, Nanjing, 210096, P.R. China.

Abstract

We develop a new method that combines a decision tree with a wavelet transform to forecast time-series data with spatial spillover effects. The method can not only improve prediction but also give good interpretability of the time series mechanism. As a feature exploration method, the wavelet transform represents information at different resolution levels, which may improve the performance of decision trees. The method is applied to simulated data, air pollution and COVID time series data sets. In the simulation, Haar, LA8, D4 and D6 wavelets are compared, with the Haar wavelet having the best performance. In the air pollution application, by using wavelet-transform based decision trees, the temporal effect of air quality index including autoregressive and seasonal effects can be described as well as the spatial correlation effect. To describe the spillover spatial effect in contiguous regions, a spatial weight is constructed to improve the modeling performance. The results show that air quality index has autoregressive, seasonal and spatial spillover effects. The wavelet-transformed variables have a better forecasting performance and enhanced interpretability than the original variables. For the COVID time series of cumulative cases, spatial weighted variables are not selected which shows the lock-down policies are truly effective.

Keywords: CART, MODWT, COVID, air pollution, time series, spatial analysis

[1]Xin Zhao. Email: mmxinzhao@hotmail.com

1. Introduction

Due to heavy usage of fossil fuels such as oil and coal, urban industry has rapidly developed, but at the cost of increasingly serious environmental problems, especially air pollution. According to the '2018 World Air Quality Report' (AirVisual, 2018), nine out of ten people worldwide are now breathing unsafe polluted air. China ranks 12^{th} out of 73 countries in the world in terms of the estimated average PM2.5 concentration with a value of 41.2 $\mu g/m^3$. PM2.5 has become one of the main contibutors to air pollution in China, leading to unhealthy air quality or even hazardous air, especially smog.

An Air Quality Index (AQI) is a generalized comprehensive way to describe air quality, based on the level of six atmospheric pollutants: sulfur dioxide (SO_2), nitrogen dioxide (NO_2), suspended particulates (PM10, PM2.5), carbon monoxide (CO), and ozone (O_3) measured at monitoring stations throughout each city (Gurjar et al., 2008; Gupta et al., 2006). AQI has been in use since 2012 and is improved from the previous metric API (Air Pollution Index), which was based only on sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and suspended particulate (PM10). However, PM2.5, the main constituent of smog, is not included in API. Each of the six metrics in AQI is standardized because of the unit difference before calculating the AQI. The rules for standardization can be found from the United States Environmental Protection Agency (EPA, 2020). AQI can represent the most serious atmospheric pollutant in one region, such as PM2.5 in China. Daily or monthly AQI is averaged from hourly AQI.

AQI is divided into six ranges, representing six levels of air quality. A higher AQI means more serious pollution. Generally, an AQI under 100 is good or moderate, while an AQI higher than 100 means unhealthy or even hazardous conditions. For detail, please refer to the AQI Basics (AirNow, 2020).

If AQI could accurately be forecast ahead by, say, one month, then government, businesses, and people would have enough time to prepare some response measures in case of serious pollution. During days with low AQI, enterprises could also carry out more strict waste management. The method developed for forecasting AQI is also applied to COVID-19 data as a second application.

2. Literature review

As an index measuring the air quality, AQI has been forecast with deterministic, statistical, and machine learning methods (Ma et al., 2019). Deterministic methods (Sivakumar et al., 2007; Finardi et al., 2008; Hoshyaripour et al., 2016) require expensive computation and specific domain knowledge for parameter identification.

Such parameter determined models may not adjust to the quick change of air quality distribution. Due to the difficulty in justifying assumptions such as linearity and problems with multicollinearity, simple statistical methods like linear models become less appropriate, requiring more sophisticated modelling strategies. One of the popular machine learning methods is neural networks (Septiawan and Endah, 2018; Patni and Sharma, 2019), which have high prediction accuracy, along with its improved variants like convolution recurrent neural networks (Zhao and Zettsu, 2019) and long short-term memory neural network (Dua et al., 2019). Other machine learning methods include support vector machines (Leong et al., 2019), random forests (Rubal and Kumar, 2018; Kaminska, 2018; Li et al., 2019) and so on. However, in addition to requiring high prediction accuracy, it is preferred that methods also have good interpretability.

Air quality has some typical features that other time series may not have, which should be taken into consideration when forecasting. One is that air pollution has spatial spillover effects (Chen et al., 2019). Pollution from one region with bad air quality is apt to diffuse and migrate into neighbouring regions, influence them and lead to lower air quality in those regions. Li et al. (2018) takes account of the spatial influence when using neural networks for forecasting by proposing a novel spatio-temporal-aware sparse denoising autoencoding neural network architecture. Spillover effects have also been considered by other authors such as Guyu et al. (2019) and Wang and Song (2018). One typical method in economics to include spatial effects is to construct a spatial weight matrix to represent the mutual influence of geographically contiguous spaces (You and Lv, 2018), which will be used in our proposed method.

Another feature is that air pollution is likely to have pronounced autocorrelation since current air quality is influenced by its previous state. One novel model of accommodating this is the spatiotemporal convolutional short-term neural network (Wen et al., 2019; Song et al., 2019), which considers both spatial and temporal effects. Methods including autocorrelation effects like ARIMA are also considered by (Emetere, 2018). Such models can include long term autocorrelation effects, but at the cost of more parameters. To explore the autocorrelation of air pollution, we propose the wavelet transform method (Percival and Walden, 2000), which can decompose the original time series into different time series on multiple resolution levels. Long term effect of air pollution can be included in decomposed time series on low resolution levels while short term effects are included in high resolution levels, as we describe in Section 4.

Air quality also has seasonal patterns (Kim et al., 2017; Du et al., 2018). This is especially true in northern China, where heating is powered by electricity and mineral

fuel in winter. Air pollution from the heating companies has a negative impact on air quality. There are also other reasons like climate change, and automobile exhaust emissions. For example, there are generally windy days during the spring in northern China, which will help disperse the suspended particulates (PM2.5, PM10), resulting in relatively good AQI.

There are many supervised learning methods designed for time series analysis, like support vector machine (Jaramillo et al., 2017), neural network (Sahoo et al., 2019) and decision trees (Li et al., 2020). Instead of using "black box" models in this paper, a decision tree model is used for AQI prediction due to its good interpretability. To incorporate the effects of autocorrelation and seasonality, the wavelet transform is applied to the original time series including the spatial weighted matrix. Wavelet transformed time series can represent information via the decomposition into different resolution levels, which will help increase the accuracy of decision tree based modeling. Wavelet transform-based decision trees have been applied to panel data by Zhao et al. (2018), who have shown that wavelet transformed variables can be better than original variables for classification. This paper will also explore whether or not wavelet transformed variables are still better than original variables for forecasting using regression methods, as Zhao et al. (2021) found in the context of forecasting streaming data.

We introduce our basic model in Section 3, and apply it to simulated time series experiments in Section 4 before analysing our AQI time series in Section 5 and COVID data in Section 6. Some concluding comments appear in Section 7. All calculations were done using R (R Core Team, 2018); 'waveslim' (Whitcher, 2019) was used for wavelet decompositions and 'rpart' (Therneau et al., 2014) for CART as the decision tree method.

3. Model description

3.1. Decision tree

The decision tree we use is the classification and regression tree (CART; Breiman et al., 1984). Regression trees are similar to classification trees, except that the response variable is numerical rather than categorical as for classification trees. In a typical binary split regression tree, the input attribute space is recursively partitioned by a sequence of binary splits leading to terminal nodes. At each terminal node, the predicted response value is the mean of the data in that node. Compared to a classification tree, the main difference in constructing the tree is the impurity criterion. For regression trees, we use mean squared error, which measures the difference between the data and the predicted value. Thus a regression tree is formed

by iteratively splitting nodes so as to maximize the decrease of mean squared error at each step.

Any variable which is used to split nodes may appear in the tree many times, either as a primary or a surrogate variable (to replace the primary variable when it is missing). An overall score to measure variable importance is the sum of the goodness of split measures for each split for which it was the primary variable, plus goodness multiplied by the adjusted agreement for all splits for which it was a surrogate. These are scaled to sum to 100 and the rounded values are shown, omitting any variable whose proportion is less than 1%, as detailed in Therneau et al. (2014).

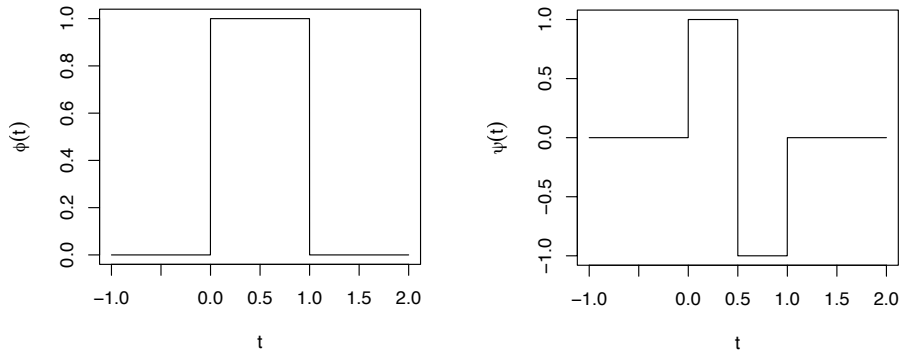
This approach gives a tree which can be used to describe the observed data, which is not necessarily optimal for making predictions on new data. A bigger, more complex, tree means better accuracy in fitting the training data but not necessarily improvement in predicting test data outcomes, so we need to prune the tree to make it robust to the test data. A tree with a fixed size has a corresponding complexity parameter α which measures the balance between tree depth and accuracy. In other words, for a data set, when the α value is given as a stopping criteria, the tree is decided. For a larger tree, each value of α can be used to determine the amount of pruning. So the objective is to find the best α value (or say the best tree size) which decides a tree that has the best prediction ability in *test* data. Typically, α is selected by k-fold cross validation.

3.2. Wavelet transform

Sometimes the predictive data have nearly the same range and mean but different frequencies like sine functions with different periods. Unfortunately, decision trees can not easily classify or do regression on such data as there is no obvious split point to separate them. Additionally, for time series data, there might be dependence between successive observations; growing trees which treat each time point as an independent observation may ignore this information. Wavelet analysis can deal with these problems by picking out patterns in short term fluctuations in data which can be exploited for prediction when consecutive observations lack independence or when key information is encoded as frequencies in the data. Using a wavelet transform to discover signal information at different resolution levels is like using a camera to enjoy landscape pictures at different scales. A camera lens can take broad landscape pictures as well as zoom in to capture microscopic detail that is not easily seen by the human eye.

The wavelet transform method we choose is the maximal overlap discrete wavelet transform (MODWT; see, for example, Percival and Walden, 2000), which is a non-decimated wavelet transform. Unlike some other wavelet transforms, the MODWT

is not constrained to data whose sample size is a power of 2. The wavelet transform describes the data in terms of a collection of basis functions, all constructed from a mother wavelet and scaling function. One such wavelet basis is constructed from the Haar wavelet, which is easy to understand and simple to interpret.



(a) (b)

Figure 1: The Haar scaling function (a) and mother wavelet (b).

As shown in Figure 1(a), the Haar scaling function is defined as

$$\phi(t) = \begin{cases} 1 & t \in [0,1) \\ 0 & \text{else.} \end{cases} \tag{1}$$

Using dilation and translation, the scaling function at resolution level j and location k is

$$\phi_{j,k}(t) = 2^{j/2}\phi\left(2^j t - k\right).$$

Note that $\phi_{j,k}(t)$ is compactly supported on $I_{j,k} = [2^{-j}k, 2^{-j}(k+1))$, where $j = 0, 1, 2, \ldots, J$ $(J = \lfloor \log_2 n \rfloor)$ and $k = 0, 1, 2, \ldots n-1$, and that

$$\phi\left(2^j t - k\right) = \begin{cases} 1 & t \in I_{j,k} \\ 0 & \text{else.} \end{cases}$$

When $j = 0$, the scaling coefficients are actually the original time series. Thus, when j is small, wavelets are highly localized at a fine scale resolution level, representing brief transient effects. Conversely, when j is large, wavelets represent lower frequency activity at coarser scale resolution levels (Aykroyd et al., 2016). Here the factor of $2^{j/2}$ ensures energy preservation, defined by

$$\text{energy} = \frac{1}{2}\int_0^{2\pi} |f(x)|^2 dx \tag{2}$$

so that the energy in the data set will be preserved and that is why wavelets are orthogonal and have inverse transforms; for more detail, see Graps (1995). Then, the scaling coefficients $s_{j,k}$ can be calculated as

$$s_{j,k} = <x(t), \phi_{j,k}> = \int_R x(t)\,\phi_{j,k}(t)\,dt = 2^{j/2}\int_{I_{j,k}} x(t)\,dt.$$

As shown in Figure 1(b), the Haar mother wavelet function $\psi(t)$ is defined as

$$\psi(t) = \begin{cases} 1 & t \in [0, 0.5) \\ -1 & t \in [0.5, 1) \\ 0 & \text{else,} \end{cases}$$

and the wavelet function at resolution level j and location k is $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$. The wavelet coefficients $d_{j,k}$ are defined as $d_{j,k} = <x(t), \psi_{j,k}>$, but in practice can be calculated as $d_{j,k} = s_{j-1,k} - s_{j-1,k-1}$. The scaling coefficients vector $s_j = (s_{j,0}, s_{j,1}, \ldots, s_{j,n-1})$ and wavelet coefficient vector $d_j = (d_{j,0}, d_{j,1}, \ldots, d_{j,n-1})$ become new variables which can be used for classification and regression. We refer to them as scale variables (or scale information) and detail variables (or detail information) respectively in the following sections.

The Haar wavelet is the simplest member of the family of Daubechies orthogonal wavelets, which also includes the D4 and D6 wavelets with filter lengths of 4 and 6 respectively, as well as LA8 which is the Daubechies least asymmetric wavelet of length 8. Simulation is conducted in Section 4 to compare the performance of different wavelet bases under different settings. The relatively best wavelet is applied in the real data application Sections 5 and 6.

The implementation of the MODWT in the R package waveslim (Whitcher, 2019) ignores the energy preservation factor $2^{j/2}$, which will not affect the results when using classification or regression trees. Since coarse-scale wavelets span large data intervals, methods are required for dealing with boundary conditions. Typical boundary correction methods are to assume periodicity of the data or to reflect data at the boundaries. In this paper, reflection is chosen, as the time series is not periodic. After reflection, the time series of length n, becomes one of length $2n$, so the first n wavelet coefficients of the transform are used as the wavelet-transformed variables. The maximum level of the wavelet transform depends on the length of the test data, n_t and should not exceed $J = \lfloor \log_2 n_t \rfloor$.

4. Simulation

In order to establish when wavelet-transformed decision trees can pick out useful information, we conduct a simulation study under different seasonal effect levels

and for a range of forecast horizons. The air quality data we wish to forecast has autoregressive (AR) properties and therefore our simulation study is based on data generated from AR-based models. It is easy to understand that the current air quality is highly correlated with the previous values and there is some seasonal property as well, so we also include seasonal effects in the simulation study. The AR time series generated follow

$$y_t^{\text{raw}} = a_1 \cdot y_{t-1}^{\text{raw}} + a_2 \cdot y_{t-2}^{\text{raw}} + a_3 \cdot y_{t-3}^{\text{raw}} + \cdots + a_{12} \cdot y_{t-12}^{\text{raw}} + \epsilon_t, \tag{3}$$

where $\epsilon_t \sim N(0,1)$, $[a_1, \ldots, a_{12}] = [0, 0, 0.1, 0.8, 0, 0, 0, 0, 0, -0.1, 0.1, -0.5]$, and $t = 1, 2, \ldots, T$. In the model for y_t^{raw}, the AR parameters were chosen to ensure a stationary time series and both short and long time lag effects are included. To make the simulated time series y_t^{raw} more similar to air quality data, we also add a seasonal effect. Supposing a time series that is collected daily, and has a sine-shaped seasonality around the year, we use the seasonal effect function

$$y_t^{\text{season}} = \alpha \sin(t' \cdot 2\pi/365), t' = t \bmod 365,\ t = 1, 2, \ldots T;$$

then y_t^{season} is added to y_t^{raw} to obtain y_t. An example realization is shown in Figure 2.

In the context of our spatial scenario, after generating y_t we also generate a time series y_t^{SW} that has some correlation with y_t, representing the influence from neighboring areas. We assume

$$y_t^{SW} \sim \mathcal{N}(y_t, \tau^2), \tag{4}$$

where τ controls the strength of the relationship between y and y^SW; lower τ, resulting in y_t^{SW} being more correlated to y_t. We now describe the process of creating wavelet-derived predictive variables from the observed data. We show the operation on y_t for illustration; in practice, the same operation will be conducted on y_t^{SW} as well.

Let $A = [y_1, y_2, \ldots, y_T]^T$, then without a wavelet transform, a schematic matrix representation of our prediction is

$$A^{\text{lag}} = \begin{bmatrix} y_1 & y_2 & \cdots & y_{k.\text{lag}} \\ y_2 & y_3 & \cdots & y_{k.\text{lag}+1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{T-k.\text{lag}-p} & y_{T-k.\text{lag}-p+1} & \cdots & y_{T-p} \end{bmatrix} \longrightarrow Y = \begin{bmatrix} y_{k.\text{lag}+p} \\ y_{k.\text{lag}+p+1} \\ \vdots \\ y_T \end{bmatrix}. \tag{5}$$

where A^{lag} is our raw explanatory variable without wavelet transform and Y is our response variable. One thing to note is, in the model training process, the data are

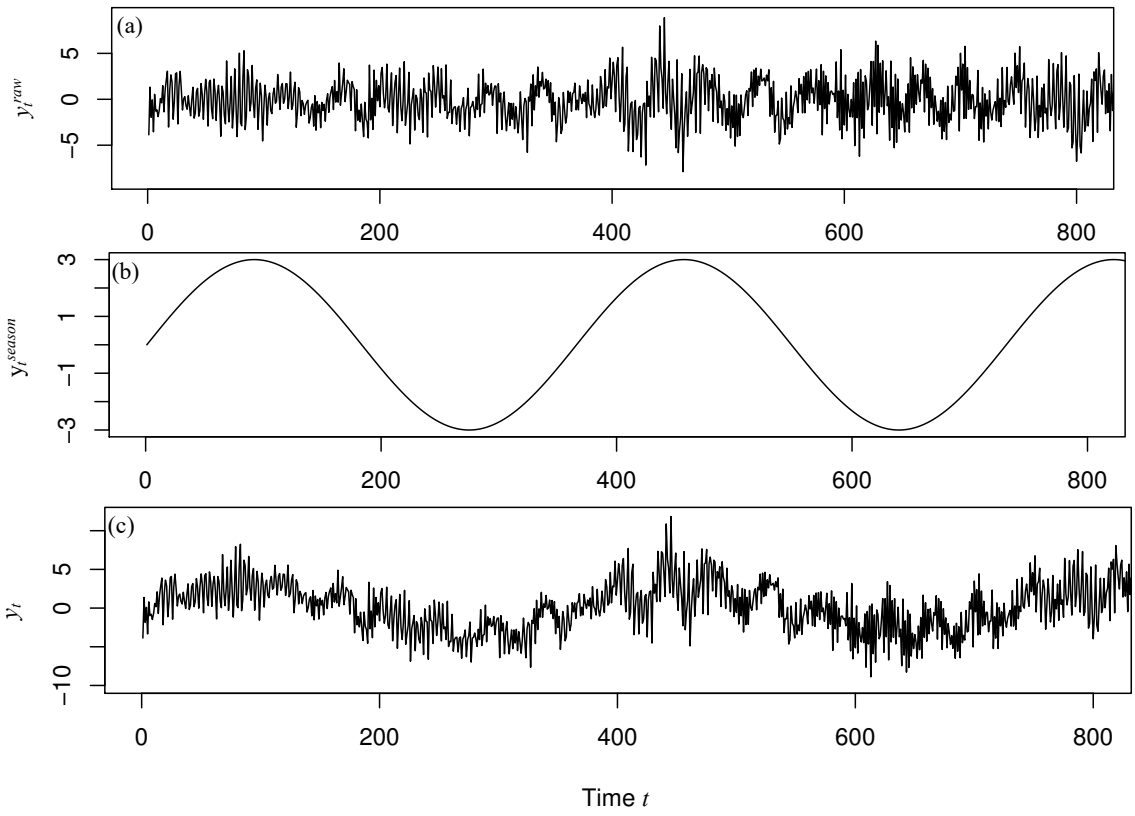


Figure 2: Generation of time series simulation data. Only 800 observations are shown for illustration. (a) is the raw time series without seasonal effect. (b) is the seasonal effect with $\alpha = 3$. (c) is the time series with seasonal effect added.

trained in batches, so some lines instead of one line of A will be trained. But actually we are not using many lines of the matrix A^{lag} at one time to predict each future y_t, but using each line to predict each time point. The matrix just represents a batch of data. By giving data y_t for $t = 1, 2, \dots T$, our predictions are implicitly of the form

$$\hat{y}_{(t+p)} = f(y_t, y_{t-1}, \dots, y_{t-k.\text{lag}}),$$

which means y will be predicted p steps ahead, by using observed y values from the current time t back to the previous k.lag values.

We obtain wavelet transformed data of each column i in A^{lag} using the MODWT for levels $j = 1, 2, \dots, J$. There are $J + 1$ scale variables $s_0, s_1, \dots s_J$ and J detail variables $d_1, d_2, \dots, d_J$, making a total of $2J+1$ variables. These variables are stacked side-by-side to obtain a $(T - k.\text{lag} - p) \times (2J + 1)$ matrix $W^{\text{lag}.i}$ for column i:

$$W^{\text{lag}.i} = \left[\begin{array}{ccc|ccc} W_i^{d_1} & \cdots & W_i^{d_J} & W_i^{s_0} & \cdots & W_i^{s_J} \\ W_{i+1}^{d_1} & \cdots & W_{i+1}^{d_J} & W_{i+1}^{s_0} & \cdots & W_{i+1}^{s_J} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ W_{T-k.\text{lag}-p+i-1}^{d_1} & \cdots & W_{T-k.\text{lag}-p+i-1}^{d_J} & W_{T-k.\text{lag}-p+i-1}^{s_0} & \cdots & W_{T-k.\text{lag}-p+i-1}^{s_J} \end{array}\right].$$

The final wavelet transformed variable can be expressed as $W^{\text{lag}} = \left[W^{\text{lag}.1}, W^{\text{lag}.2}, \cdots, W^{k.\text{lag}}\right]$. In the same way, we can also get $(W^{SW})^{\text{lag}}$ for A^{SW}.

We have now finished constructing the explanatory variables as W^{lag} and $(W^{SW})^{\text{lag}}$. With the response variable $Y = A_{(k.\text{lag}+p:T)}$, the prediction function can be represented as

$$Y = f\{W^{\text{lag}}, (W^{SW})^{\text{lag}}, k.\text{lag}, p\}$$

with the requirement $k.\text{lag} + p + 1 \leq T$.

The accuracy of the results is measured by R-squared, RMSE and MAE. Suppose $y_{t\in S}$ is the data set for measurement with number of observations as m. Then we have

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}},$$

where $SS_{res} = \sum_{t\in S}(y_t - \hat{y}_t)^2$ and $SS_{tot} = \sum_{t\in S}(y_t - \bar{y})^2$, in which $\hat{y}_t$ is the fitted value of y_t and $\bar{y}$ is the mean of data set $y_{t\in S}$. RMSE is the root mean squared error:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{t\in S}(y_t - \hat{y}_t)^2}, \tag{6}$$

and MAE is the mean absolute error:

$$\frac{1}{m}\sum_{t\in S}\left|y_t - \hat{y}_t\right|. \tag{7}$$

Table 1: Values used in the simulations for parameters p, τ, α and $k.lag$

parameter	values
p	1, 5, 10, 15, 20
τ	1, 3, 5
α	1, 3, 5
k.lag	1, 5, 10, 15, 20

By generating 2000 observations from Equation 3, we use 80% of the data for training and the rest for testing. The parameters chosen have the options specified in Table 1.

When $p = 1$, other parameters will run through all their possible values. Under each different parameter settings, for example the setting ($p = 1$, $\tau = 1$, $\alpha = 1$ and k.lag $= 1$), we conduct the simulation 50 times, with the results shown in Figure 3. When k.lag increases, the longer the lag, the more information is used in the regression, which leads to a better performance with higher R-squared, lower RMSE and MAE. When p increases, the model will predict an observation with wider forecast horizon. From Figure 3, the trend is generally decreasing in R-squared, and increasing in RMSE and MAE, which means the model is decreasing in performance when the forecast horizon is wider. When p has the values 5, 10 or 15, the performance tends to remain unchanged. The reason is that the simulated data are based on AR process with lags 3, 4, 10, 11 and 12, p with values 5, 10 or 15 are nearby so result in better performance. When the seasonal effect level α increases, the data have more obvious pattern, which results in better performance. If $J = 0$, the data are not wavelet transformed, leading to markedly worse performance than when wavelet transformed data are used. Even with only two resolution levels of wavelet transformation, there is a marked increase in performance. But when J continues to increase, except Haar wavelet, D4, D6 and LA8 all decrease in performance which might due to the overfitting phenomenon with higher vanishing moments. The parameter τ has little influence on the forecast performance, but it has influence on the variables selected as shown in Figure 4. Overall, Haar wavelet shows the best performance relatively, with higher R-squared, lower RMSE and MAE. The rest the simulation and real data analysis choose Haar wavelet in the wavelet transform.

For interpretation, it is helpful to assess which predictor variables are the most

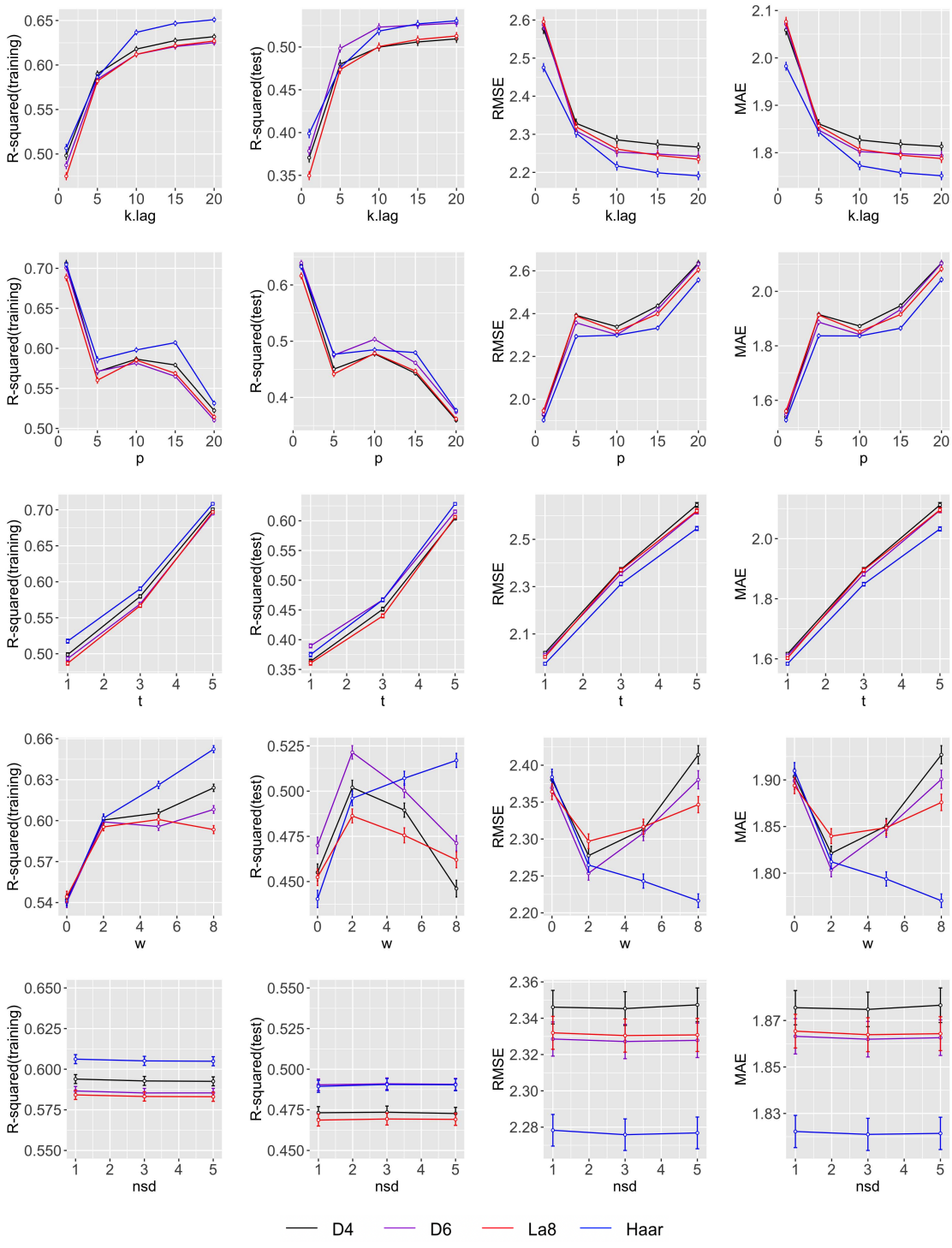


Figure 3: Forecast accuracies with 95% confidence interval under different time lag length k.lag, forecast horizon p, seasonal effects α, wavelet resolution level J and spatial influence parameter τ using simulated data. Four wavelet basis applied are D4, D6, La8 and Haar.

useful ones in the forecasting process. In Figure 4, we average the importance score for the predictor variables across all the trials under the wavelet transform resolution level $J = 8$ (maximum allowance). D4, D6 and LA8 are also tried but result with bad performance.

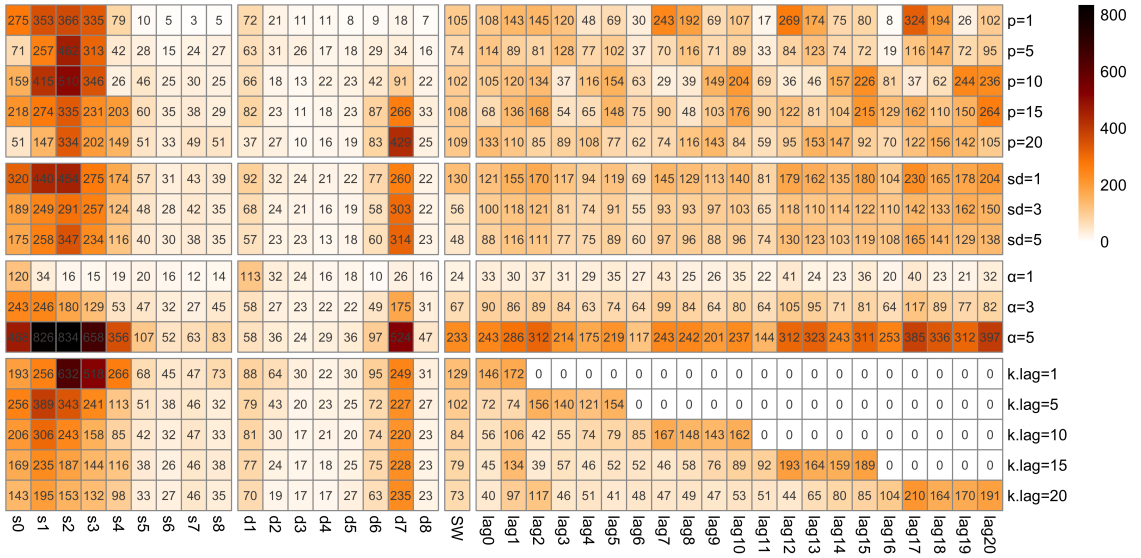


Figure 4: Variables selected by the model under different time lag length k.lag, forecast horizon p, seasonal effects α and spatial parameter τ using Haar-wavelet transformed simulated data. The number in each cell is the averaged variable importance score.

When the forecast horizon p changes from 1 to 20, the wavelet transformed variables on coarse resolution level become more important while those on fine resolution levels become less important. For example, the importance of s_8 increases from 5 to 51 and that of d_7 changes from 18 to 429, but for s_0 and d_1, their importance scores decrease from 275 and 72 to 51 and 37 respectively. When $p = 20$, d_7 has a high importance score, which means the data information of past 128 observations plays a more signficant role than others. As the time series data contains a seasonal effect following a sine function with a period of 365 observations, this is represented at resolution level 7 since 128 is closer to the middle of 365 than $2^6 = 64$ or $2^8 = 256$. For the information provided by the scale variables, models choose more recent scale variables like s_1 to s_4 instead of long scale information. For spatial weighted information SW and lag information, they change little when p changes.

When the spatial parameter τ changes, the scale and detail information changes little. As the spatial variable is simulated by using the original variable, there is no

extra information added. The τ of the noise added to the spatial variables influences whether the spatial information will be selected by the model or not. That leads to the result that the importance score of SW decreases when τ increases.

When the seasonal effect α increases, the sine trend of the time series becomes obvious. The importance score of most variables has a sharp increase, especially when α increases from 3 to 5, including the spatial information SW, the scale information, lag information and part of the detail information. As α increases, the model tends to choose relatively more coarse scale variables like s_2 and s_1. This is because, with a higher seasonal effect level, the sine pattern becomes the main trend in the data compared to the AR pattern and relatively more coarse scale variables can weaken the AR pattern but also can keep the sine trend.

When the permitted lag length k.lag increases, the importance of scale and detail information of low resolution level weakens and most of that on high resolution level performs similar as that when k.lag is small. This is because scale and detail information on low resolution level share similar information with lag information, so their importance weakens as the permitted lag length increases. The variable which retains its importance is d_7, which has the detail information of 128 observations. For the lag information in use, lag_1 and lag_2, as well as the largest four lag variables, like lag_{17} to lag_{20} when $k.\text{lag} = 20$, are more frequently selected by the model in terms of importance. This is consistent with that the model takes s_0 to s_4 more frequently as important variables.

These simulations have shown that forecasts based on wavelet transformed variables can be better than those using the original variables under different conditions. The wavelet-transform based decision trees have a better interpretability than when using only the original variables. We now apply our wavelet-based forecasting to real data.

5. Application to air pollution data

The AQI dataset collected comes from the China Air Quality Online Monitoring and Analysis Platform (`https://www.aqistudy.cn/`) which summaries information from the data centre of the Ministry of Environmental Protection of the People's Republic of China (`http://www.mep.gov.cn/`). Since the data are shown in maps, which are not available for downloading directly, we collect the data manually. We obtained daily data of 31 provinces in China (except Hong Kong, Macao and Taiwan) from 13th May 2014 to 14th December 2019, making a total of 2021 observations, except some missing data. The data can be described as

$$A = [A_{.,1}, A_{.,2}, \ldots, A_{.,T}]$$

where $A_{.,t} = [y_{1,t}, y_{2,t}, \ldots, y_{N,t}]^T$, for $t = 1, 2, \ldots, T$ and $n = 1, 2, \ldots, N$ with $T = 2021$ and $N = 31$.

Since air pollution has spatial spillover effects, one province's AQI can be influenced by that of neighbouring provinces. We incorporate these spatial spillover effects as a separate variable in our forecast models. For detail of spatial regression analysis, see Ward and Gleditsch (2018). There are many ways to construct a spatial weight matrix. In this section, it is assumed that only geographically contiguous provinces share spatial influence. In this scenario, we have an adjacency matrix, B, with elements

$$B_{i,j} = \begin{cases} 1 & \text{if provinces } i \text{ and } j \text{ are neighbours } (i, j \in \{1, 2, \ldots, N\}) \\ 0 & \text{otherwise including } i = j. \end{cases}$$

Then, a standardized matrix, $B^{(s)}$ is constructed by making the sum of each row equal to 1. So the elements of the standardized adjacency matrix are

$$B^{(s)}_{i,j} = B_{i,j} / \sum_k B_{i,k}, \;\; i, j \in 1, \ldots, N$$

So, we define variables to capture spatial spillover effect as

$$(A^{SW}_{.,t})^T = A^T_{.,t} \, B^{(s)},$$

which measures the spillover effect from all the neighbouring provinces to the current province and the combined spatial weighted variable matrix is

$$A^{SW} = [A^{SW}_{.,1}, A^{SW}_{.,2}, \cdots, A^{SW}_{.,T}].$$

In order to detect whether wavelet transformed variables have a better performance in forecasting, we apply MODWT to both $A_{n,.}$ and $A^{SW}_{n,.}$ for each separate province, n at level j. (Since the time length T is 2021 and the training data is set as 80% of the whole data, considering the data cost in function, we choose the maximum resolution level J as 9 and hence $j = 1, 2, \ldots, 9$.) The wavelet transformed data MODWT$(A_{n,.})$ for province n are denoted as the $T \times (2J + 1)$ matrix

$$W_{n,.} = \left[\begin{array}{ccc|ccc} W^{d_1}_{n,1} & \cdots & W^{d_J}_{n,1} & W^{s_0}_{n,1} & \cdots & W^{s_J}_{n,1} \\ W^{d_1}_{n,2} & \cdots & W^{d_J}_{n,2} & W^{s_0}_{n,2} & \cdots & W^{s_J}_{n,2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ W^{d_1}_{n,T} & \cdots & W^{d_J}_{n,T} & W^{s_0}_{n,T} & \cdots & W^{s_J}_{n,T} \end{array} \right].$$

Here, s_0 is also included as it is actually the original variable. The wavelet transformed data are then

$$W = [W_{1,.}^T, W_{2,.}^T, \cdots, W_{N,.}^T]^T.$$

Similarly, we get the wavelet transformed data W^{SW} corresponding to spatial spillover effects variable A^{SW}.

Taking the original data $A_{n,.}$ as an example, with time lag i included, it becomes

$$A_{n,.}^{\text{lag}.i} = [y_{n,i+1}, y_{n,(i+2)}, \cdots, y_{n,(T-k.\text{lag}+i-p)}]^T,$$

where p is the forecast horizon and $A_{n,.}$ with time lag becomes

$$A_{n,.}^{lag} = [A_{n,.}^{\text{lag}.0}, A_{n,.}^{\text{lag}.1}, \cdots, A_{n,.}^{k.\text{lag}}].$$

We set the maximum time lag as 20 (lag.max) and other values can be also considered. After combination, we have A^{lag}. In the same way, we can also get W^{lag} and $(W^{SW})^{\text{lag}}$.

We have now finished constructing the explanatory variables as W^{lag} and $(W^{SW})^{\text{lag}}$. The response variable is

$$Y = [A_{1,(k.\text{lag}+p+1:T)}, A_{2,(k.\text{lag}+p+1:T)}, \cdots, A_{N,(k.\text{lag}+p+1:T)}]^T.$$

The function can be constructed as

$$Y = f\{W^{\text{lag}}, (W^{SW})^{\text{lag}}, k.\text{lag}, p\}$$

with the requirement $k.\text{lag} + p + 1 \leq T$

Figure 5 shows how the model performance changes according to different combinations of J, k.lag and p. In the first column, under different k.lag and J settings, both R-squared (training) and R-squared(test) decrease as the forecast horizon p increases while RMSE and MAE increases. This result is the same as expected since a larger p means a longer AQI will be predicted which will have a relatively worse performance. This is also true in general when J changes except some points. A larger J means longer AQI information is processed for forecasting. But when it comes to k.lag, the results show a different trend. The performance metrics changes little when k.lag changes. The reason is J means as long as 2^J observations of AQI are included in the modeling, however, k.lag only means k.lag observations are included. When $J = 0$, the data are the original data without being wavelet transformed. The original data have worse performance than the wavelet transformed data measured by the R-squared (training), but not obviously worse in R-squared (test), RMSE and

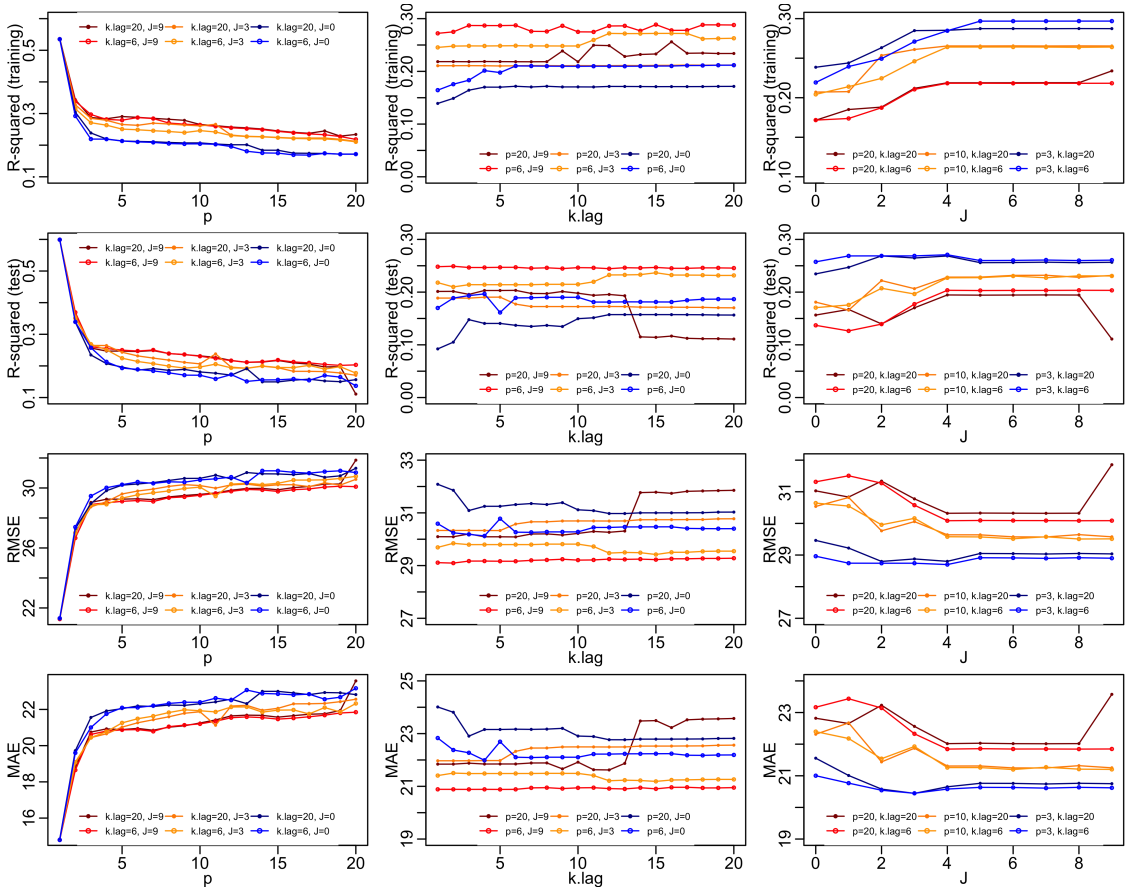


Figure 5: AQI prediction accuracy results with respect to different J, k.lag and p.

Table 2: Important variables chosen by the decision tree. The importance measurement is combined by the importance of variable as first split variable and surrogate variable.

important variables	lag	scale information	detail information
AQI	lag_{20}	$s_0 - s_5$	$d_1\ d_2\ d_9$
	lag_{19}	$s_0 - s_5$	d_9
	lag_{18}	$s_2 - s_5$	d_9
	lag_{17}	$s_2 - s_5$	d_9
	lag_{16}	$s_2 - s_5$	d_9
	lag_{15}	$s_4 - s_5$	d_9
	$\text{lag}_{14} - lag_{10}$	s_4	–
AQIW	$\text{lag}_{15} - \text{lag}_{20}$	–	d_9

information with "–" means no such information is selected as important variables by the tree.

MAE especially when p is small like 3. This is because with a small p, recent data will be sufficient to provide the forecasting information.

When we set $J = 9$, $k.\text{lag} = 20$ and p ranges from 1 to 20, the important variables selected by the model are shown in Table 2.

The results in Table 2 show that when long lag information is permitted, scale information with a period no longer than two months is more preferred, like s_5, which means data with a lag of 32 days is selected. For detail information, the resolution level of 9 is chosen which means information as long as 512 days is better than others. Since long lag information is selected instead of short lag information which means the future AQI is more likely to be influenced by its information of weeks or months instead of recent days. The spatial weighted AQI variables are also chosen as important variables, which confirms the assumption air pollution has spatial spill over effect and one province's AQI will be influenced by its conjugate provinces' AQI.

6. Application to COVID-19 data

COVID-19, caused by coronavirus SARS-Cov-2, is a novel pneumonia firstly noticed in December 2019 as some hospitals in Wuhan diagnosed several cases of unexplained pneumonia. It shares similar properties with the AQI data such as spatial effect (Guliyev, 2020) and time autocorrelation effect (Liu et al., 2021).

As an infectious disease caused by the most recently discovered coronavirus, COVID-19 spreads across the world. Since there is no effective treatment therapy, the disease gets quite strong spatial transmission ability, which causes terrible result. But with the effective and strict government control policies, the disease is limited to some extent. Similar to the AQI data analysis, the spatial effect is also included in the COVID-19 data analysis to test whether the government policies play a role in disease control or not.

As there is no instant medical method to confirm the infection, there is a time gap between the test and confirmation, which boosts the spread of the disease. The number of confirmed cases today autocorrelate with its previous number. Also similar to the AQI data, the autocorrelation effect is included in the analysis.

The data of cumulative confirmed cases in 31 provinces and cities in Mainland China during 16th of January 2020 to 1st February 2020 are collected from http://2019ncov.chinacdc.cn/2019-nCoV/ and analyzed with the proposed method. As the disease has complex unseen changing patterns which are hard for any forecasting method to learn and predict, test accuracy is lower than the training accuracy as shown in Figure 6. As the forecast length p increases, the model shows little change in performance, which means forecasting with small interval gap can be effective.

With the k.lag increases, both R-squared (training) and R-squared(test) increases a little bit. When wavelet transform level J increases, the model has small increase in R-squared (training) but show little difference on other metrics.

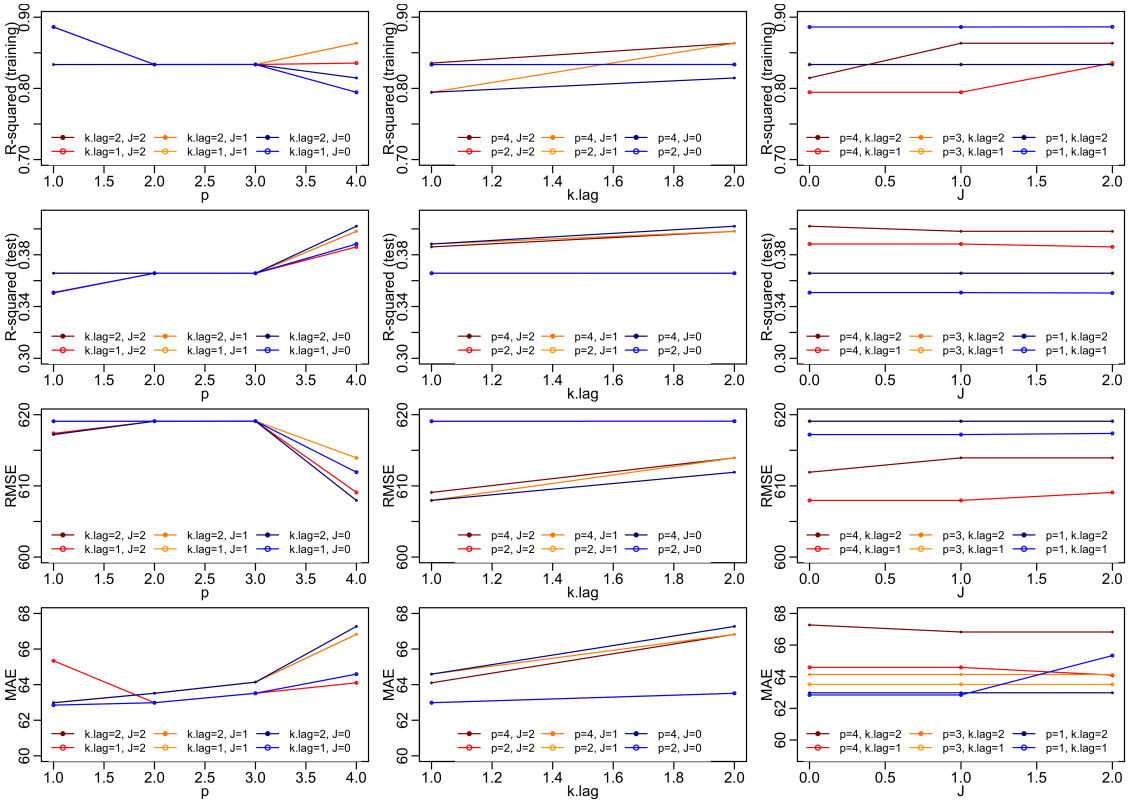


Figure 6: COVID prediction accuracy results with respect to different J, k.lag and p.

The variables selected in the importance list include lag_0 and lag_1, both with d_1 and $s_0 - s_2$ as detail and scale information. Compared to the AQI data, the variable importance list does not include the spatial weighted variables, which may be patially due to the lock-down policies in China during that period being effective in preventing mutual influence.

7. Conclusion

In conclusion, we built a model by using original or wavelet transformed variables in forecasting spatio-temporal data and compared the performance of the original

variables with wavelet transformed variables. Wavelet transformed data can pick out scale information and detail information of the original variables, which result in a generally better accuracy measured by R-squared (both training and test), RMSE and MAE. Interpretation via considering which variables are used in the tree is also obtained. In simulation, the performance of Haar, LA8, D4 and D6 wavelets are compared, and Haar wavelet shows the best performance in most cases.

Specifically, MODWT based CART can detect true lag information and has much better performance when only short lag information is permitted (k.lag is small). When the forecast horizon increases, performance of MODWT based forecasting decays more slowly in performance than that of original-data forecasting. In the real data analysis, MODWT based CART also performs better in the air pollution data of China. This also shows air pollution time series has autocorrelation effect. There is a spillover effect of the air pollution from neighbouring provinces as the spatial weighted variable is selected by the model in the variable importance list.

For the air pollution analysis, we find that long lag effects seem to exist, for example d_9 is selected in the important variables list, which contains information of 512 days in length confirming the intuitive explanation that long lag effects exist which partly explains the annual seasonal effect.

For the COVID analysis, spatial weighted variables are not selected by the model in the importance list. The conjugate areas have little influence to each other mutually during the lock-down period.

Further research can be conducted in the context of streaming data analysis with spatial and seasonal effects included in the time series. For the wavelet method, we used the Daubechies families of compactly-supported wavelets (Daubechies, 1992) in the simulation section, a further direction would be to consider when different basis functions are preferred. Other alternative wavelet transforms such as the Haar-Fisz wavelet transform (Fryzlewicz and Nason, 2004) can also be considered if the distribution of the time series is partly accessible.

Acknowledgements

This research is funded by the Fundamental Research Funds for the Central Universities (2242020R40073, MCCSE2021B02, 2242020R10053), Guangdong Basic and Applied Basic Research Foundation (2020A1515110129), Natural Science Foundation of Jiangsu Province (BK20200347, BK20210218), Nanjing Scientific and Technological Innovation Foundation for Selected Returned Overseas Chinese Scholars (1107010306, 1108000241), Jiangsu Foundation for Innovative and Entrepreneurial Doctor (1108000245). National Natural Science Foundation of China (62103105, 12171085).

Declarations

The authors declare that they have no conflict of interest.

References

AirNow (2020). Aqi basics, https://www.airnow.gov/aqi/aqi-basics/. Accessed: 2020-06-03.

AirVisual, I. (2018). 2018 world air quality report, https://www.airvisual.com/world-most-polluted-cities/world-air-quality-report-2018-en.pdf. Accessed: 2020-04-03.

Aykroyd, R. G., Barber, S. and Miller, L. R. (2016). Classification of multiple time signals using localized frequency characteristics applied to industrial process monitoring, *Computational Statistics & Data Analysis* **94**: 351–362.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Wadsworth and Brooks, Monterey, CA.

Chen, S., Zhang, Y., Zhang, Y. and Liu, Z. (2019). The relationship between industrial restructuring and China's regional haze pollution: A spatial spillover perspective, *Journal of Cleaner Production* **239**.

Daubechies, I. (1992). *Ten Lectures on Wavelets*, SIAM, Philadelphia.

Du, P., Du, R., Ren, W., Lu, Z. and Fu, P. (2018). Seasonal variation characteristic of inhalable microbial communities in PM2.5 in Beijing city, China, *Science of the Total Environment* **610-611**: 308 – 315.

Dua, R. D., Madaan, D. M., Mukherjee, P. M. and Lall, B. L. (2019). Real time attention based bidirectional long short-term memory networks for air pollution forecasting., *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), Big Data Computing Service and Applications (BigDataService), 2019 IEEE Fifth International Conference on* p. 151.

Emetere, M. E. (2018). Environmental data retrieval and prediction using the autoregression moving average and polynomial experimentation, *Aerosol Science and Engineering* **2**(3): 99–108.

EPA (2020). Aqi breakpoints, https://aqs.epa.gov/aqsweb/documents/codetables/aqi_breakpoints.html. Accessed: 2020-06-03.

Finardi, S., De Maria, R., D'Allura, A., Cascone, C., Calori, G. and Lollobrigida, F. (2008). A deterministic air quality forecasting system for torino urban area, italy, *Environmental Modelling & Software* **23**(3): 344–355.

Fryzlewicz, P. and Nason, G. P. (2004). A Haar-Fisz algorithm for Poisson intensity estimation, *Journal of Computational and Graphical Statistics* **13**(3): 621–638.

Graps, A. (1995). An introduction to wavelets, *IEEE Computational Science and Engineering* **2**(2): 50–61.

Guliyev, H. (2020). Determining the spatial effects of covid-19 using the spatial panel data model, *Spatial Statistics* **38**: 100443.

Gupta, P., Christopher, S. A., Wang, J., Gehrig, R., Lee, Y. and Kumar, N. (2006). Satellite remote sensing of particulate matter and air quality assessment over global cities, *Atmospheric Environment* **40**(30): 5880–5892.

Gurjar, B., Butler, T., Lawrence, M. and Lelieveld, J. (2008). Evaluation of emissions and air quality in megacities, *Atmospheric Environment* **42**(7): 1593–1606.

Guyu, Z., Huang, G., He, H., He, H. and Ren, J. (2019). Regional spatiotemporal collaborative prediction model for air quality, *IEEE Access* **PP**: 1.

Hoshyaripour, G., Brasseur, G., Andrade, M., Gavidia-Calderón, M., Bouarar, I. and Ynoue, R. Y. (2016). Prediction of ground-level ozone concentration in são paulo, brazil: deterministic versus statistic models, *Atmospheric environment* **145**: 365–375.

Jaramillo, J., Velasquez, J. D. and Franco, C. J. (2017). Research in financial time series forecasting with svm: Contributions from literature, *IEEE Latin America Transactions* **15**(1): 145–153.

Kaminska, J. A. (2018). The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wroclaw, *Journal of Environmental Management* **217**: 164 – 174.

Kim, S. E., Honda, Y., Hashizume, M., Kan, H., Lim, Y.-H., Lee, H., Kim, C. T., Yi, S.-M. and Kim, H. (2017). Seasonal analysis of the short-term effects of air pollution on daily mortality in Northeast Asia, *Science of the total environment*, p. 850.

Leong, W., Kelani, R. and Ahmad, Z. (2019). Prediction of air pollution index (API) using support vector machine (SVM), *Journal of Environmental Chemical Engineering* **8**(3): 103208.

Li, L., Dai, S., Cao, Z., Hong, J., Jiang, S. and Yang, K. (2020). Using improved gradient-boosted decision tree algorithm based on kalman filter (gbdt-kf) in time series prediction, *The Journal of Supercomputing* pp. 1–14.

Li, R., Cui, L., Meng, Y., Zhao, Y. and Fu, H. (2019). Satellite-based prediction of daily SO2 exposure across China using a high-quality random forest-spatiotemporal kriging (RF-STK) model for health risk assessment, *Atmospheric Environment* **208**: 10–19.

Li, Y., Shen, X., Han, D., Sun, J. and Shen, Y. (2018). Spatio-temporal-aware sparse denoising autoencoder neural network for air quality prediction, *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pp. 96–100.

Liu, L., Hu, T., Bao, S., Wu, H., Peng, Z. and Wang, R. (2021). The spatiotemporal interaction effect of covid-19 transmission in the united states, *ISPRS International Journal of Geo-Information* **10**(6): 387.

Ma, J., Cheng, J. C., Lin, C., Tan, Y. and Zhang, J. (2019). Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques., *Atmospheric Environment* **214**: 116885.

Patni, J. and Sharma, H. (2019). Air quality prediction using artificial neural networks, *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, pp. 568–572.

Percival, D. B. and Walden, A. T. (2000). *Wavelet Methods for Time Series Analysis*, Cambridge University Press, Cambridge.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Rubal and Kumar, D. (2018). Evolving differential evolution method with random forest for prediction of air pollution., *Procedia Computer Science* **132**(International Conference on Computational Intelligence and Data Science): 824 – 833.

Sahoo, B. B., Jha, R., Singh, A. and Kumar, D. (2019). Long short-term memory (lstm) recurrent neural network for low-flow hydrological time series forecasting, *Acta Geophysica* **67**(5): 1471–1481.

Septiawan, W. and Endah, S. (2018). Suitable recurrent neural network for air quality prediction with backpropagation through time, *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)*, pp. 1–6.

Sivakumar, B., Wallender, W. W., Horwath, W. R. and Mitchell, J. P. (2007). Nonlinear deterministic analysis of air pollution dynamics in a rural and agricultural setting, *Advances in Complex Systems* **10**(04): 581–597.

Song, X., Huang, J. and Song, D. (2019). Air quality prediction based on lstm-kalman model, *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, pp. 695–699.

Therneau, T., Atkinson, B. and Ripley, B. (2014). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-8.
URL: *http://CRAN.R-project.org/package=rpart*

Wang, J. and Song, G. (2018). A deep spatial-temporal ensemble model for air quality prediction., *Neurocomputing* **314**: 198 – 206.

Ward, M. D. and Gleditsch, K. S. (2018). *Spatial regression models*, Vol. 155, Sage Publications.

Wen, C., Liu, S., Yao, X., ng, L., Li, X., Hu, Y. and Chi, T. (2019). A novel spatiotemporal convolutional long short-term neural network for air pollution prediction., *Science of the Total Environment* **654**: 1091 – 1099.

Whitcher, B. (2019). *waveslim: Basic Wavelet Routines for One-, Two- And Three-Dimensional Signal Processing*. R package version 1.7.5.1.
URL: *https://CRAN.R-project.org/package=waveslim*

You, W. and Lv, Z. (2018). Spillover effects of economic globalization on CO2 emissions: A spatial panel approach, *Energy Economics* **73**: 248 – 257.

Zhao, P. and Zettsu, K. (2019). Convolution recurrent neural networks based dynamic transboundary air pollution prediction., *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA), Big Data Analytics (ICBDA), 2019 IEEE 4th International Conference on* p. 410.

Zhao, X., Barber, S., Taylor, C. C. and Milan, Z. (2018). Classification tree methods for panel data using wavelet-transformed time series., *Computational Statistics and Data Analysis* **127**: 204 – 216.

Zhao, X., Barber, S., Taylor, C. C. and Milan, Z. (2021). Interval forecasts based on regression trees for streaming data, *Advances in Data Analysis and Classification* **15**(1): 5–36.