Boundary problem and data leakage: A caveat for wavelet-based forecasting

Preprint · December 2018

DOI: 10.13140/RG.2.2.28234.21446

Yuto Kajita
Deloitte Touche Tohmatsu LLC

JCER DISCUSSION PAPER

No.148

Boundary problem and data leakage: A caveat for wavelet-based forecasting

Ryo Hasumi

Yuto Kajita

December 2018

Japan Center for Economic Research

Boundary problem and data leakage: A caveat for wavelet-based forecasting

Ryo Hasumi* Yuto Kajita†

December 2018

Abstract

The application of machine learning to economics has drawn much attention in recent years. Forecasting economic data based on machine learning needs feature extraction to obtain better performance. In time series forecasting, researchers often use the wavelet transform to process time series data, and have reported that the combination of a neural network model with the wavelet transform improves the accuracy of the prediction. There are, however, many papers relating to wavelet-based forecasting that do not provide sufficient information on how the time-series data was processed. We show that inappropriate procedures for applying the wavelet decomposition to time series data easily lead to data leakage, which uses unobserved data and so its forecasting results would be of extremely high precision. We find that wavelet-based forecasting in which the time series data are processed appropriately cannot outperform even a naive prediction. Prediction performance based on wavelets is unreliable if the researcher does not specify the data processing method.

Keywords: Wavelet Transformation, Data Leakage, Boundary Problem

1 Introduction

It has often been reported in the literature that the application of the wavelet decomposition to time series data enhances the predictive power of a forecasting model. The early studies of Aussem and Murtagh (1997) and Aussem et al. (1998) show the compatibility of the wavelet transform with a recurrent neural network (RNN) since the wavelet transform decomposes the time series data into periodic components and trend and the RNN is suitable for handling the regularity of the signals. A number of papers have proposed combining neural network models and the wavelet decomposition, and have reported a reasonable accuracy (Pahasa and Theera-Umpon (2007), Minu et al. (2010), Hsieh et al. (2011), Ortega and Khashanah (2014), Jothimani et al. (2015), Yu et al. (2017), Bao et al. (2017)). The wavelet decomposition has also been applied to a factor-augmented model (Rua (2011)) and the GARCH model (Tan et al. (2010)), as well as the canonical ARIMA model (Fernandez (2008), Al Wadia et al. (2011), Kriechbaumer et al. (2014), Zhang et al. (2017)). In this way, many researchers have constructed forecasting models based on the wavelet transform.

Although many papers have used the wavelet decomposition to process time series data, the boundary problem involved here is often ignored. This boundary problem involves the variation of the wavelet coefficients near the end point of the transformation window with its shifts, and is caused by the assumption of circularity, or by padding with artificial data after the endpoint. The boundary problem is closely related to forecasting since it concerns the data yet to be observed, located beyond the boundary.

In this study, we show that disregarding the boundary problem and taking inappropriate procedures for employing the wavelet transform leads to a serious problem, especially in forecasting: data leakage. Data leakage is the mishandling of a model or data, in which information

*Corresponding author, Japan Center for Economic Research. E-mail: hasumi@jcer.or.jp

†Graduate School of Economics, Waseda University.

not observed in that period is used. In the presence of this problem, the obtained model and results are unrealistic and often seem to be unrealistically good. In fact, there are studies in which data leakage occurs unintentionally or intentionally and the results are seemingly biased in favor of the wavelet transform. Without getting rid of this mishandling, we cannot deny the possibility that the usefulness of the wavelet transform has been overstated or even does not exist.[1]

This paper is organized as follows. In Section 2, we describe a simple example of the wavelet transform and the cause of the boundary problem. In Section 3, we construct simple forecast models to clarify the data leakage associated with the wavelet transform. Section 4 is a discussion of how to manage the boundary problem. The last section is the Conclusion.

2 The wavelet transform and the boundary problem

The discrete wavelet transform (DWT) is an orthogonal transformation of time series data X by a discrete wavelet matrix $\mathcal{W}$. The resulting matrix $W = \mathcal{W}X$ is called the DWT coefficients. If we conduct the level-2 Daubechies(4) discrete wavelet transform (D(4) DWT), for which the length of X, denoted by N, must be a multiple of 4, we first define two kinds of orthogonal matrices

$$\underbrace{\mathcal{B}_J}_{N/2J\times N/J} = \begin{bmatrix} h_1 & h_0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & h_3 & h_2 \\ h_3 & h_2 & h_1 & h_0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 \\ & & & & & \vdots & & & & & \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & h_3 & h_2 & h_1 & h_0 \end{bmatrix}, \tag{1}$$

$$\underbrace{\mathcal{A}_J}_{N/2J\times N/J} = \begin{bmatrix} g_1 & g_0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & g_3 & g_2 \\ g_3 & g_2 & g_1 & g_0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 \\ & & & & & \vdots & & & & & \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & g_3 & g_2 & g_1 & g_0 \end{bmatrix}, \tag{2}$$

where

$$h_0 = \frac{1-\sqrt{3}}{4\sqrt{2}}, \quad h_1 = \frac{-3+\sqrt{3}}{4\sqrt{2}}, \quad h_2 = \frac{3+\sqrt{3}}{4\sqrt{2}}, \quad h_3 = \frac{-1-\sqrt{3}}{4\sqrt{2}},$$
$$g_0 = -h_3,\ g_1 = h_2,\ g_2 = -h_1,\ g_3 = h_0,$$

and then define

$$\mathcal{W} = \begin{bmatrix} \mathcal{B}_1 \\ \mathcal{B}_2\mathcal{A}_1 \\ \mathcal{A}_2\mathcal{A}_1 \end{bmatrix}. \tag{3}$$

Since $\mathcal{W}^\top\mathcal{W} = I$, we can decompose X as

$$X = \underbrace{\mathcal{B}_1^\top\mathcal{B}_1X}_{\mathcal{D}_1} + \underbrace{\mathcal{A}_1^\top\mathcal{B}_2^\top\mathcal{B}_2\mathcal{A}_1X}_{\mathcal{D}_2} + \underbrace{\mathcal{A}_1^\top\mathcal{A}_2^\top\mathcal{A}_2\mathcal{A}_1X}_{\mathcal{S}}. \tag{4}$$

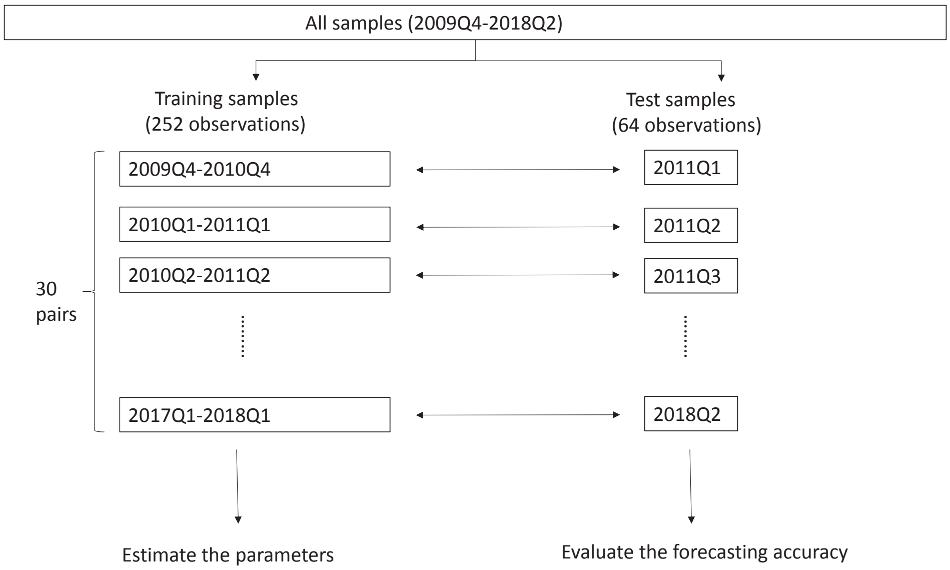
Here, $\mathcal{D}_j$ and $\mathcal{S}$ are called the wavelet details and smooth, respectively.

The circularity assumption appears in the last two elements of the first row of B_j, h_3, and h_2, which make a certain length of elements near the end of the wavelet details and smooth dependent on the beginning of the data X. For example, the last two elements of $\mathcal{D}_1$ are affected by the first two observations of X.

Under the circularity assumption, the data yet to be observed, $X(N+1), X(N+2), \dots$, are replaced by $X(1), X(2), \dots$. In case the circularity assumption is inappropriate for the reason

[1] As for deliberate data leakage, Herwartz and Schlüter (2017) refer to it as perfect foresight and Kriechbaumer et al. (2014) as calibration.

Figure 1: Description of data processing



that there exists a discrepancy between $X(N)$ and $X(1)$, for example, another assumption is often used: reflection, in which $X(N+1), X(N+2), \ldots$ are replaced by $X(N), X(N-1), \ldots$. Constant padding is also an option in which a constant $X(N)$ substitutes for $X(N+1), X(N+2), \ldots$.

As is apparent from this explanation, if we obtain new observations, which must be a multiple of 4 in this case, the elements near the end of the wavelet details $\mathcal{D}_j$ and smooth $\mathcal{S}$ that we already have will change, since $X(1), X(2), \ldots$ under circularity or $X(N), X(N-1), \ldots$ under reflection are replaced by the actual data $X(N+1), X(N+2), \ldots$. This drawback of the wavelet transform is referred to as the boundary problem. It occurs in all the wavelet transforms, including the maximal overlap discrete wavelet transform (MODWT), except for the most primitive Haar DWT, where the width of the wavelet is 2.

3 Wavelet-based forecasting and data leakage: A simple example

We now construct a wavelet-based forecasting model for predicting the S&P500 index. The sample data are the daily closing prices covering a period of seven years, from January 2010 to June 2018. We split the whole data into 30 subsets with the pairs of training sample $X^{(1)}$ and test sample $X^{(2)}$ as shown in Figure 1. The sample length of $X^{(1)}$ is 252 trading days before the beginning of each quarter (from 2011Q1 to 18Q2) and that of $X^{(2)}$ is 64 trading days from the beginning of the quarter.

For each subset, we firstly take the logarithm of the training and test sets by the D(4) DWT and then obtain two sets, wavelet details and smooth, $(\mathcal{D}_1^{(1)}, \mathcal{D}_2^{(1)}, \mathcal{S}^{(1)})$ and $(\mathcal{D}_1^{(2)}, \mathcal{D}_2^{(2)}, \mathcal{S}^{(2)})$. In the second step, we apply the AR(1) model to $\mathcal{D}_1^{(1)}$, $\mathcal{D}_2^{(1)}$ and the first differences of $\mathcal{S}^{(1)}$,

$$\mathcal{D}_1^{(1)}(t) = \alpha_1 + \beta_1 \mathcal{D}_1^{(1)}(t-1) + \varepsilon^{\mathcal{D}_1}(t), \tag{5}$$

$$\mathcal{D}_2^{(1)}(t) = \alpha_2 + \beta_2 \mathcal{D}_2^{(1)}(t-1) + \varepsilon^{\mathcal{D}_2}(t), \tag{6}$$

$$\Delta\mathcal{S}^{(1)}(t) = \alpha_3 + \beta_3 \mathcal{S}^{(1)}(t-1) + \varepsilon^{\mathcal{S}}(t), \tag{7}$$

and estimate the coefficients by OLS. In the third step, we perform out-of-sample predictions at each trading day by using the estimated models and the wavelet details and smooth of the test

Table 1: Summary of prediction performance (1)

method	MDA	RMSE	MAE	ARR
overall_DWT(circ.)_AR1	0.7302	0.0063	0.0048	0.3778
overall_DWT(circ.)_naive	0.7143	0.0064	0.0048	0.3545
dln(X)_AR1	0.5238	0.0074	0.0056	0.0146
dln(X)_naive	0.4841	0.0108	0.0083	-0.0059

Note: MDA (mean direction accuracy), RMSE (root mean squared error) and MAE (mean absolute error) are the medians across the 30 sets of training and test series. ARR is the median of the average return rate corresponding to each forecasting method, i.e., buy if a one-period-ahead forecast is positive and sell if one is negative.

set $(\mathcal{D}_1^{(2)}, \mathcal{D}_2^{(2)}, \mathcal{S}^{(2)})$:

$$\widehat{\mathcal{D}}_1^{(2)}(t) = \alpha_1 + \beta_1 \mathcal{D}_1^{(2)}(t-1), \tag{8}$$

$$\widehat{\mathcal{D}}_2^{(2)}(t) = \alpha_2 + \beta_2 \mathcal{D}_2^{(2)}(t-1), \tag{9}$$

$$\Delta\widehat{\mathcal{S}}^{(2)}(t) = \alpha_3 + \beta_3 \Delta\mathcal{S}^{(2)}(t-1). \tag{10}$$

By summing the predicted values of the wavelet details and smooth $(\widehat{\mathcal{D}}_1, \widehat{\mathcal{D}}_2, \widehat{\mathcal{S}})$, we have a prediction of the logarithm of the test series,

$$\ln(\widehat{X}^{(2)}(t)) = \widehat{\mathcal{D}}_1^{(2)}(t) + \widehat{\mathcal{D}}_2^{(2)}(t) + \widehat{\mathcal{S}}^{(2)}(t). \tag{11}$$

We call this forecasting method *overall_DWT(circ.)_AR1*.

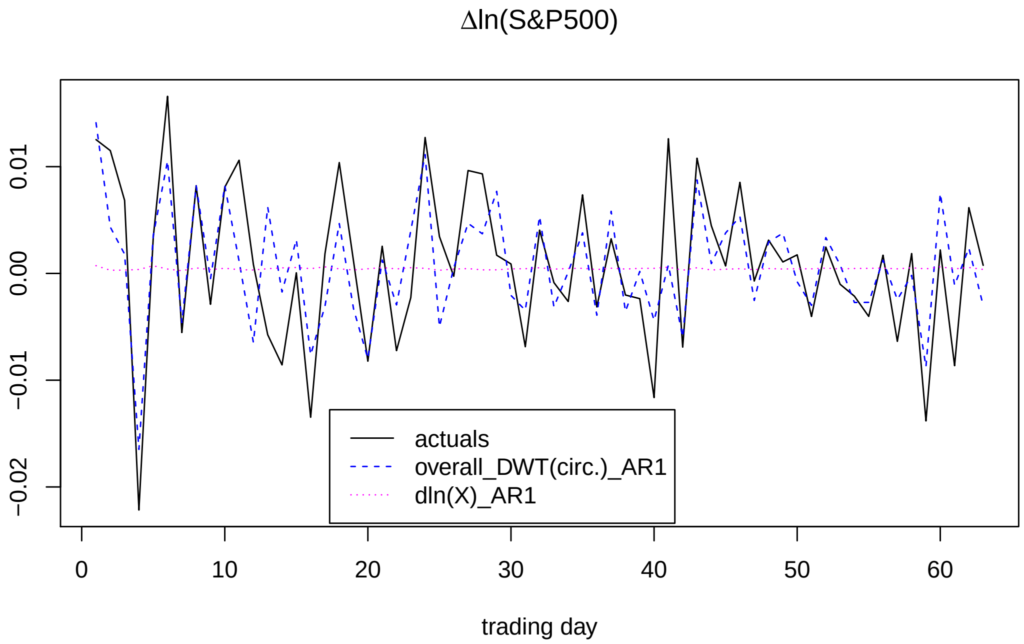
For comparison, we make a prediction assuming the one-period-ahead $\mathcal{D}_1$ to be zero ($\widehat{\mathcal{D}}_1^{(2)}(t+1) = 0$), and $\mathcal{D}_2$ and $\mathcal{S}$ keep the same level ($\widehat{\mathcal{D}}_2^{(2)}(t+1) = \mathcal{D}_t^{(2)}(t)$ and $\widehat{\mathcal{S}}^{(2)}(t+1) = \mathcal{S}_t^{(2)}(t)$) and name this model *overall_DWT(circ.)_naive*. We also estimate the AR(1) model directly applied to $\Delta\ln(X^{(1)})$ and make one-period-ahead predictions, which is denoted by *dln(X)_AR1*. *dln(X)_naive* assumes that the percentage change of a one-day-ahead is the same as that of today.

The prediction performance is evaluated by four performance measures: mean directional accuracy (MDA), root mean squared error (RMSE), mean absolute error (MAE), and average rate of return (ARR). We calculate ARR as the average of the returns from the results of the trading strategy based on the model's forecasting.

Table 1 shows a summary of the prediction performance of the models. As a whole, the precision of *overall_DWT(circ.)_AR1* and *overall_DWT(circ.)_naive* is much higher than the direct predictions, *dln(X)_AR1* and *dln(X)_naive*. The facts that $\mathcal{D}_1$ takes over high-frequency components and that the coefficient of the AR(1) is negative increase the accuracy of the AR(1) prediction compared to the naive prediction. Since $\mathcal{S}$ plays the role of a centered moving average, knowing its level greatly increases the prediction accuracy. The same logic applies to $\mathcal{D}_2$ since it takes over longer period components. Figure 2 depicts an example of predictions based on *overall_DWT(circ.)_AR1* and *dln(X)_AR1* during 2018: Q1 as well as its actual values. Apparently, this figure also confirms the extremely high accuracy of the prediction based on the DWT.

Although the procedures employing the wavelet transform seem to be quite sensible, the result is spurious and impossible to be applied in actual practice. The seemingly high accuracy is caused by the fact that the wavelet details and smooth, $(\mathcal{D}_1^{(2)}, \mathcal{D}_2^{(2)}, \mathcal{S}^{(2)})$, obtained by applying the level 2 D(4) DWT to the whole test series $X^{(2)}$ containing future information. More specifically, $\mathcal{D}_1(t)$, $\mathcal{D}_2(t)$, and $\mathcal{S}(t)$ depend on the data up to $X(t+3)$, $X(t+7)$, and $X(t+7)$, respectively, by the definition exhibited in Equation (4) and the values of the higher order wavelet details and smooth, $\mathcal{D}_2$ and $\mathcal{S}$, are adjusted in line with the trend, which makes it easy to guess a one-period-ahead value of the original series to be forecasted. This is typical data leakage: a

Figure 2: An example path of the log difference of the S&P 500 index and its predictions (2018:Q1)



Note: The solid black line is a test set of log differenced S&P 500 index from 04/02/2018 to 06/29/2018 (64 trading days). The dashed blue line is a one-period-ahead forecast based on the wavelet details and smooth converted from the whole test set and the application of the AR(1) model. The dotted pink line indicates those based on the AR(1) model directly applied to $\Delta \ln(X^{(1)})$.

forecast based on information not yet obtainable at that time. It is similar to the drawback of the Hodrick–Prescott filter shown by Hamilton (2017), that the detrended cyclical component obtained by the two-sided HP filter is highly predictable since it depends on the future error terms.

In the next section, we explain an appropriate procedure to make the out-of-sample prediction in wavelet-based forecasting and examine its forecasting accuracy.

4 How to manage the boundary problem in forecasting

That the wavelet details and smooth are affected by the data from future times is closely related to the boundary problem. If the wavelet transform is sequentially performed in decomposing the test data, as suggested in Aussem et al. (1998), we can avoid data leakage. Table 2 is a summary of the prediction results when the DWT is sequentially applied to the series of 64 trading days before the timing of the prediction, as follows:

Step 1 : Set the reference date τ and initialize $i = 0$.

Step 2 : Set a rolling window from $\tau - 63 + di$ to $\tau + di$ where d stand for the interval to employ the DWT.

Step 3 : Employ the DWT and store the wavelet details and smooth from $\tau + (i-1)d + 1$ to $\tau + di$.

Step 4 : Increase i by 1 and go back to Step 2 until the window goes beyond the end of the period of the test set.

Table 2: Summary of prediction performance (2)

method	MDA	RMSE	MAE	ARR
sequential_DWT(circ.)_AR1	0.4841	0.0384	0.0322	0.0110
sequential_DWT(circ.)_naive	0.4841	0.0168	0.0143	0.0065
sequential_DWT(ref.)_naive	0.5238	0.0085	0.0063	0.0130
sequential_DWT(con.)_naive	0.5238	0.0081	0.0060	0.0250

Note: See the note to Table 1.

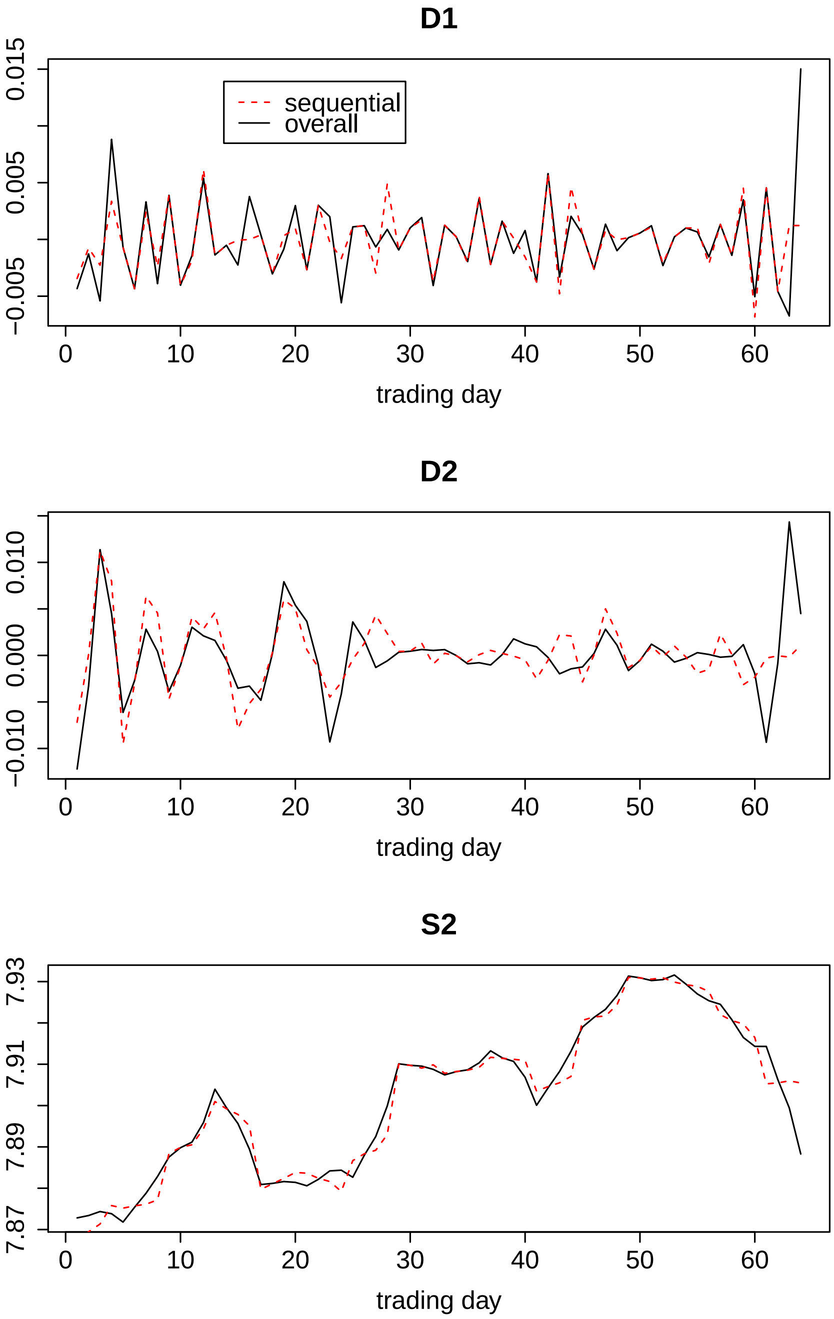
sequential_DWT(circ.)_AR1 uses the same AR(1) coeficients as *overall_DWT(circ.)_AR1*. *sequential_DWT(circ.)_naive*, *sequential_DWT(ref.)_naive* and *sequential_DWT(con.)_naive* are naive predictions employing a common DWT, but each assumes different boundary conditions: circular, reflection, and constant padding. The interval d is set to 1. This procedure ensures that the results are based on out-of-sample predictions.

It is clear that neither method in Table 2 explicitly outperforms *dln(X)_AR1* or *dln(X)_naive* shown in Table 1. *sequential_DWT(con.)_naive* may perform relatively well, but is much worse than *overall_DWT(circ.)_AR1* or *overall_DWT(circ.)_naive*. This exercise suggests that assuming reflection or constant padding instead of circularity may help to some extent, but not decisively.[2]

Figure 3 depicts example paths of $(\mathcal{D}_1^{(2)}, \mathcal{D}_2^{(2)}, \mathcal{S}^{(2)})$ obtained by the above procedure assuming reflection and setting the reference date τ to the beginning of the period of 2018Q2 and the interval $d = 4$ (the red dotted lines). For comparison, we also depict *overall_DWT(circ.)_naive* (the solid black lines) corresponding to the period. The difference between the two lines reflects the boundary problem in a broad sense as these are based on different information. It is worth noticing that one cannot obtain the black solid lines, which has much richer information, until the end of the period. Although the boundary problem occurs only near the beginning and end of the wavelet details and smooth, the above exercises have clarified that a researcher should not naively discard the unstable part near the end of the series and use only the stable part since their stability is the result of incorporating future information.

[2]For aliviating the boundary problem, Arino (1995) proposes padding based on an ARIMA model. Herwartz and Schlüter (2017) use future prices for padding as their objective of forecasting is foreign exchange rates.

Figure 3: Example paths of wavelet details and smooth extracted by two different methods (2018:Q1)



Note: The solid black lines are the wavelet details and smooth obtained by transforming a whole test set beginning from 04/02/2018. The dashed red lines are those obtained by sequentially conducting the DWT assuming reflection, in which the number of samples is increased by 4 while the sample size is restricted to 64, and the last 4 values of each calculation are stored.

5 Conclusion

In this study, we have shown that there is a close relationship between forecasting through the wavelet transform and the boundary problem this involves. How to decompose the test data into the wavelet details and smooth is not unique, especially in relation to the boundary problem, and is important for ensuring reproducibility in a practical environment. As discussed in the previous section, the wavelet transform should be applied repeatedly at each time of prediction: otherwise, data leakage may occur. Although the above example of data leakage concerns a misuse of the DWT, it is also the case with the MODWT, where the length of the data to be transformed is not restricted to be a multiple of 2^J. Despite its being called "perfect foresight," such prediction is often not perfect but of good accuracy, which makes the problem difficult to be noticed.

References

Al Wadia, S., Mohd Tahir Ismailb, M. H. Alkhahazaleh, and Samsul Ariffin Abdul Karim (2011) 'Selecting wavelet transforms model in forecasting financial time series data based on arima model.' *Applied Mathematical Sciences* 5(7), 315–326

Arino, Miguel A. (1995) 'Time series forecasts via wavelets: an application to car sales in the spanish market'

Aussem, Alex, and Fionn Murtagh (1997) 'Combining neural network forecasts on wavelet-transformed time series.' *Connection Science* 9(1), 113–122

Aussem, Alex, Jonathan Campbell, and Fionn Murtagh (1998) 'Wavelet-based feature extraction and decomposition strategies for financial forecasting.' *Journal of Computational Intelligence in Finance* 6(1), 5–12

Bao, Wei, Jun Yue, and Yulei Rao (2017) 'A deep learning framework for financial time series using stacked autoencoders and long-short term memory.' *PloS One* 12(7), e0180944

Fernandez, Viviana (2008) 'Traditional versus novel forecasting techniques: how much do we gain?' *Journal of Forecasting* 27(7), 637–648

Hamilton, James D. (2017) 'Why you should never use the hodrick-prescott filter.' National Bureau of Economic Research Working Paper 23429

Herwartz, Helmut, and Stephan Schlüter (2017) 'On the predictive information of futures' prices: A wavelet-based assessment.' *Journal of Forecasting* 36(4), 345–356

Hsieh, Tsung-Jung, Hsiao-Fen Hsiao, and Wei-Chang Yeh (2011) 'Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm.' *Applied Soft Computing* 11(2), 2510–2525

Jothimani, Dhanya, Ravi Shankar, and Surendra S. Yadav (2015) 'Discrete wavelet transform-based prediction of stock index: a study on national stock exchange fifty index.' *Journal of Financial Management and Analysis* 28(2), 35–49

Kriechbaumer, Thomas, Andrew Angus, David Parsons, and Monica Rivas Casado (2014) 'An improved wavelet–arima approach for forecasting metal prices.' *Resources Policy* 39, 32–41

Minu, K.K., M.C. Lineesh, and C. Jessy John (2010) 'Wavelet neural networks for nonlinear time series analysis.' *Applied Mathematical Sciences* 4(50), 2485–2495

Ortega, Luis, and Khaldoun Khashanah (2014) 'A neuro-wavelet model for the short-term forecasting of high-frequency time series of stock returns.' *Journal of Forecasting* 33(2), 134–146

Pahasa, Jonglak, and Nipon Theera-Umpon (2007) 'Short-term load forecasting using wavelet transform and support vector machines.' In 'Power Engineering Conference, 2007. IPEC 2007. International' IEEE pp. 47–52

Rua, António (2011) 'A wavelet approach for factor-augmented forecasting.' *Journal of Forecasting* 30(7), 666–678

Tan, Zhongfu, Jinliang Zhang, Jianhui Wang, and Jun Xu (2010) 'Day-ahead electricity price forecasting using wavelet transform combined with arima and garch models.' *Applied Energy* 87(11), 3606–3610

Yu, Lean, Yang Zhao, and Ling Tang (2017) 'Ensemble forecasting for complex time series using sparse representation and neural networks.' *Journal of Forecasting* 36(2), 122–138

Zhang, Keyi, Ramazan Gençay, and M. Ege Yazgan (2017) 'Application of wavelet decomposition in time-series forecasting.' *Economics Letters* 158, 41–46