JAPAN EXCHANGE GROUP

JPX WORKING PAPER

Analysis of Investors' Behavior through Non-Time Series Analysis of Stock Prices

Kenichi YOSHIDA

April 1, 2019

Vol. 28

Note

This material was compiled based on the results of research and studies by directors, officers, and/or employees of Japan Exchange Group, Inc., its subsidiaries, and affiliates (hereafter collectively the "JPX group") with the intention of seeking comments from a wide range of persons from academia, research institutions, and market users. The views and opinions in this material are the writer's own and do not constitute the official view of the JPX group. This material was prepared solely for the purpose of providing information, and was not intended to solicit investment or recommend specific issues or securities companies. The JPX group shall not be responsible or liable for any damages or losses arising from use of this material. This English translation is intended for reference purposes only. In cases where any differences occur between the English version and its Japanese original, the Japanese version shall prevail. This translation is subject to change without notice. The JPX group shall accept no responsibility or liability for damages or losses caused by any error, inaccuracy, misunderstanding, or changes with regard to this translation.

Analysis of Investors' Behavior through Non-Time Series Analysis of Stock Prices

Kenichi YOSHIDA *

April 1, 2019

Abstract

After the war began with the special demand economy in the early 1950's, Japan was followed by a reactionary recession and repeated business cycles such as the "izanami economy" and the "Global Financial Crisis." In this study, we attempt to analyze the investors' behavior behind these business cycles by analyzing TOPIX price movements from 1954 to 2016 using simple data mining methods.

We use the gradient boosting decision tree (GBDT) as the data mining method to analyze TOPIX price movements; this method can analyze mixed distribution. Furthermore, as a method to detect the change in investors' behavior, we optimize the analyzing period of the data mining process by considering only the fluctuation in the monthly TOPIX price and the standard deviation of the daily price within the month.

The academic contributions of this paper are 1) showing the existence of a long-term investors' behavior over the business cycle; 2) pointing out the existence of anomaly where stock prices can be predicted in multiple markets; and 3) pointing out the importance of the non-time-series analysis of stock prices.

1) The relationship between the identified training period and the business cycles studied by previous researches, and 2) whether the prediction ability found this time will continue, are left as future research issues.

*Graduate School of Business Science, University of Tsukuba, yoshida.kenichi.ka@u.tsukuba.ac.jp

Contents

List of Tables

List of Figures

1 Introduction

After the war began with the special demand economy in the early 1950's, Japan was followed by a reactionary recession and repeated business cycles such as the "izanami economy" and the "Global Financial Crisis." In this study, we attempt to analyze the investors' behavior behind these business cycles by analyzing TOPIX price movements from 1954 to 2016 using simple data mining methods.

We use the gradient boosting decision tree (GBDT) as the data mining method to analyze TOPIX price movements; this method can analyze mixed distribution. Furthermore, as a method to detect the change in investors' behavior, we optimize the analyzing period of the data mining process by considering only the fluctuation in the monthly TOPIX price and the standard deviation of the daily price within the month. The characteristics of this simple analysis are: 1) non-time series data representation; 2) use of method that can handle mixed distribution; and 3) optimization of the training period. Although this analysis uses only past price information, the experimental results demonstrate its ability to stably predict the future price of representative indexes.

Here, an important byproduct is the training period used to create the prediction model. This period suggests the existence of a long-term investors' behavior over the business cycle based on the business standard date that the Cabinet Office, Government of Japan defines [1]. Therefore, the academic contributions of this paper are 1) showing the existence of a long-term investors' behavior over the business cycle; 2) pointing out the existence of anomaly where stock prices can be predicted in multiple markets; and 3) pointing out the importance of non-time-series analysis of stock prices.

The remainder of this paper is organized as follows. Section 2 presents a survey of existing approaches and analyzes their limitations in clarifying the characteristics of the method used in this research. Section 3 describes the data representation and method for analysis using the experimental results, and Section 4 shows 1) the existence of a long-term investors' behavior over the business cycle, and 2) an analysis of why the data mining method used in this study has not been reported to date. Finally, Section 5 summarizes our findings.

2 Related Works

2.1 Business Cycles in Japan

The Cabinet Office, Government of Japan defines the business cycle by considering major economic indicators [1]. Ogawa and Kitasaka quantitatively analyzed the formation and collapse of Japan's economic bubble from late 1980 to 1990s [2], and Miyao analyzed inter-dependencies among macro variables such as supply and demand gap, productivity, inflation rate, and stock price, using a vector auto-regression (VAR) approach [3]. These studies analyze the characteristics of the business cycle based on the major economic indicators.

In this paper, we quantitatively analyze the characteristics of investors' behavior over the business cycle based on the stock price information only. Although our approach cannot analyze the movement of the entire economy, it can be characterized by the analysis focusing on the stock price.

Watanabe [4] and Otsuka [5] use the Markov Switching Model. Watanabe quantitatively analyzed the structural change in business cycles using the Indexes of Business Condition defined by The Cabinet Office, Government of Japan [4]. Otsuka carried out a quantitative analysis of the regional business cycle and the spatial interaction among the regions using the regional indices of industrial production (IIP), published by the Ministry of Economy, Trade and Industry [5].

Their Markov switching models have variables that represent structure of economy. Our approach does not have such variables. However, the optimization of the training period and the generation of independent models on each training period enable us to analyze the investors' behavioral changes in our study.

2.2 Stock price analysis

In our study, prediction of future stock price is the key of the investors' behavior analysis. We use a simple prediction method for stock prices. Here, the widely accepted efficient market hypothesis (EMH [6]) entails the unpredictability of future stock prices. Although several studies have challenged this hypothesis, to date no one has reported a predictive model that yields a stable return over 16 years from price information alone. Time-series analysis using, for instance, auto regression (AR) and moving average (MA) [7], is the standard method used for analyzing market data. The results of a survey used to develop the present study found that the EMH, to which the unpredictability of future stock prices is foundational, is widely accepted and that there is no currently accepted method for predicting future prices solely from past prices (for details on the survey, see [8, 9]).

However, some researchers have challenged the underlying EMH assumption that it is difficult to predict future prices (see, e.g., [10, 11, 12, 13, 14, 15, 16, 17]). For example, [10] and [12] analyzed the relationship between individual stock returns and order imbalance using, respectively, extracted prediction rules and performed regression modeling. However, to date no one has reported a predictive model that can yield a stable return over 16 years based solely on price information.

Risk-based portfolios that consider the difficulty of predicting future stock prices, such as minimum variable (MV), risk parity (RP), and maximum dispersion (MD), are attracting attention from practitioners and have been shown (e.g., [18]) to outperform traditional portfolios that attempt to predict future prices [19].

Information symmetry is an important assumption of EMH, and some researchers are attempting to exploit this asymmetry to challenge EMH using social network services (SNS) and news data (e.g., [20, 21, 22, 23, 24]). Such studies seek to analyze the processing of information through SNS and news feeds to find situations in which information asymmetry is present.

Past research [25] shows that simple analysis of high-frequency trading order book information can be used to classify short-term stock price fluctuations with an accuracy of 82.9%, although the method was found to be inapplicable for monthly stock price fluctuations.

In this study, we demonstrate a simple method for analyzing prices that can be used to predict future stock prices. Back-testing of this method against representative indexes over a window of 16 years demonstrates the validity of the methodology, i.e., non-time-series analysis of stock prices.

The important byproduct of this method is the training period used to create the prediction model. The use of the method that has long stable predictability enables us analyzing investors' behavior from 1954 to 2016 through found periods. These periods seem to suggest the existence of a long-term investors' behavior over the business cycle.

3 Non-Time Series Analysis

This section explains the data mining method used in this study with the experimental results obtained using TOPIX data. Further discussion based on additional experimental results is then presented in Section 4.

3.1 Data Representation

We use the following simple vector representation, v_t, of the data:

$$
\begin{aligned}
p_{t,d} &= \text{Price of stock on d-th day of month } t \\
p_t &= \text{Price of stock on last day of month } t \\
d_t &= (p_t - p_{t-1})/p_{t-1} \\
d_{2,t} &= (d_t + 1) * (d_{t-1} + 1) \\
d_{3,t} &= (d_t + 1) * (d_{t-1} + 1) * (d_{t-2} + 1) \\
s_t &= \sqrt{\sum_{d=1}^{\text{last day of month } t} \frac{(p_{t,d}/p_{t-1} - 1)^2}{\text{number of days of month } t}} \\
v_t &= (d_t, s_t, \\
&\quad d_{t-1}, d_{2,t}, s_{t-1}) \\
&\quad d_{t-2}, d_{3,t}, s_{t-2})
\end{aligned}
$$

where p_t is the stock price of month t, d_t is the price fluctuation between the current and the previous month, s_t is the standard deviation of the daily stock price in month t, and $d_{n,t}$ is the price fluctuation over n months.

Although this vector representation is simple and naive, it is quite different from the standard method of data representation used in conventional time series analysis (see Figure 1). Conventional methods apply linear differential equations such as

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

to represent the relationship between data such as y_t and y_{t-p} or $_q$ based on the assumption that the continuous states from time $t-p$ or q to $t-1$ contain information that can be used to predict states at time t. By tuning ϕ, θ, and C, the linear differential equation can be applied as a prediction model of y_t from continuous past states.

The vector representation used in this study diverges from conventional approaches in that it assumes that discrete time points contain the necessary predictive information and that this information can be obtained using data mining methods. Although this method uses a data structure similar to those used in conventional time series analysis, its novel modeling mechanism enables the use of discrete time-point information.

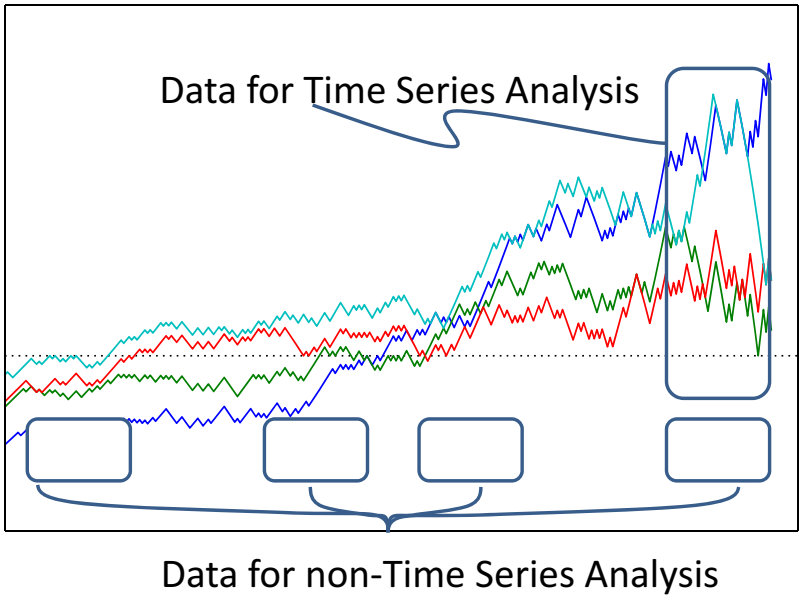


Figure 1: Data Representation

3.2 Algorithm

Algorithm 1 estimates a future stock price fluctuation d_{t+1} based on previous stock price information, v_t, by selecting the training period length i (months) that produces a maximum correlation between past data d_t and estimated data e_t of length *test* (line 11), where *test* is a tuning parameter that specifies the length of the test period used to calculate correlation. Based on the results of initial testing, we eventually selected a value of $test = 12$. We also set the maximum training period i to 300 under the assumption that 25 years is a sufficient time window for analyzing the investors' behavior. Using previous stock price information from time v_{t-i} to v_t, the algorithm constructs a prediction model m (line 15), and finally, it returns $m(v_t)$ as its prediction results (line 16).

The value of i that gives a maximum correlation between d_t and e_t is found by the SELECT function, which iterates i from one up to its maximum value, constructing models m up to the maximum length for each iterated value of i (line 7) from which estimates of the price fluctuation e_i are obtained (line 8). Here, the variable j ($1 \leq j \leq test$, line 6) specifies the time of analysis using the training data of length i. Finally, the value of i corresponding to the maximum correlation between the actual data d_t of length *test* and the estimated data e_t is selected.

Note that this algorithm uses the data for period $i + 1$ to produce the final model m (line 15). Because the investors' behavior is evolving, data that are too old should not be used for constructing the estimate e of future d; however, it can be assumed that their behavior is evolving slowly enough to justify using the relationship in the current month to project the relationship in the succeeding month; in other words, if i is optimal for estimating d_t, $i + 1$ will be close to optimal for estimating d_{t+1}.

Algorithm 1 Non-Time Series Analysis

```
function MODEL(d, v)
    return Model for estimating d_t from v_{t-1}
end function
▷
function SELECT(d, v)
    for i in [1:length(d)-test] do
        for j in [1:test] do
            m = MODEL(d[data of length i for position j],
                      v[data of length i for position j])
            e[i, j] = m(v[for j])
        end for
    end for
    return i that gives max correlation
                between d[t-test+1:t] and e[i,1:test]
end function
▷
function ESTIMATE(d, v)
    i = SELECT(d, v)
    m = MODEL(d[last part of length (i+1)],
                  v[last part of length (i+1)])
    return m(v_t)                          ▷ i.e., return estimation of d_{t+1}
end function
```

3.3 Experimental results on TOPIX

Our experimental investment strategy is simple: we buy if the estimated e_{t+1} is positive and sell if e_{t+1} is negative (See Algorithm 2). Figure 2 shows the results obtained by applying the proposed method to TOPIX data acquired from a Bloomberg terminal. The top graph, in which the X-axis is the month and the Y-axis is the selected period, shows the selected training period i. The graph directly below shows the correlation between the real data $d_{t:t+11}$ and the estimated results $e_{t:t+11}$. To calculate the correlation at time t, data from t to $t + 11$ were used, as described above. The third graph down shows the accuracy, where the accuracy at t is obtained using data from t to $t + 11$, and the bottom graph shows the results of investment.

Algorithm 2 Investment strategy

```
procedure INVESTMENT STEP(d, v)
    e_{t+1} = ESTIMATE(d, v)
    if e_{t+1} > 0 then
        Buy
    else
        Sell
    end if
end procedure
```

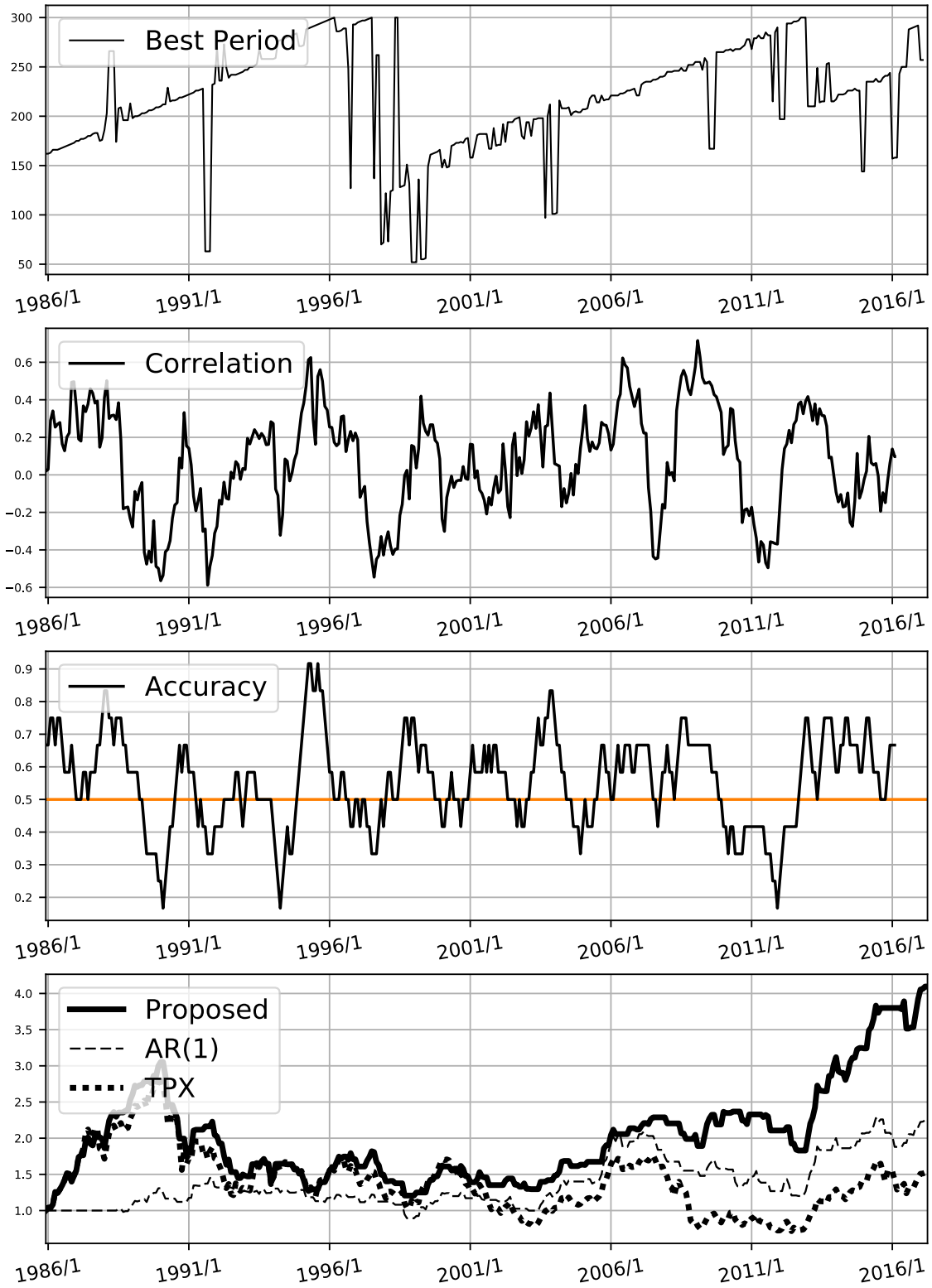


Figure 2: Results obtained using TOPIX

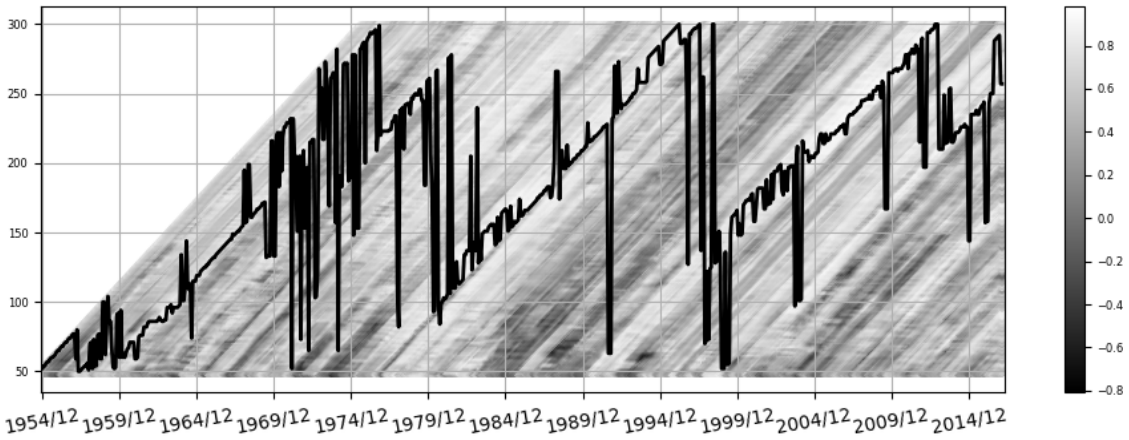


Figure 3: Changing the Optimal Training Period

Here, a gradient boosting decision tree (GBDT)[1] is applied as the learner, with d_t, the price fluctuation between the current and previous month, and s_t, the standard deviation, used as the elements to construct vector v_t:

$$v_t = (d_t, s_t)$$

Figure 2 compares the investment efficiency of the proposed method (solid line) with the results obtained using index investment (dotted line) and AR(1) (dashed line). Index investment and AR modeling were selected for comparison here as representative standard investment and time series analysis methods, respectively. It is seen from the figure that the proposed method shows clear profits. The fact that AR(1) also realizes some profits suggests the presence of so-called anomalies [26] in the TOPIX data, which will be discussed later (see Section 4). Nevertheless, Figure 2 indicates the clear advantage of the proposed method over the other methods.

Although the results in Figure 2 indicate an unstable correlation between the real data d_t and the estimated results e_t, the selected training period i reflects the change in the investors' behavior. The correlations calculated during the estimation process (line 11 in Algorithm 1) and shown in Figure 3, in which the X- and Y-axes are the month and length of training period i respectively, suggest this change more clearly. In the figure, the black and white shading represent regions of high and low correlation, while the solid line shows the selected training period i; the diagonally aligned numerical values clearly suggest a changing investors' behavior. Section 4.1 gives further discussion on this changing investors' behavior.

Our simplified model representation, which uses only d_t and s_t, can be displayed graphically (see Figure 4) by plotting d_t on the X-axis and s_t on the Y-axis. The dotted regions indicate where the learned model indicates the price goes down in the next month, while the other regions indicate prices going up. Although Figure 4 is only a retrofitted interpretation of the data, it can be interpreted as follows:

- When prices crash, the price in the next month will go down.
- In the adjustment phase, an increase in price will go down in the next month. A decrease in price will go up in the next month.

[1]In particular, the lightgbm package in python (https://github.com/Microsoft/LightGBM/) was used in the experiments.

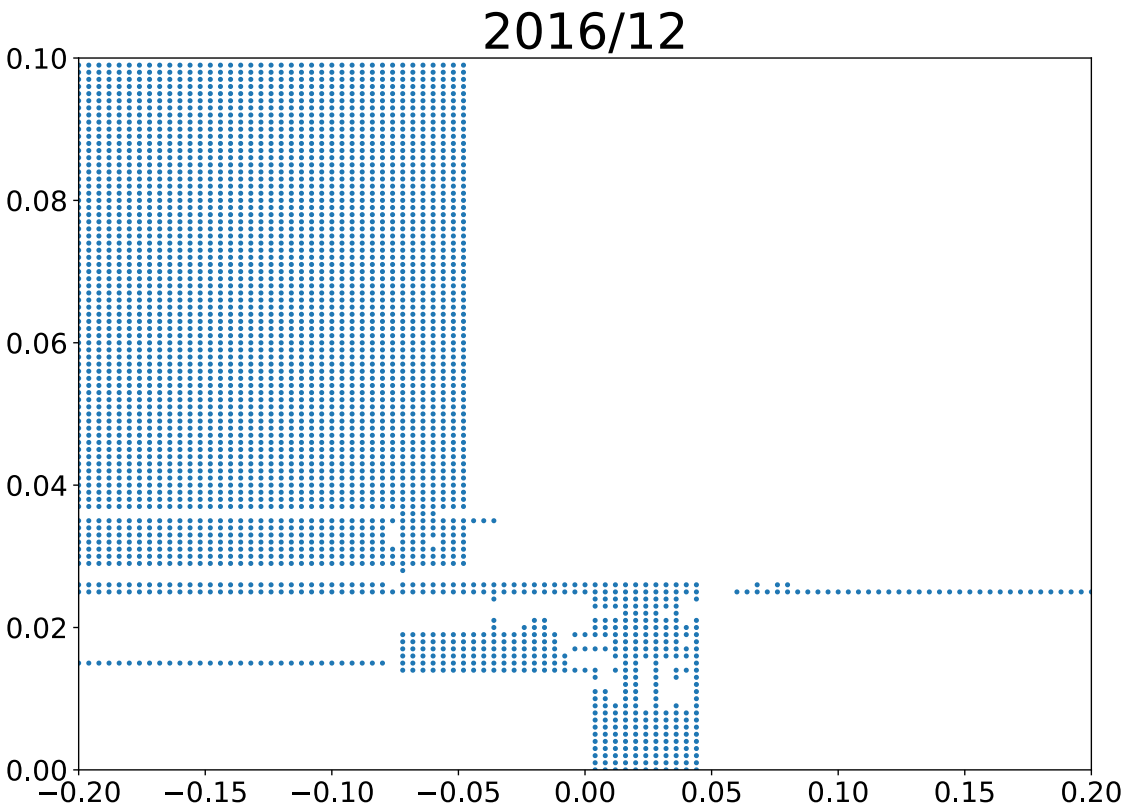


Figure 4: Decision Surface of GBDT

- Otherwise, the price will go up.

where the terms "price crash" and "adjustment phase" are defined numerically as follows:

$$\begin{array}{ll} \text{price crash:} & d_t < -0.05 \text{ and } s_t > 0.029 \\ \text{adjustment phase:} & |d_t| < 0.05 \text{ and } s_t < 0.029 \end{array}$$

In the application to past TOPIX data, the accuracy of this model is about 60%. This relatively high accuracy seems to be the cause of the high return shown in Figure2.

Note that an out-of-sample testing framework is used to produce the results shown in Figure 2; in other words, we always use past data to generate a model for predicting future prices. However, intermediate models generated in this process have always shown similar results if interpreted using this retrofitting method. Figure 5 shows the estimations produced by generated models at selected sample points. While the initial models appear to use only d_t, later models give indications of refinement by s_t. The boundaries 0.05 and 0.029 for d_t and s_t, respectively, appear to be stable.

Other important implications can be drawn from Figure 4, including:

- It indicates a mixed distribution underlying the data. We interpret Figure 4 as follows: the data in the "adjustment phase" area are generated by different distributions from other areas. Although the GBDT can model the mixed distribution behind the data, the regularity is weak and cannot be found unless an appropriate method is used (see Section 4 for further discussion).

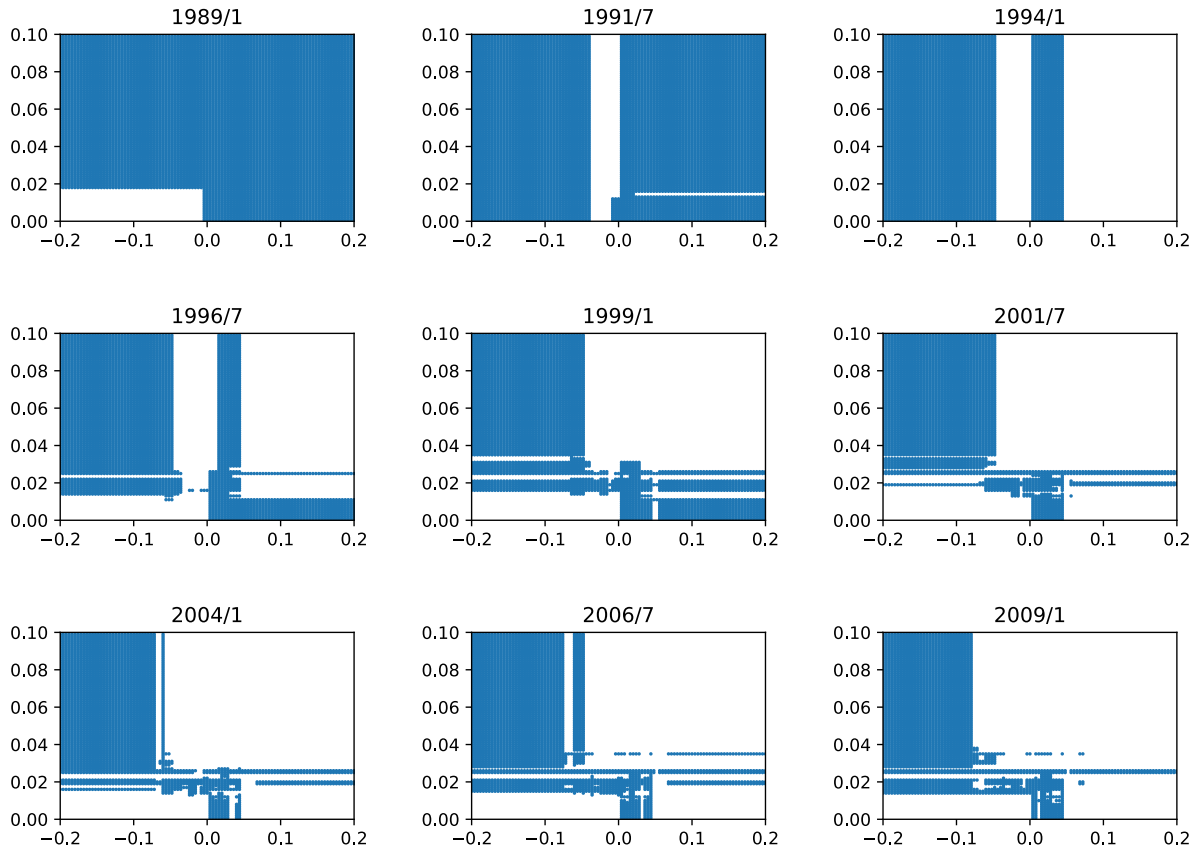


Figure 5: Changing the Decision Surface

- The linear differential equations used in conventional time-series analysis is not appropriate for these data. The regularity discussed so far requires the analysis of discrete time points, and we contend that linear differential relations are not well suited to this task. For example, the simple rule "**If** $d_t < -0.05$ **and** $s_t > 0.029$ **then** price will go down" cannot be simply handled by a linear differential equation. The non-time series analysis used in this paper seems to be more appropriate for handling this type of regularity.

4 Discussion

Experimental results reported above suggested three important research questions: 1) if we can apply this method to analyze long term investors' behavior behind business cycles, and 2) why has this simple method not been previously reported, and 3) can we apply this method to other indexes? This section answers these questions through further experimental assessments.

4.1 Analysis of Investors' Behavior

After the war began with the special demand economy in the early 1950's, Japan was followed by a reactionary recession and repeated business cycles such as the "izanami

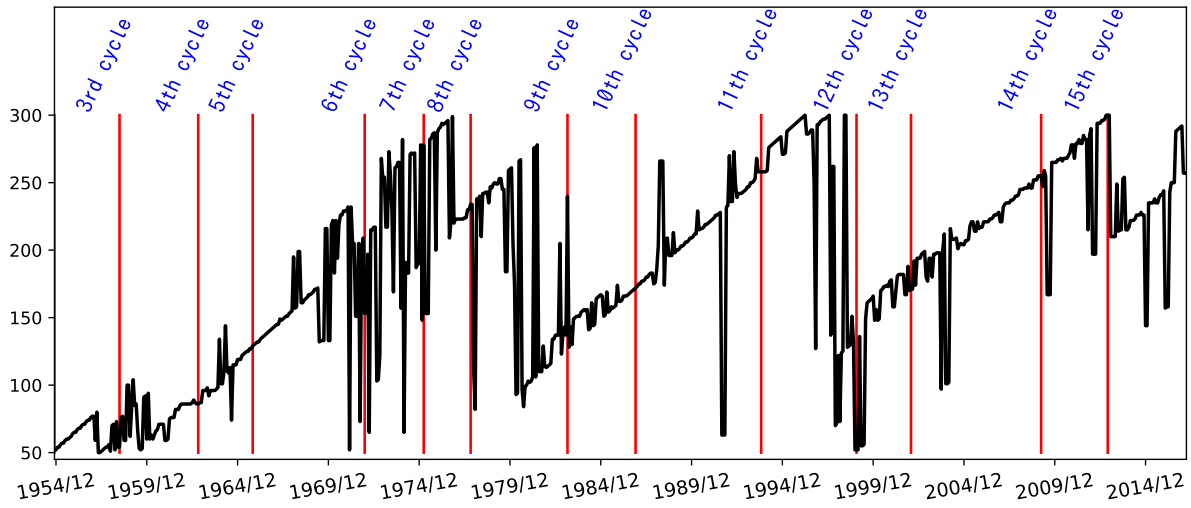


Figure 6: Change of Investors' Behavior

economy" and the "the Global Financial Crisis." The Cabinet Office, Government of Japan defines the business cycle by considering major economic indicators [1].

Figure 6 shows the business cycle defined by the Cabinet Office with the training period shown in Figure 2. Although the business cycle defined by the Cabinet Office directly reflects up and down movement of economic indicators, i.e., stock prices, and has an average length of 4–5 years, the training period is longer than the business cycle. This difference in the period length seems to indicate a fact that the behavior of investors does not change at every limit of the economic indicator. They seem to change their behavior after several fluctuations of economic indicators.

As shown in Figure 8 (See later), prior to 1990, the Sharpe ratio of the proposed method is worse than that of index investment. After 1990, the results are reversed, and the method used in this study outperforms index investment. Although our results, such as those shown in Figure 2, only show the existence of an anomaly after 1990, Figure 6 shows the existence of stable training periods even before 1990 where EMH does hold. This seems to show that the proposed method is effective for analyzing the investors' behavior even in a period when the method does not have prediction ability.

The relationship between this training period and the business cycles studied by previous research, e.g.,[1, 2, 3, 4, 5], is left as future research issues.

4.2 Why this method has not been reported

To answer the second question, we checked to see if the data mining method used in this study could earn profit on past data. Figure 7 shows the results. We found that, in contradiction to the results in the previous section, the method does not work on historical TOPIX data from 1954 to 1984. Although the obtained i suggests the beginning of the bubble economy of the 1980s (see the value of i shown in the top of Figure 7), the proposed method does not produce a profit when using these historical data. We conjecture that this might reflect a difference between the current investors' behavior and that of the late 20th century.

To clarify the above results, we compared the Sharpe ratio produced by the proposed method [27] from 1954 to 2016 (see Figure 8) with that produced by standard index

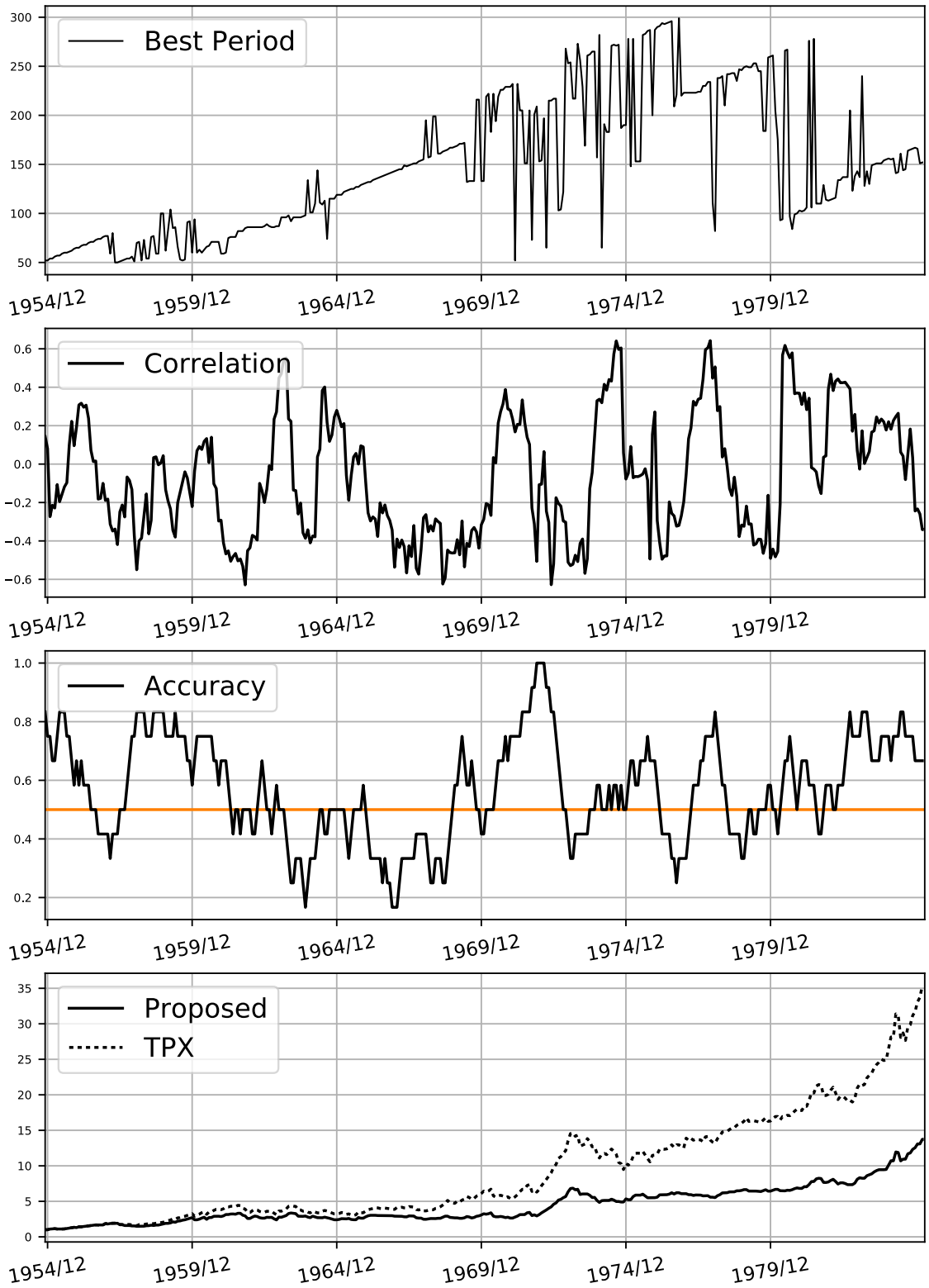


Figure 7: Results on Historical Data

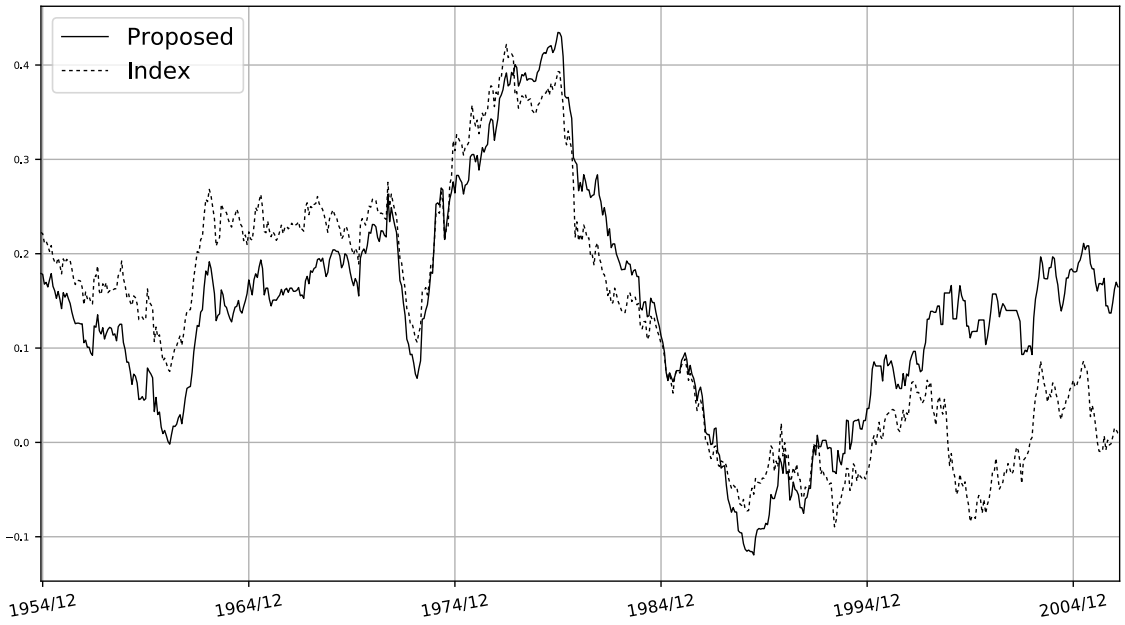


Figure 8: Change in the Sharpe Ratio

investment over the same time period. As it is more commonly applied than correlation or accuracy, the Sharpe ratio was chosen as a representative index to evaluate the performance of investment. To calculate the Sharpe ratio at time point t, we used data from time t to time $t + 119$, i.e., data over a 10-year window, and assumed a risk-free rate of zero:

$$\text{Sharpe Ratio} = \frac{\overline{R} - \text{risk-free rate}}{\sigma}$$

where

$$\begin{aligned} r_i &= \text{return at } i \\ \overline{R} &= \frac{1}{120}\sum_{i=t}^{t+119} r_i \\ \sigma &= \sqrt{\sum_{i=t}^{t+119}(r_i - \overline{R})^2} \\ \text{risk-free rate} &= 0 \end{aligned}$$

It is seen from the results in Figure 8 that, prior to 1990, the Sharpe ratio of the proposed method is worse than that of index investment; however, after 1990, the results are reversed and the proposed method outperforms the index investment.

On the other hand, there is no major change in the representative statistical indicators over this time period. For example, Figure 9 shows the relationship between the current (d_t) and next-month (d_{t+1}) price fluctuations (X- and Y-axes, respectively), which indicates that there is no correlation between the two values even after 1986. Note that Figure 4 suggests the existence of a mixed distribution underlying the data. As correlation cannot be simply calculated for mixed-distribution data, inappropriate

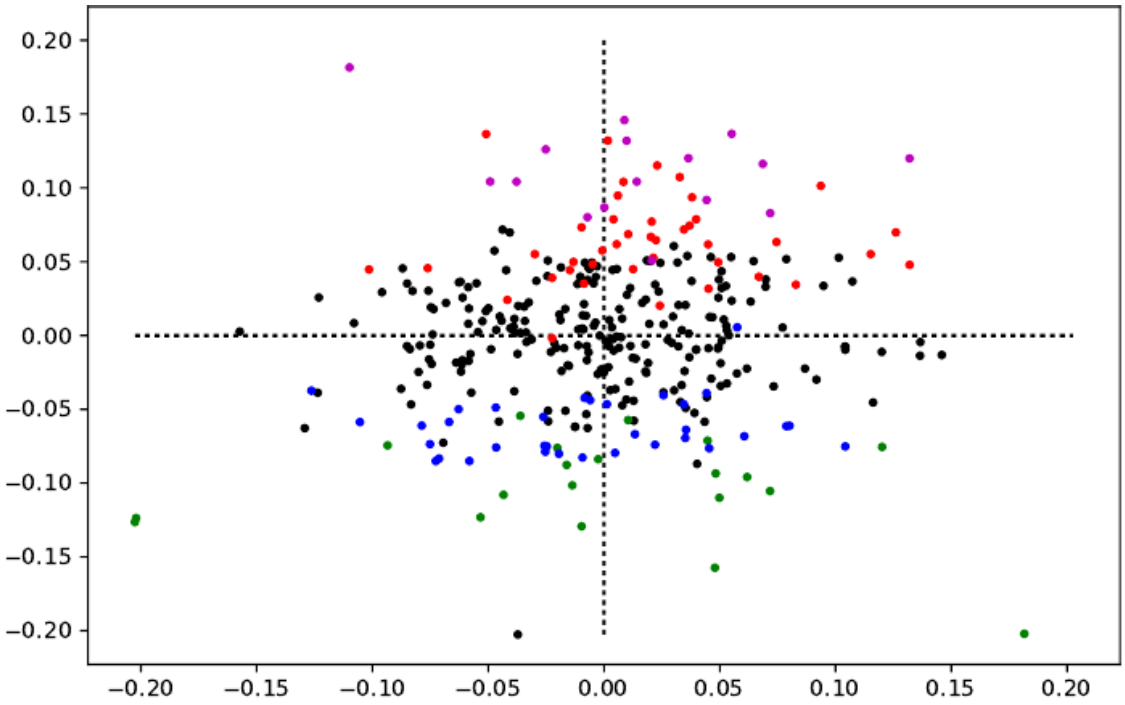


Figure 9: Monthly TOPIX Movement (1986~2016)

results might have misled previous researchers and prevented them from noticing this change in the investors' behavior.

The mixed distribution in the historical data might have had other inhibiting effects on prior research. Figure 10 shows the return obtained using support vector regression (SVR) without the use of a GBDT. Although SVR is frequently used for regression analysis, it is not useful in the analysis of TOPIX data.

At first glance, decision tree learning results also appear to be unenlightening. Figure 11 shows the tree generated by a classification and regression tree (CART) learning system. Although the tree can be used to produce profit when used for prediction,[2] its large number of nodes might serve to hide its more useful results.

The last hypothesis is that the factors used are too simple. Although the method to analyze data, e.g., time series analysis, is important to analyze stock prices, the factor used for the analysis is also important and various factors have been studied, .e.g., [28], [29], [30], [31]. Among them, it is worth examining the relationship between the data used in the proposed method and the VIX index [29]. The VIX index is a 30-day expected volatility, i.e., standard deviation of returns, of the U.S. stock market. This causes a suspicion that the proposed method is only a reconstruction of already known methods.

In fact, the factors used in the proposed method are not new. We found 1) the advantage of a method that can handle mixed-distribution, and 2) the effect of data period tuning. As shown in Figure 12, naive use of the VIX index alone cannot make any investment strategy. The typical way of the VIX index usage is

- Sell if $\mathrm{VIX}_{max} < \mathrm{VIX}$

[2]We omit the experimental results of investment by CART because they deviate from main argument.

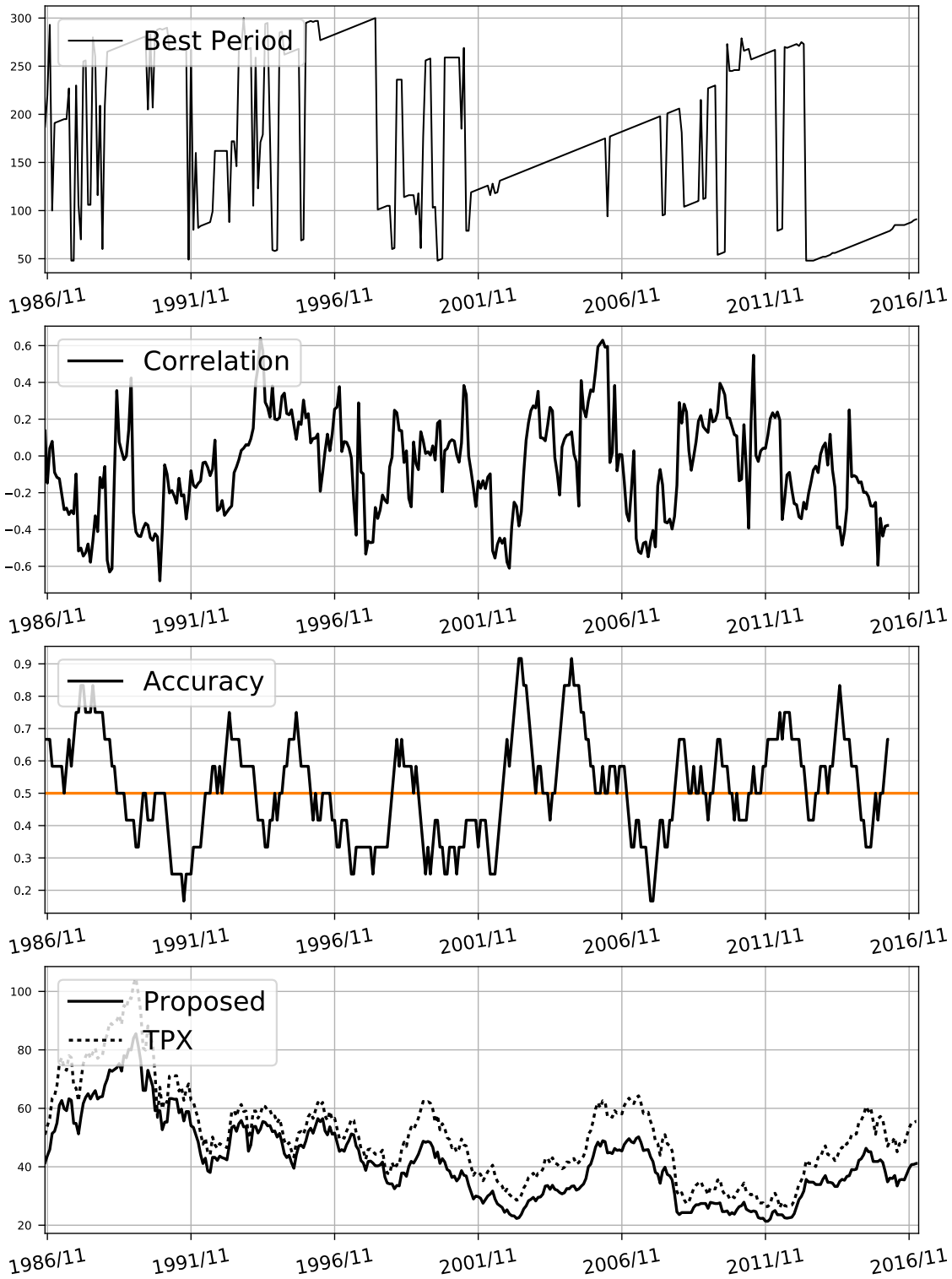


Figure 10: SVR Results

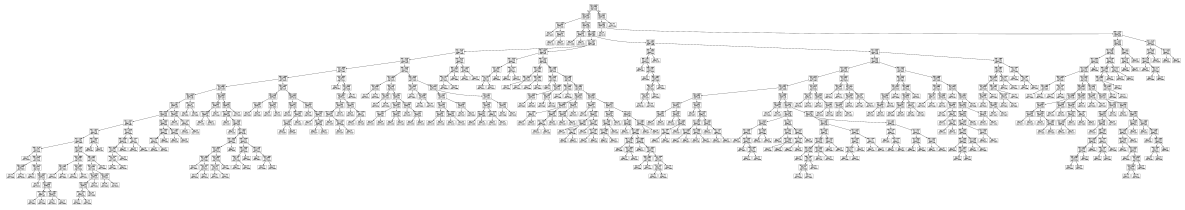
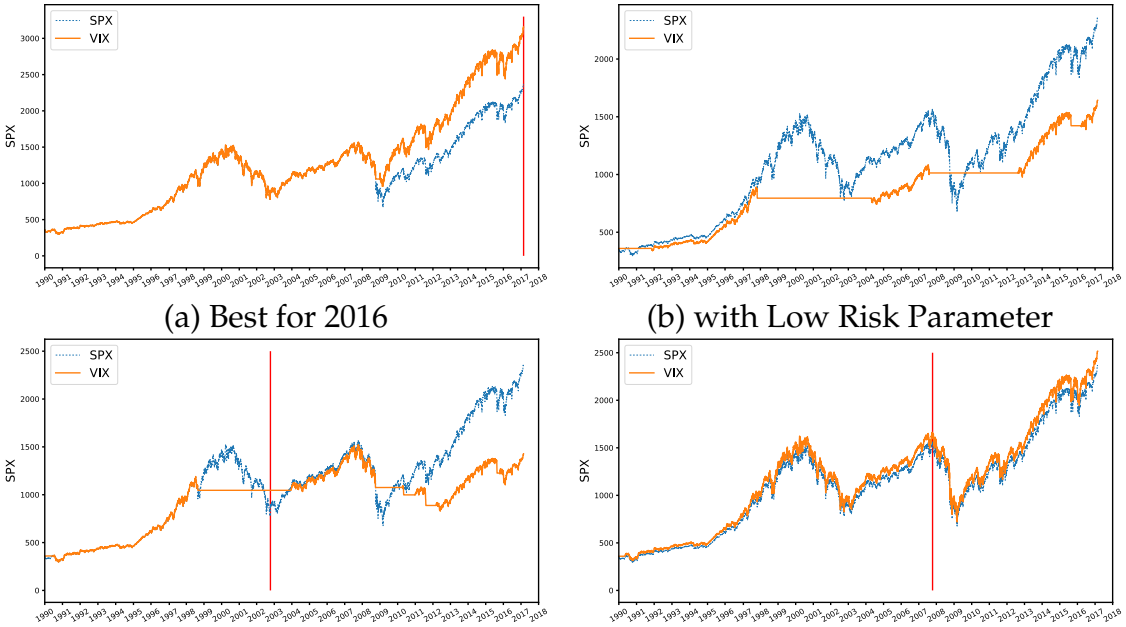


Figure 11: CART Decision Tree



(a) Best for 2016

(b) with Low Risk Parameter

(c) Best for 2002

(d) Best for 2007

Figure 12: Results of VIX

- Buy if $\text{VIX}_{min} > \text{VIX}$

Here, VIX_{max} and VIX_{min} are tuning parameters.

Figure 12 (a) shows the investment results on SPX with parameters that obtain the best return at the end of 2016. Figure 12 (c) shows the results with parameters that obtain the best return at 2012. Figure 12 (d) shows the results with parameters that obtain the best return at 2007.

Although the best parameters make profits at the end of 2016 (Fig 12 (a)), it implies that the investors have to predict the best parameter before 1990. Parameters that give the best result when SPX shows the bottom value in 2002 cannot make profit at 2016 (Fig 12 (c)). A strategy with parameters that gives the best result when SPX shows the ceiling value in 2007 is similar to index investment (Fig 12 (d)). Profit by low-risk parameters is less than that of index investment (Fig 12 (b)). The use of sophisticated parameter tuning is necessary. The handling of mixed-distribution and the effect of data period tuning were not reported so far.

4.3 Results on Other Indexes

Our last research question was whether we could apply the proposed method to other indexes. As the results shown in Figure 8 seem to suggest a change in the investors' behavior after 1990, we applied data from 2001 onwards to the CAC, DAX, SPX, and UKX indexes, with the results shown in Figures 13, 14, 15, 16, and 17, respectively. It is seen that the proposed method earns profit on all these indexes during this time period, with Sharpe ratios that are universally higher than the corresponding index investment results. Table 1 summarizes the results. Although its overall prediction accuracy is not very high ($0.52 \sim 0.6$), the proposed method outperforms the conventional methods on all indexes.

In these comparisons, we used the following model:

$$v_t = (d_t, s_t, d_{t-1}, s_{t-1}, d_{2,t}, d_{t-2}, s_{t-2}, d_{3,t})$$

Because the GBDT approach could not earn stable profits, we selected appropriate alternatives for each index. CAC was analyzed using CART; DAX, NKY, and SPX were analyzed using K-NN; and Random Forest was used to analyze UKX. All analysis was conducted using the scikit-learn package in python.

Although we could not find a method to select an optimal learning system, K-NN and tree based-methods such as GBDT, Random Forest, and CART were found to work well in general, as such methods appear to be compatible with data containing mixed distributions. SVR, by contrast, was found to be completely unsuitable. We were unable to definitively determine why (d_t, s_t) is sufficient to analyze TOPIX data while other attributes, such as d_{t-1}, d_{t-2}, are required for other indexes. These differences seem to reflect underlying difference in the investors' behavior; however, further study is required.

Note that the training periods selected for TOPIX, CAC, SPX, and UKX all seem to suggest changes in investors' behavior, while CAC and NKY use all available data, i.e., select the longest training period. The NKY and CAC data start at 1975 and 1993, respectively, and the left-side slope of the selected training period (Figure 13) reflects the lack of CAC data prior to 1993. The NKY data from 1975 always produce 300-month training periods (Figure 15).

While the changes in the best training periods for TPX, CAC, SPX, and UKX suggest the importance of optimization of the training period, the unchanging best training periods for CAC and NKY are more difficult to explain; it is possible that the investors' behavior represented by these indexes are stable or that our analysis was insufficient. Investors may change their behavior in the future, and found anomaly, i.e., the prediction ability of the proposed method, may disappear. These issues have to be studied in future work.

4.4 Other Research Issues

As shown in Table 1, the proposed method can produce profits from various market indexes after 2001. Because these indexes are related, combining one or more seems to be a promising approach to formulating investment strategies. For example, SPX and TOPIX appear to be tightly related, suggesting the usefulness of a combined SPX-TOPIX framework.

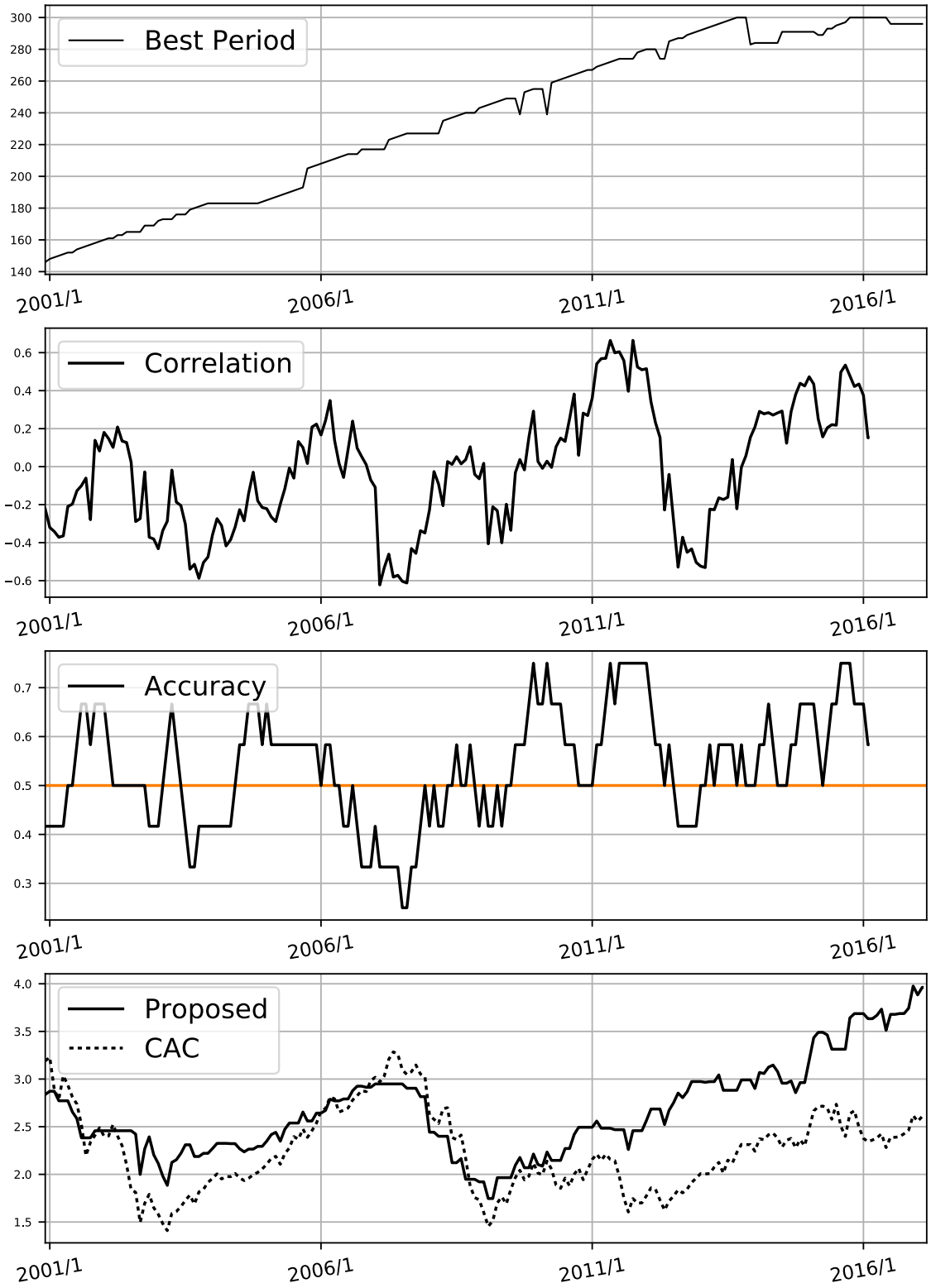


Figure 13: Results on CAC

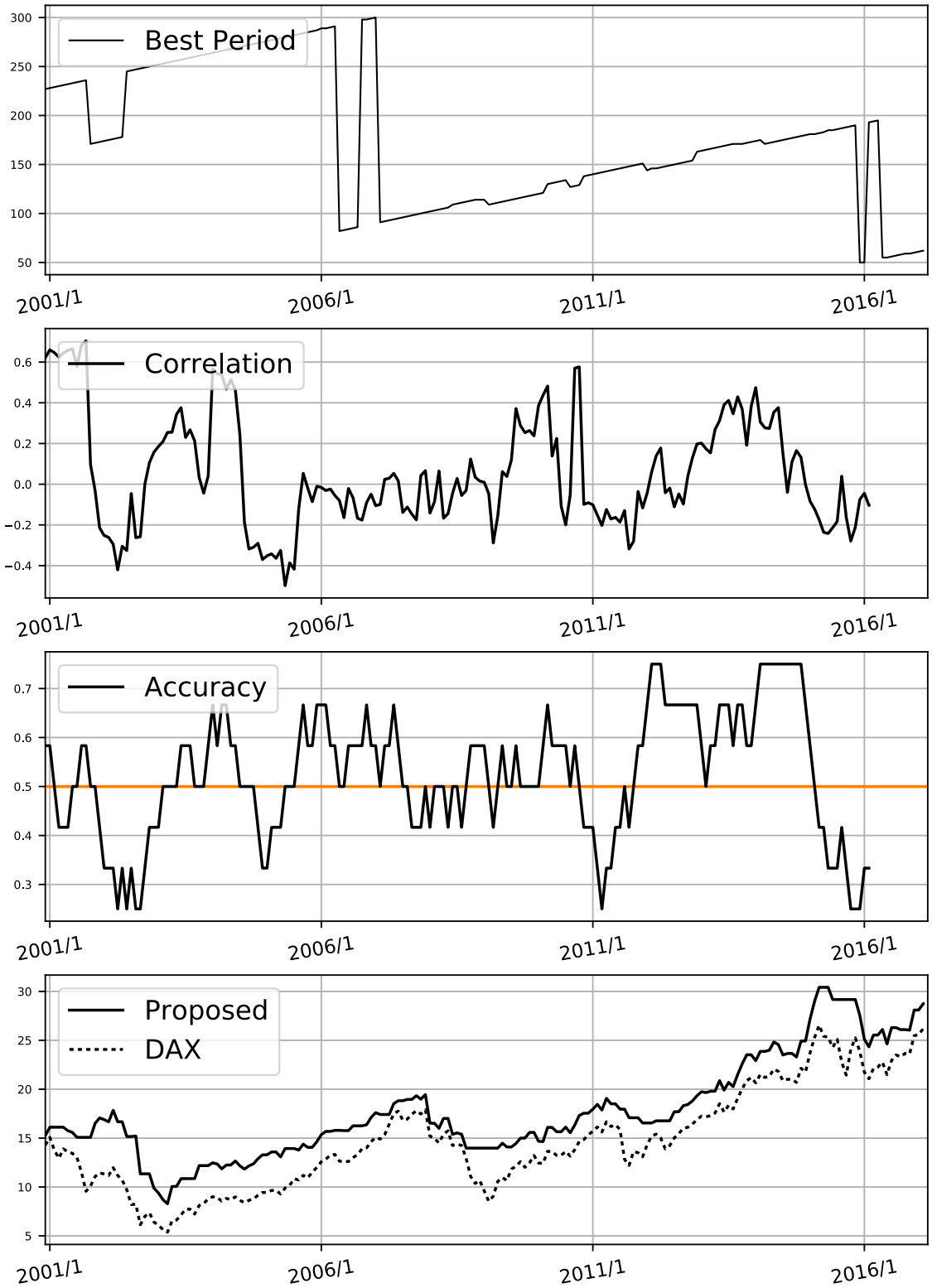


Figure 14: Results on DAX

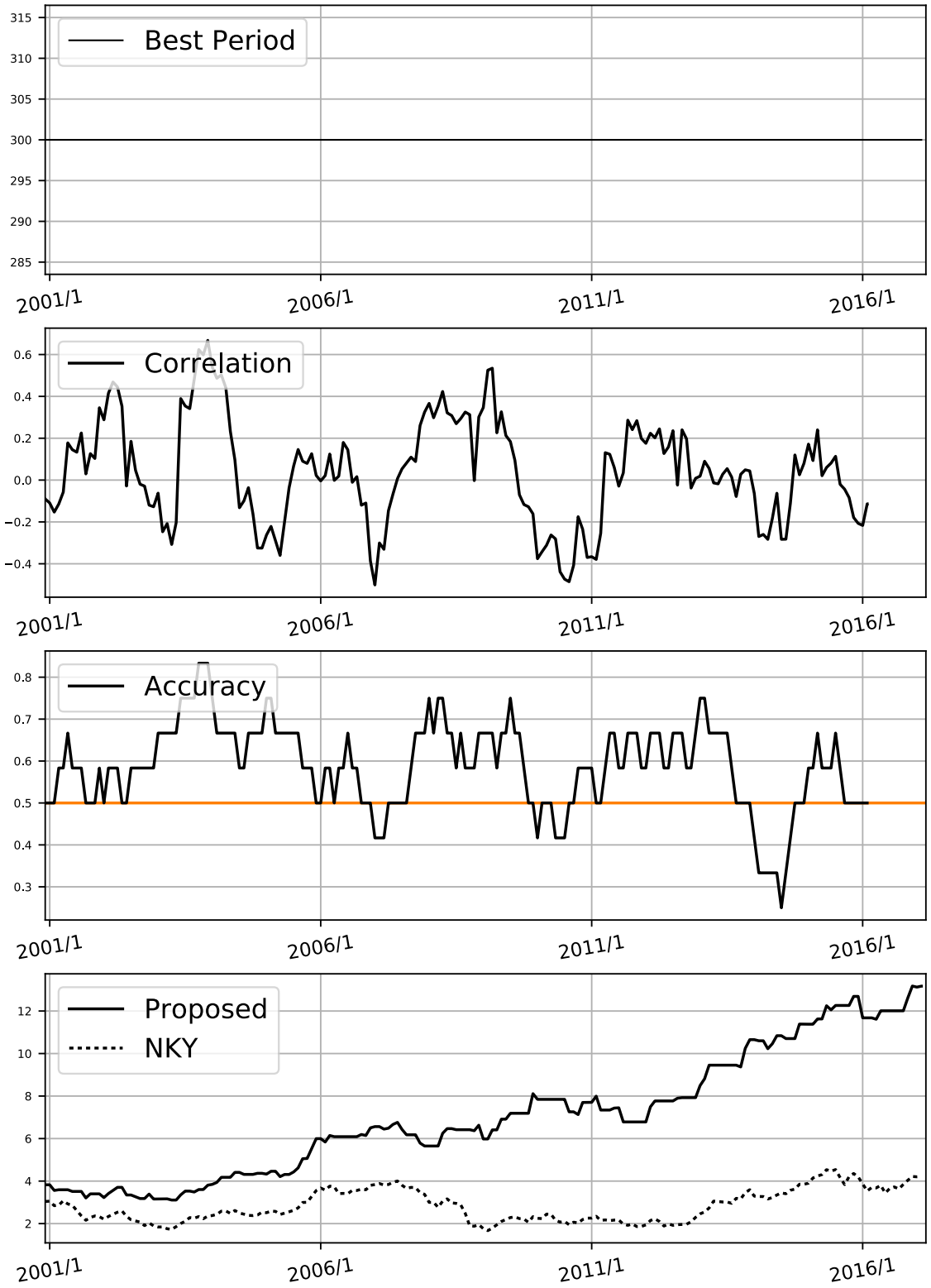


Figure 15: Results on NKY

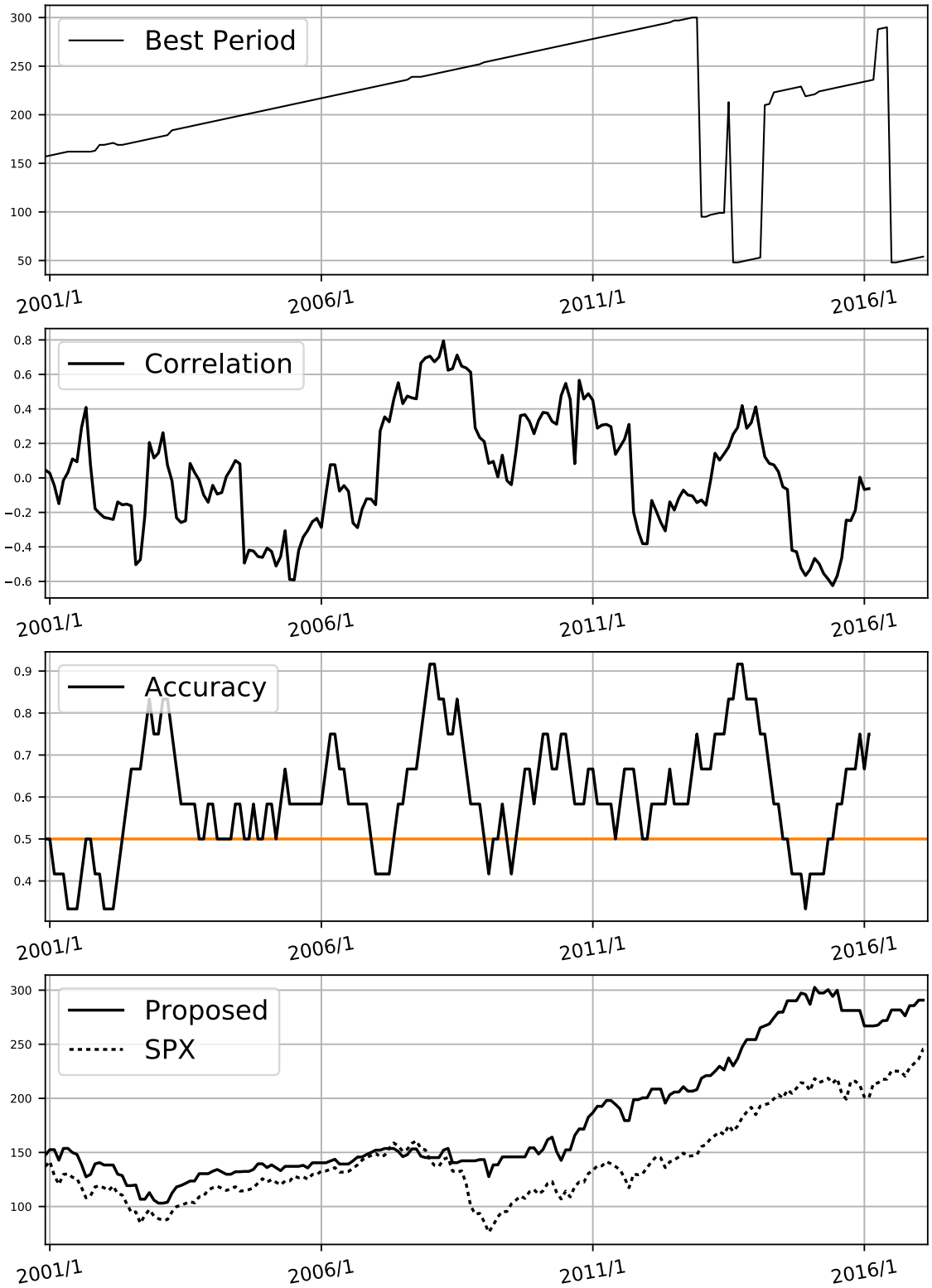


Figure 16: Results on SPX

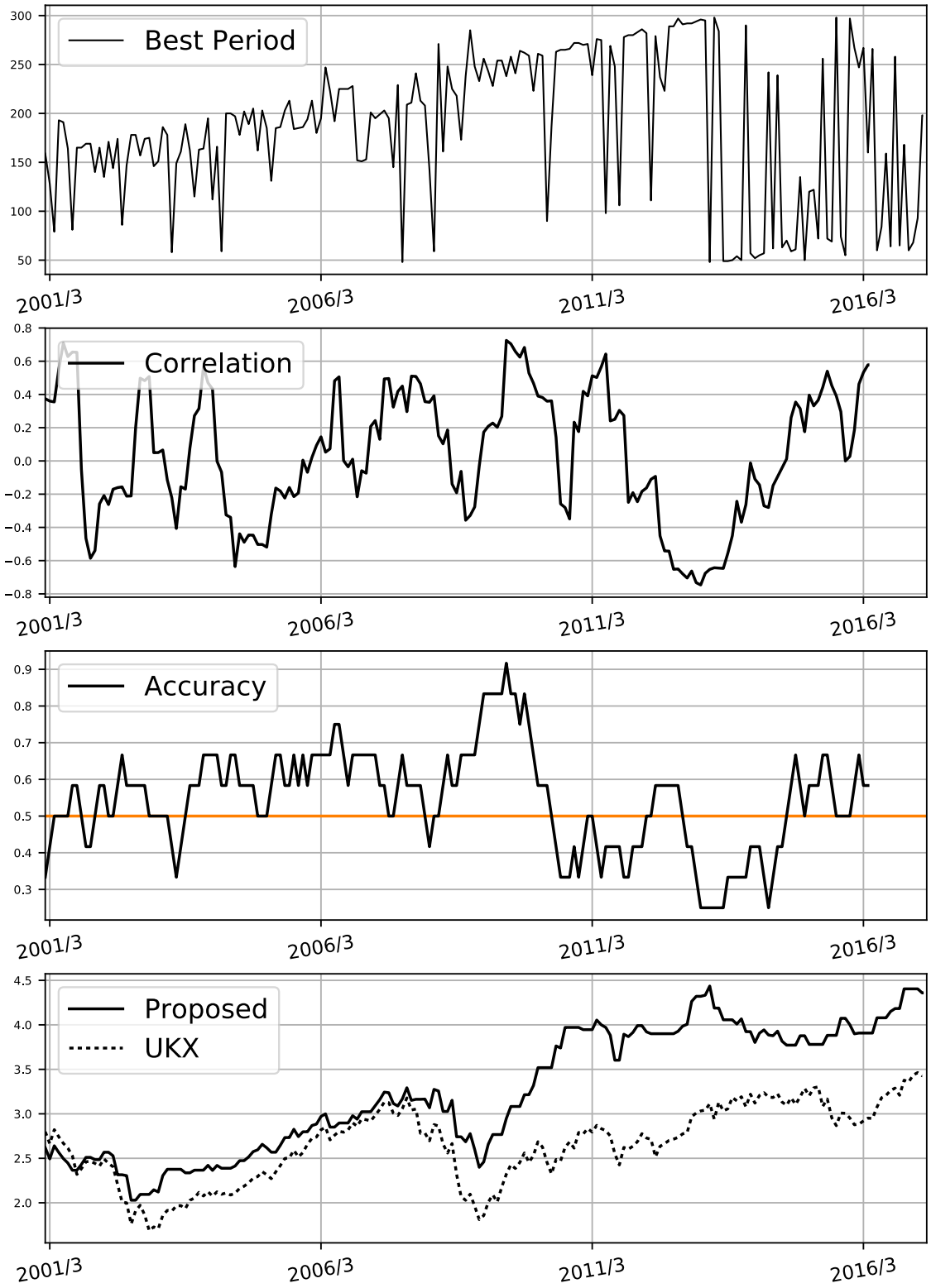


Figure 17: Results on UKX

Table 1: Sharpe Ratio of Representative Indexes

Index	Period	Proposed Method			Index	
	~2017/2	Accuracy	Sharp Ratio	Return *	Sharp Ratio	Return *
TPX	1986/1~	0.5525	0.1102	5.41%	0.0465	3.15%
CAC	2001/1~	0.5375	0.0639	2.99%	0.0062	0.39%
DAX	2001/1~	0.5234	0.0963	5.15%	0.0825	6.15%
NKY	2001/1~	0.5879	0.1960	8.44%	0.0587	3.93%
SPX	2001/1~	0.6012	0.1234	4.86%	0.0928	4.71%
UKX	2001/1~	0.5417	0.0909	3.30%	0.0398	1.88%

$*$: Annualized Return

This, however, is outside the scope of the current study, which focused on demonstrating the potential of non-time series analysis. The selection of optimal learning methods and attributes are left as future work. Here, our principal methodological findings are 1) learning methods that can handle mixed-distribution data are promising, and 2) the inter-month price fluctuation and the standard deviation of the daily stock price—both of which are non-time series data representations—are rich sources of information.

Many researchers have recently become interested in the application of deep learning methods to the analysis of stock prices [9] by using the excellent function-fitting abilities of structures such as convolutional neural networks and recurrent neural networks. Such studies have a complementary relationship to our study, as deep learners might be able to improve their analytic abilities using the attributes we have discovered. Conversely, we might be able to use deep learning methods as an improved learning framework; however, these are also issues for future research.

The final question this study raises is why it has become recently possible to make predictions using prices alone. Although we have no firm answers to this, it is possible that the advances in financial engineering in recent years have introduced a degree of regularity to the investors' behavior, or the prediction ability found this time may disappear in the future. This should be studied through future research.

5 Conclusion

In this study, we attempted to analyze the investors' behavior behind the business cycles by analyzing TOPIX price movements from 1954 to 2016 using simple data mining methods.

The characteristics of the data mining method used in this study are: 1) non-time-series data representation; 2) the use of methods that can handle mixed data distributions; and 3) optimization of the training period. The importance of these approaches were demonstrated from experimental results obtained from applying the proposed method to various index data, e.g., TOPIX, CAC, DAX, NKY, SPX, and UKX. The experimental results show the existence of anomaly where prices of these indexes can be predicted.

The important byproduct is the training period used to create the prediction model. This period seems to suggest the existence of a long-term investors' behavior over the

business cycles. Although various studies have been made to clarify the mechanism behind business cycles, these studies shows stable investors' behavior over the business cycles solely on stock price information.

1) The relationship between the identified training period and the business cycles studied by previous researches, and 2) whether the prediction ability found this time will continue, are left as future research issues.

References

[1] Cabinet Office, Goverment of Japan. Standard Date of Japanese Economy, 2018. [accessed 7-Nov-2018].

[2] Kazuo Ogawa and Shinichi Kitasaka. *Asset Market and Business Cycle: Empirical Analysis of Modern Japan Economy (in Japanese)*. Nikkei Inc., 1998.

[3] Ryuzo Matsuo. Factors behind the economic fluctuation in japan (in japanese). In Kenzo Abe, Masao Oogaki, Kazuo Ogawa, and Takatosi Tabuchi, editors, *Trends in modern economics 2011 (in Japanese)*, chapter 2, pages 35–65. Toyo Keizai Inc., 2011.

[4] Toshiaki Watanabe. Econometric analysis of business cycle in japan using markov switching model. *Economic Research*, 60:253–265, 2009.

[5] Yoshihiro Ohtsuka. Estimation of regional bussiness cycle in japan with markov switching spatial autoregressive-ar model (in japanese). *Journal of Japan Statistical Society*, 40(2):89–109, 2011.

[6] Sanjoy Basu. Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The Journal of Finance*, 32(3):663–682, 1977.

[7] James D. Hamilton. *Time Series Analysis*. Princeton university press, 1994.

[8] Rodolfo C Cavalcante, Rodrigo C Brasileiro, Victor LF Souza, Jarley P Nobrega, and Adriano LI Oliveira. Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications*, 55:194–211, 2016.

[9] Eunsuk Chong, Chulwoo Han, and Frank C Park. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, 83:187–205, 2017.

[10] Roger D Huang and Hans R Stoll. Market microstructure and stock return predictions. *Review of Financial studies*, 7(1):179–213, 1994.

[11] Liam A Gallagher and Mark P Taylor. Permanent and temporary components of stock prices: Evidence from assessing macroeconomic shocks. *Southern Economic Journal*, pages 345–362, 2002.

[12] Tarun Chordia and Avanidhar Subrahmanyam. Order imbalance and individual stock returns: Theory and evidence. *Journal of Financial Economics*, 72(3):485–518, 2004.

[13] Eric Gilbert and Karrie Karahalios. Widespread worry and the stock market. In *ICWSM*, pages 59–65, 2010.

[14] Robert D Gay Jr et al. Effect of macroeconomic variables on stock market returns for four emerging economies: Brazil, russia, india, and china. *International Business & Economics Research Journal (IBER)*, 7(3), 2011.

[15] Po-Hsuan Hsu, Yu-Chin Hsu, and Chung-Ming Kuan. Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. *Journal of Empirical Finance*, 17(3):471–484, 2010.

[16] Lukas Menkhoff. The use of technical analysis by fund managers: International evidence. *Journal of Banking & Finance*, 34(11):2573–2586, 2010.

[17] Shangkun Deng, Kazuki Yoshiyama, Takashi Mitsubuchi, and Akito Sakurai. Hybrid method of multiple kernel learning and genetic algorithm for forecasting short-term foreign exchange rates. *Computational Economics*, 45(1):49–89, 2015.

[18] Gian Piero Aielli. Dynamic conditional correlation: on properties and estimation. *Journal of Business & Economic Statistics*, 31(3):282–299, 2013.

[19] Harry Markowitz. Portfolio selection. *The journal of finance*, 7(1):77–91, 1952.

[20] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[21] Timm O Sprenger, Andranik Tumasjan, Philipp G Sandner, and Isabell M Welpe. Tweets and trades: The information content of stock microblogs. *European Financial Management*, 2013.

[22] Hongkee Sul, Alan R Dennis, and Lingyao Ivy Yuan. Trading on twitter: The financial information content of emotion in social media. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 806–815. IEEE, 2014.

[23] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In *Ijcai*, pages 2327–2333, 2015.

[24] Huy D Huynh, L Minh Dang, and Duc Duong. A new model for stock price movements prediction using deep neural network. In *Proceedings of the Eighth International Symposium on Information and Communication Technology*, pages 57–62. ACM, 2017.

[25] Kenichi Yoshida and Akito Sakurai. Short-term stock price analysis based on order book information. *Information and Media Technologies*, 10(4):521–530, 2015.

[26] George M Frankfurter and Elton G McGoun. Anomalies in finance: What are they and what are they good for? *International review of financial analysis*, 10(4):407–429, 2001.

[27] William F Sharpe. The sharpe ratio. *Journal of portfolio management*, 21(1):49–58, 1994.

[28] Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56, 1993.

[29] Cboe Global Markets, Inc. VIX Index Charts & Data, 2018. [accessed 20-Dec-2018].

[30] Andrew W Lo, Harry Mamaysky, and Jiang Wang. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The journal of finance*, 55(4):1705–1765, 2000.

[31] Clifford S Asness, Tobias J Moskowitz, and Lasse Heje Pedersen. Value and momentum everywhere. *The Journal of Finance*, 68(3):929–985, 2013.