Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 7 Issue 7, July - 2020

Financial Time Series Forecasting Based On Motif Discovery

A case study in foreign exchange rate

Min Shen
Quantitative Researcher
Trendalyze
New Jersey, United States
rachel.shen@trendtrade.io

Rado Kotorov
CEO
Trendalyze
New Jersey, United States
rado.kotorov@trendalyze.com

Lianhua Chi
Department of Computer Science and IT
La Trobe University
Melbourne, Australia
L.Chi@latrobe.edu.au

***Abstract*—The objective of this research is to provide empirical evidence that motif discovery can be applicable to predict financial time series. Two prediction methods based on motif discovery (One Motif Approach and Integrated Motif Approach) are proposed, which apply adaptive dissimilarity index [6] with Complexity-Invariant Distance (CID) [2] as the similarity measure. This paper extends the work previously introduced by Ismailaja [12]. Tests are conducted based on relatively large financial time series datasets for foreign exchange rate, and result shows that the new prediction model is more efficient with less computational complexity and higher forecasting accuracy compared to previous model.**

Keywords — motif discovery; financial time series; forecasting; adaptive dissimilarity index; CID; foreign exchange rate

I. INTRODUCTION

Time series forecasting is a common problem in various domains, including manufacturing, agriculture, retail, and tourism, etc. Among them, financial time series forecasting is extremely challenging and has been studied extensively for years. Except traditional statistical models, there are also models based on machine learning techniques, including artificial neural networks [17] and support vector machines [16] to predict financial time series.

This paper aims to provide empirical evidence that motif discovery with a novel similarity measure can be applicable to predict financial time series data as well. A new prediction approach is proposed based on Ismailaja's model [12] to make it more suitable in larger datasets. The similarity search algorithm is constructed applying the adaptive dissimilarity index [6] with Complexity-Invariant Distance (CID) [2] measure. According to test results, our method is more efficient with less computational cost and higher forecasting accuracy. In this case study, foreign exchange rate is used as testing data since the foreign exchange market is the largest and most liquid financial market globally, and it has been considered a very challenging task to predict forex data due to its nonlinearity and uncertainty [11].

II. LITERATURE REVIEW

Motif discovery has been widely studied in bioinformatics for detecting biosequences for years [23]. It can also be applied to medical data to detect anomalies in heart rhythm and blood pressure [22]. A time series motif is defined as a frequently recurrent pattern throughout the time series [20]. And motif discovery is the process of detecting and locating previously defined patterns in time series datasets [21]. An efficient motif discovery algorithm can be used as a data mining tool to summarize and analyze massive time series data.

Many of the motif discovery methods are based on searching a discrete approximation of the time series. To achieve dimensionality reduction, Agrawal et al. [1] used Discrete Fourier Transform (DFT) for processing similarity queries. Chan and Fu [4] contended that Discrete Wavelet Transform (DWT) can be effective in replacing DFT in many areas of study, including image [9], speech [14] and signal processing [13]. There are also algorithms utilizing Piecewise Aggregate Approximation (PAA) as the discretization technique [15][18][27]. Tanaka et al. [25] applied Principal Component Analysis (PCA) to reduce dimensions of data and discovered a motif based on Minimum Description Length (MDL) principle. More recently, there are series of algorithms based on Matrix Profile technique, which can improve the performance and increase the scalability of data [19][26]. In order to achieve motif discovery, distance or similarity measures between time series sequences are calculated to give a numerical value that indicates how similar or dissimilar two sequences are. Except the prevailing methods of Euclidean distance and dynamic time warping (DTW) [24], Cha [3] gave a comprehensive survey of distance and similarity measures, including Jaccard distance, Bhattacharyya similarity and cosine similarity, etc. The existing measures have divergent algorithm complexity and can achieve different levels of accuracy and quality for motif discovery. In their article, Batista and Keogh [2] introduced a new distance measure, Complexity-Invariant Distance (CID), which is proved to be an efficient measure to improve classification and clustering accuracy. In order to identify genes expression profile, Chouakria et al. [5] proposed a novel dissimilarity index based on an automatic

adaptive tuning function to include proximity measures with respect to value and behavior. In [6], Chouakria and Nagabhushan compared the adaptive dissimilarity index with other conventional measures and concluded that the adaptive dissimilarity is a more suitable measure to reflect the expected behavior and dynamics of the data. Furthermore, Dhamo et al. [8] provided empirical evidence that combining the adaptive dissimilarity index with CID can achieve better results in motif discovery than CID alone.

Moving forward to time series forecasting, Ismailaja [12] applied the adaptive dissimilarity index (Chouakria's index in [12]) with CID as the distance measure in motif discovery and claimed that the prediction model can achieve better forecasting results than ARIMA model. However, the author only worked with relatively small datasets with less than a thousand observations. The results may vary for relatively large datasets in a different domain.

This research extends the previous work by proposing a more efficient prediction approach with greater accuracy and less computational complexity for relatively larger financial time series applying motif discovery.

III. Background and Notation

In this section, we introduce the definitions and notations of key terms in this paper, then we propose our methodology and build the forecasting model in next section.

A. Basic Concepts

A time series may be defined as a sequence of real numbers. In this case study, a time series is observed at successive times.

Definition 1: A *time series* is a sequence $T = [t_1, t_2, \dots, t_n]$, which is an ordered set of n real valued numbers obtained in n regular intervals of time.

Definition 2: A *subsequence* of length m of a time series $T = [t_1, t_2, \dots, t_n]$ is a time series $T_{i,m} = [t_i, t_{i+1}, \dots, t_{i+m-1}]$ for $1 \le i \le n - m + 1$.

Definition 3: A *time series motif* in T of length m is a repeated subsequence of T.

Definition 4: In a *similarity search*, two subsequences $T_{i,m} = [t_i, t_{i+1}, \dots, t_{i+m-1}]$ and $T_{j,m} = [t_j, t_{j+1}, \dots, t_{j+m-1}]$ of length m in time series T are similar if the distance between them is within absolute error ε.

Definition 5: The *k^{th} time series motif* of length m is the k^{th} most similar non-overlapping subsequence to a given subsequence $T_{i,m} = [t_i, t_{i+1}, \dots, t_{i+m-1}]$ in the time series T.

Definition 6: The *best match motif* is the first time series motif, i.e., the most similar non-overlapping subsequence of length m to a given subsequence $T_{i,m} = [t_i, t_{i+1}, \dots, t_{i+m-1}]$ in the time series T.

B. Distance Measures

The similarity measure in a similarity search is based on the distance between two subsequences in a time series. The most prevailing measure is the Euclidean distance.

Definition 7: The *Euclidean distance* between two time series $Q = [q_1, q_2, \dots, q_m]$ and $C = [c_1, c_2, \dots, c_m]$ of length m is

$$ED(Q, C) = \sqrt{\sum_{i=1}^{m} (q_i - c_i)^2}\,. \quad (1)$$

While Euclidean distance is a prevalent method, in many domains the data are distorted in some way and a more robust measure is necessary. Comparing to Euclidean distance, applying the adaptive dissimilarity index proposed by Chouakria et al. [6] can capture more expected behaviors and dynamics of data. Also, the Complexity-Invariant Distance (CID) introduced by Batista and Keogh [2] has a correction factor which accounts for complexity differences between two time series. Considering the complexity of financial time series, the result might be better if the similarity measure can reflect the behavior of data. In this case study, we combine the adaptive dissimilarity index with CID in order to improve the quality of motif search. The definitions of the above similarity measures are introduced as following.

Definition 8: The *Complexity-Invariant Distance (CID)* between two time series $Q = [q_1, q_2, \dots, q_m]$ and $C = [c_1, c_2, \dots, c_m]$ of length m is

$$CID(Q, C) = ED(Q, C) * \frac{\max(CE(Q), CE(C))}{\min(CE(Q), CE(C))} \quad (2)$$

where CE is defined as a complexity estimate:

$$CE(Q) = \sqrt{\sum_{i=1}^{m-1} (q_i - q_{i+1})^2} \quad (3)$$

This similarity measure has a complexity correction factor $\frac{\max(CE(Q), CE(C))}{\min(CE(Q), CE(C))}$, which accounts for the complexity differences between the time series Q and C. This factor can force the distance of time series with high level of complexity differences be further. If two time series have the same level of complexity, the correction factor will be 1 and the CID simply become the Euclidean distance. Also, since $\frac{\max(CE(Q), CE(C))}{\min(CE(Q), CE(C))}$ is greater or equal to 1, $CID(Q, C) \ge ED(Q, C)$.

In order to find behavior proximity measure of time series data, the temporal correlation coefficient [7] is defined. It measures the monotonicity and growth rate features of two subsequences.

Definition 9: The *temporal correlation coefficient* of two time series $Q = [q_1, q_2, \dots, q_m]$ and $C = [c_1, c_2, \dots, c_m]$ of length m is

$$CORT(Q, C) = \frac{\sum_{i=1}^{m-1} (q_{i+1} - q_i)(c_{i+1} - c_i)}{\sqrt{\sum_{i=1}^{m-1} (q_{i+1} - q_i)^2} \sqrt{\sum_{i=1}^{m-1} (c_{i+1} - c_i)^2}} \quad (4)$$

The value of $CORT(Q, C)$ belongs to the interval $[-1, 1]$. $CORT(Q, C) = -1$ indicates that in the observed period where Q increases, C decreases and

vice versa with a same growth rate (in absolute value). $CORT(Q,C)=1$ means that Q and C have similar behaviors in the observed period. They increase and decrease simultaneously at the same growth rate. A value of $CORT(Q,C)=0$ indicates that Q and C exhibit different behaviors (neither similar nor opposite), and their growth rates are stochastically linearly independent.

Definition 10: The *adaptive dissimilarity index with CID* measure between two time series $Q=[q_1,q_2,\dots,q_m]$ and $C=[c_1,c_2,\dots,c_m]$ of length m is

$$AD(Q,C)=\frac{2}{1+e^{k*CORT(Q,C)}}*\delta_{Q,C}\,,k\geq 0, \quad (5)$$

where $\delta_{Q,C}=CID(Q,C)$. $CORT(Q,C)$ is defined in (4).

The adaptive dissimilarity index is composed by two factors, one is responsible for behavior ($CORT(Q,C)$), and the other accounts for proximity of values ($\delta_{Q,C}$). In this measure, the first term $\frac{2}{1+e^{k*CORT(Q,C)}}$ is an exponential adaptive tuning function. According to [6], setting $k=2$ can captures about 20% of behavior and 80% of the value. We decide to choose $k=2$ in this measure in order to capture some degrees of the behavior of the time series.

IV. Methodology

In this section, we will introduce the basic logic and detailed methodology of forecasting approaches applying motif discovery. The prediction methods are based on the assumptions that there are similar subsequences throughout the time series, and we can regard them as recurrent patterns, i.e., motifs.

Before the distance between subsequences in a time series is computed, we need to normalize the time series data in order to transform the data to comparable scales and offset invariance. Standardization is often used to normalize a time series.

Definition 11: Standardization of the time series $T=[t_1,t_2,\dots,t_n]$ is defined as $T'=[t_1',t_2',\dots,t_n']$, where

$$t_i'=\frac{t_i-\mu}{\sigma}, i\epsilon[1,n] \quad (6)$$

In this equation, μ is the mean of the values in the time series T, and σ is the standard deviation of T. The standardization rescales the values in the time series to have mean of 0 and standard deviation of 1. It does not ensure that all points of the resulting time series are in the $[0,1]$ interval.

After preprocessing and normalizing the data, we can compute the distance between subsequences in the time series and conduct similarity search. Given a subsequence M of fixed length m, the subsequence $T_{i,m}=[t_i,t_{i+1},\dots,t_{i+m-1}]$ of length m in time series T is considered similar with the given subsequence M if the distance between M and $T_{i,m}$ is within a defined absolute error ε.

A. Previous Forecasting Approach

In a similarity search, we are able to find similar patterns of a given subsequence by setting an absolute error for distance measure. However, in order to apply motif discovery in forecasting, we focus on the most similar subsequence of a given motif in the time series and make prediction based on the trend after this subsequence.

Considering a time series $T=[t_1,t_2,\dots,t_n]$ with length n, the prediction algorithm for the next data point t_{n+1} works as following:

1. Keep the last subsequence of length m as the motif for similarity search, which is $M=[t_{n-m+1},t_{n-m+2},\dots,t_n]$.
2. Choose a distance measure for the time series (adaptive dissimilarity index with CID in this case).
3. Conduct similarity search from t_1 to t_{n-m}, i.e., calculate the distance between the motif M and other subsequences of length m in the time series and keep track of the index of the most similar subsequence.
4. Suppose the most similar subsequence we found in step 3 is at position i, then a dependency factor is created for prediction:
$$dep_{factor}=\frac{T[i+m]}{T[i+m-1]} \quad (7)$$
5. The prediction for $t_{n+1}$is computed as:
$$t_{n+1}=t_n*dep_{factor} \quad (8)$$

In order to predict the next k points, we need to append the forecasted point to the original time series and loop the above method for k times. For example, the prediction of t_{n+2} is based on the result of similarity search for motif $M'=[t_{n-m+2},\dots,t_{n+1}]$.

The logic of creating a dependency factor is that the trend for the next point of the last subsequence in the time series is similar to the trend for the following point of the most similar motif that discovered. However, this approach can be inefficient to predict more datapoints for relatively large datasets, since another similarity search is required for the renewed motif after appending the last prediction result to the original time series. To make the algorithm more efficient for long-term prediction, we propose two approaches with less computational complexity: One Motif Approach and Integrated Motif Approach. We will introduce them in part B and part C.

B. One Motif Approach

Suppose we are predicting the next k points of a time series $T=[t_1,t_2,\dots,t_n]$ with length n. The One Motif Approach works as following:

1. Choose the last subsequence of length m in the time series, $M=[t_{n-m+1},t_{n-m+2},\dots,t_n]$ as the motif for similarity search, where $m\geq k$.
2. Apply a suitable distance measure for subsequences in the time series (same as part A, adaptive dissimilarity index with CID is used in this case).

3. Calculate the distance between motif M and other subsequences of length m from t_1 to t_{n-2m} and keep track of the distance and the corresponding index of all the subsequences of length m.
4. Suppose the most similar non-overlapping subsequence, i.e., the best match motif we found in step 3 is at position i, then a list for dependency factors is created for prediction:
$$dep_{factor}[j] = \frac{T[m+i+j]}{T[m+i+j-1]}, j \in [0,k). \quad (9)$$
5. The prediction points $[t_{n+1}, \dots, t_{n+k}]$ are computed as:
$$t_{n+1+j} = t_{n+j} * dep_{factor}[j], j \in [0,k). \quad (10)$$

For example, the first point for prediction is t_{n+1}, which is calculated as:

$$t_{n+1} = t_n * dep_{factor}[0] \quad (11)$$

where $dep_{factor}[0]$ is the first element of the list of dependency factors.

Different from the previous method, after appending the prediction result to the original time series, this prediction approach does not require to search for the most similar subsequence again based on the renewed last subsequence (motif) of length m. Instead, we make the second prediction based on the first prediction result and the second element of the dependency factor list, which is

$$t_{n+2} = t_{n+1} * dep_{factor}[1] \quad (12)$$

The above step is conducted for k times in order to get k prediction points based on k dependency factors.

The One Motif Approach focuses on the most similar non-overlapping subsequence (best match motif) to the given motif in order to make predictions. Instead of conducting the similarity search and calculating the dependency factor for k times in the previous approach, we only do similarity search once to find out the best match motif. The elements in the dependency factor list are based on the next k points of the best match motif.

In order to get a list of dependency factors, the best match motif is searched from t_1 to t_{n-2m}, which covers all non-overlapping subsequences. Also, this approach requires $m \geq k$, i.e., the selected motif length for similarity search should be greater than or equal to the number of points for prediction to prevent index out of range problem.

C. Integrated Motif Approach

Similar to the One Motif Approach, Integrated Motif Approach takes the first, second and third time series motifs (the three most similar subsequences to the given motif) and creates three dependency factor lists. Step 1-3 are the same as that in part B. Step 4 and 5 are revised as the following:

Suppose the position of the first, second and third time series motifs (the three most similar non-overlapping subsequences) are i, p, q, respectively. Three dependency factor lists are created for prediction:

$$dep_{fact1}[j] = \frac{T[m+i+j]}{T[m+i+j-1]}, dep_{fact2}[j] = \frac{T[m+p+j]}{T[m+p+j-1]},$$
$$dep_{fact3}[j] = \frac{T[m+q+j]}{T[m+q+j-1]}, j \in [0,k). \quad (13)$$

where dep_{fact1}, dep_{fact2} and dep_{fact3} are dependency factor lists for the first, second and third time series motifs, respectively.

The forecasted points $[t_{n+1}, \dots, t_{n+k}]$ are calculated as:

$$t_{n+1+j} = [(dep_{fact1}[j] + dep_{fact2}[j] + dep_{fact2}[j])/3] * t_{n+j}, \ j \in [0,k). \quad (14)$$

The Integrated Motif Approach focuses on the three most similar subsequences (top three best match motifs) in the time series and creates three dependency factor lists based on the trend after those motifs for prediction. The three lists are then integrated (averaged) to calculate the forecasted points.

V. Results and Analysis

In this section, multiple tests are conducted with foreign exchange rate data for three forecasting approaches stated in section IV. We will introduce the data source and preprocessing in part A, compare the results of different approaches in part B, and do further analysis in part C.

A. Data

All the tests in this section are conducted using datasets of EUR/USD exchange rate (close price) from 01/01/2015 to 12/31/2018. The data is publicly available from Dukascopy [4]. We use three different time units for testing: minutely, hourly, and daily. The data is processed to only include those during trading sessions, i.e., exclude flat prices during weekends and holidays, so that there is volatility through the time series. The time series data are normalized according to formula (6) before the motif search.

B. Results

In this case study, we firstly test the models using minutes data. We randomly select a dataset of length 1000 (01.01.2017 22:00 to 01.02.2017 14:40 in this case) and forecast the price of next 15 and 30 minutes with motif length of 15 and 30, respectively.

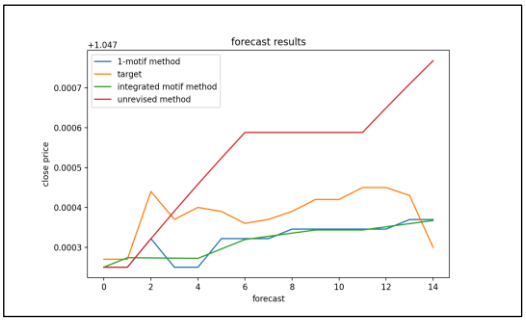

Fig. 1. *15/15 forecast result (forecast the next 15 minutes with motif length of 15)*

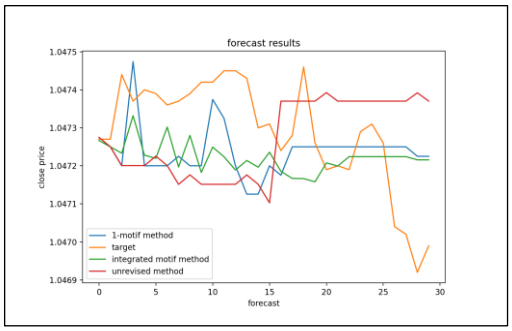

Fig. 2. *30/30 forecast result (forecast the next 30 minutes with motif length of 30)*

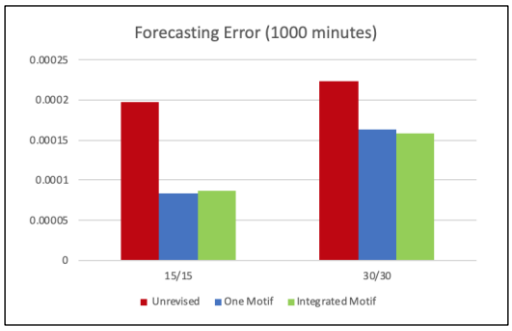

Fig. 3. *MSE for three forecasting methods (based on 1000 minutes data)*

From Fig.1 and Fig.2, we can conclude that One Motif Approach and Integrated Motif Approach perform better than the unrevised method in both 15/15 and 30/30 cases as they are closer to the real data. Specifically, in 15/15 case (Fig.1, forecast next 15 minutes with motif length of 15), the One Motif Approach successfully captures some patterns and trend in the prediction, while the line for Integrated Motif Approach is more smooth. The unrevised method fails to capture the overall trend as well as the patterns of the data in this case. In 30/30 case (Fig.2, forecast next 30 minutes with motif length of 30), the unrivised method and the One Motif Approach capture more patterns of the real data than Integrated Motif Method. However, the unrevised approach predicts an upward trend while the real data is downward.

The forecasting error is calculated using mean squared error, or MSE, as shown in Fig.3. The One Motif Approach and Integrated Motif Approach have lower MSE than unrevised method in both cases, and the forecasting errors are very close for two proposed methods.

Then we randomly select an hourly dataset with length 1000 (06.25.2017 22:00 to 08.22.2017 15:00 in this case) and forecast the next 15 hours and 30 hours price with motif length of 15 and 30, respectively. As shown in Fig. 4 and Fig. 5, the Integrated Motif Method have better performance in both 15/15 and 30/30 cases. The One Motif Method performs the worst in 15/15 case, however, it successfully predicts the upward trend of the data in 30/30 case. That might due to the great dependence of One Motif Approach on the top result of similarity search given that it only uses datapoints after the best match motif. As Fig. 6 shows, the One Motif Approach has the highest forecasting error in 15/15 case, but the lowest in 30/30 case. In both cases the MSE of the Integrated Motif Approach is lower than that of the previous method.

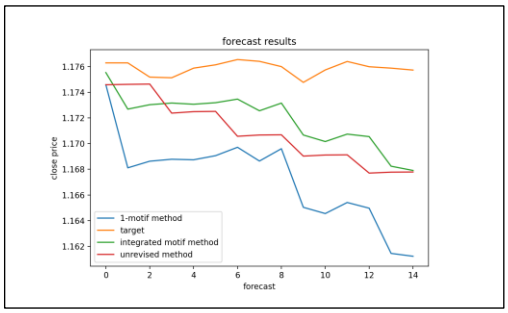

Fig. 4. *15/15 forecast result (forecast the next 15 hours with motif length of 15)*

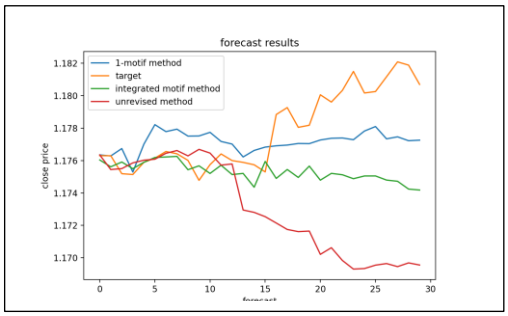

Fig. 5. *30/30 forecast result (forecast the next 30 hours with motif length of 30)*

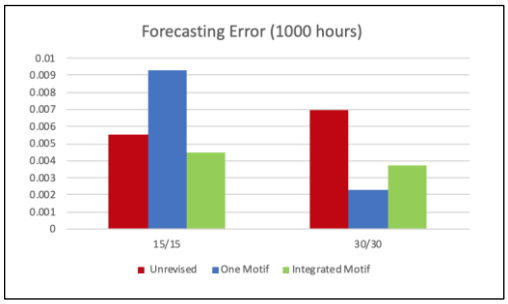

Fig. 6. *MSE for three forecasting methods (based on 1000 hours data)*

For daily exchange rate, we test the models using a dataset of length 500 (03.15.2016 to 03.13.2018 in this case) and forecast the next 10 days and 15 days exchange rate with motif length of 10 and 15, respectively. The results are shown in Fig. 7-10.

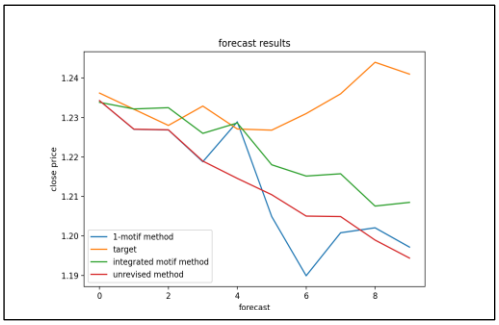

Fig. 7. *10/10 forecast result (forecast the next 10 days with motif length of 10)*

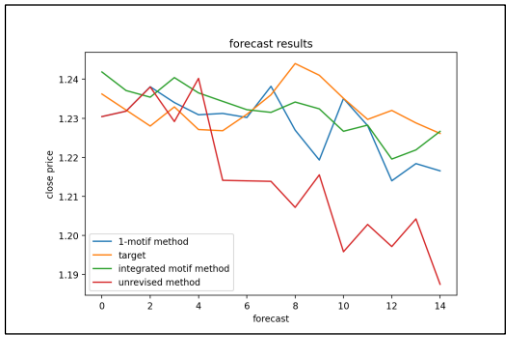

Fig. 8. *15/15 forecast result (forecast the next 15 days with motif length of 15)*

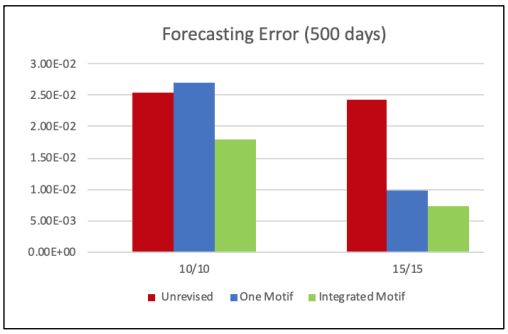

Fig. 9. *MSE for three forecasting methods (based on 500 days data)*

We can see from Fig. 7 that all three methods fail to predict the upward trend in the 10/10 case (forecast the next 10 days price with motif length of 10). However, the Integrated Motif Approach perfoms relatively the best in this case, since its forecast is the closest to the target. All three approaches perform better in 15/15 case, espacially the One Motif Method and Integrated Motif Method. They are able to capture the downward trend of the data as well as some specific features. We can see from Fig. 9 that the MSE of Integrated Motif Approach is the lowest in both cases, the One Motif Approach has the highest MSE in 10/10 case, but then it drops dramatically in 15/15 case. The MSE of unrevised method does not change drastically for the 15/15 case compare to the 10/10 case.

In summary, according to the forecasting error (MSE) of three prediction methods in all cases above, the Integrated Motif Approach has relatively lower MSE than the One Motif Approach and unrevised method in most of the cases. Moreover, although the performance of One Motif Approach fluctuates, it achieves the lowest forecasting error in some cases (Fig. 3 and Fig. 6) and can capture some specific patterns in the prediction (Fig. 1 and Fig. 2).

Also, as mentioned in section IV, since the One Motif Approach and Integrated Motif Approach only conduct similarity search once after the motif length is selected, they have much less time complexity comparing to the unrevised method. The running time of three forecasting methods is computed in Table I.

TABLE I. RUNNING TIME

Time series length	Forecasting Methods		
	Unrevised	***One Motif***	***Integrated Motif***
300	17.71	0.48	0.51
500	43.90	1.47	1.52
1000	168.80	5.36	5.43

Time unit is in seconds. Tests are based on minutes data used in Fig.2 (forecast 30 datapoints with motif length of 30).

C. Discussion

According to the empirical results in part B, the Integrated Motif Approach performs generally better in almost all cases with lower forecasting error than previous method. The prediction error of One Motif Approach is the lowest in some cases, whereas in some cases it is greater than previous approach. This result makes sense because the One Motif Approach only selects the best match motif to calculate dependency factors for forecasting, making the result mostly depend on the data after the best match motif. Comparing to Integrated Motif Approach, although One Motif Method has higher forecasting error in most of the cases, it can sometimes catch more specific patterns of the time series in forecasting. Therefore, it depends on the purpose of the prediction for choosing a suitable method.

As Table I shows, both One Motif and Integrated Motif approach have much less computational cost than the original method, with about 30 times less running time. The two proposed methods do not require a new similarity search for the renewed motif after appending the prediction data. However, the accuracy of the prediction is not decreased but improved in most of the cases.

In addition, we will detect the factors that might affect the prediction results of two proposed approaches (One Motif and Integrated Motif) in this section. Two parameters of the forecasting method will be tested, including the time series length and the selected motif length.

Firstly, the length of the time series can be a critical factor that influences the forecasting result. It might not be true that the more data in the time series, the more accurate the forecast is. We will focus on hourly EUR/USD exchange rate data (from 01/01/2015 to 12/31/2018) to do further analysis. The tests forecast the next 15 hours of the exchange rate with fixed motif length of 15. Fig. 10 illustrates the forecasting error (MSE) of two proposed apporaches with increased length of data in the time series.

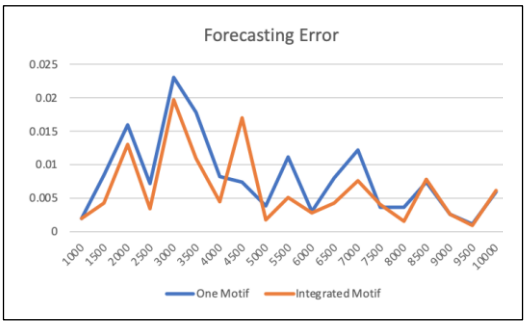

Fig. 10. *MSE of different length of datasets. Based on hourly forex data.*

In 80% of the cases, the Integrated Motif Approach has lower forecasting error than the One Motif Method. As the time series enlarged, the forecasting error fluctuates but there is no decreasing trend as the length of the time series increased.

The length of the motif can be another key factor in forecasting. As introduced in section IV, the length of the motif should be greater than or equal to the number of points to forecast in order to create lists for dependency factor for proposed approaches. We will fix the length of the time series to 1000 (1000 hours data in this case, from 01.01.2017 22:00 to 02.28.2017 13:00), and detect the influence of motif length. Fig. 11 shows the result of forecasting the next 15 hours exchange rate with different motif length. We can see that the motif length of 40 achieves the lowest forecasting error, and the prediction result of One Motif Approach is affected more by the motif length than the Integrated Motif Approach.

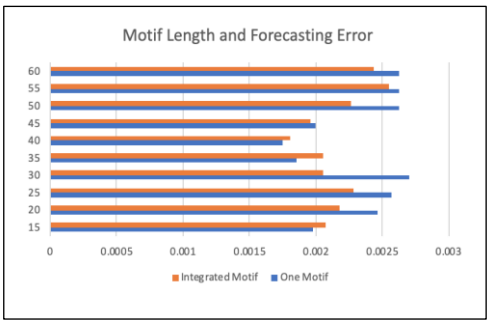

Fig. 11. *MSE of different length of motif. Fixed time series length n = 1000.*

In sum, the result of the forecasting model depends on a lot of factors. The first one is the length of the time series. However, it is not the case that the more data, the more accurate the prediction is. Also, according to Ismailaja [12], the lower error in prediction can be achieved with motif length equal to the periodicity of the time series. However, with great uncertainty of the forex data, it is hard to determine the periodicity of the time series. In this case, we should consider other factors such as the variation of the given time series in order to choose an optimal length of the motif.

VI. Conclusion And Future Work

In this research, we extend previous work by Ismailaja [12] and propose two methods to predict the foreign exchange rate applying motif discovery: One Motif Approach and Integrated Motif Approach. Comparing with the previous forecasting method, the Integrated Motif Approach performs generally better with lower forecasting error in almost all cases. The performance of One Motif Approach is more volatile, but it sometimes can achieve better results than the Integrated Motif Approach.

In the future, we will improve the model with more sophisticated means to detect the optimal length of the motif. We will also further explore the model with other financial time series data, such as equity and commodity price and implement the model into trading strategies.

References

[1] Agrawal, R., Faloutsos, C., Swami, A. (1993, October). “Efficient similarity search in sequence databases”. In International conference on foundations of data organization and algorithms (pp. 69-84). Springer, Berlin, Heidelberg.

[2] Batista, G. E., Keogh, E. J., Tataw, O. M., & De Souza, V. M. (2014). “CID: an efficient complexity-invariant distance for time series”. Data Mining and Knowledge Discovery, 28(3), 634-669.

[3] Cha, S. H. (2007). “Comprehensive survey on distance/similarity measures between probability density functions”. City, 1(2), 1.

[4] Chan, K. & Fu, W. (1999). “Efficient time series matching by wavelets”. Proceedings of the 15 th IEEE International Conference on Data Engineering.

[5] Chouakria, A. D., Diallo A., Giroud F., (2007) “Adaptive clustering of time series”. International Association for Statistical Computing (IASC), Statistics for Data Mining, Learning and Knowledge Extraction, Aveiro, Portugal

[6] Chouakria, A. D., & Nagabhushan, P. N. (2007). “Adaptive dissimilarity index for measuring time series proximity”. Advances in Data Analysis and Classification, 1(1), 5-21.

[7] Chouakria-Douzal, Ahlame. "Compression technique preserving correlations of a multivariate temporal sequence." International symposium on intelligent data analysis. Springer, Berlin, Heidelberg, 2003.

[8] Dhamo, E., Ismailaja, N., Kalluçi, E., (2015): “Comparing the efficiency of CID distance and CORT coefficient for finding similar subsequences in time series”, Sixth International Conference ISTI, 5-6 June.

[9] Eric J. Stollnitz, Tony D. Derose, and David H. Salesin. Waveletsfor Computer Graphics. Morgan Kaufmann, 1996.

[10] “EUR/USD Historical Data – dukascopy.com.” [Online]. Aviailable:

https://www.dukascopy.com/swiss/english/marketwatch/historical/. [Accessed: 15-March-2020].

[11] Hafner, C. (2013). Nonlinear time series analysis with applications to foreign exchange rate volatility. Springer Science & Business Media.

[12] Ismailaja, Nertila. "Motifs in time series for prediction." Journal of Multidisciplinary Engineering Science and Technology 2.11 (2015).

[13] John J. Benedetto and Michael W. Frazier. Wavelets – Mathematics and Applications. CRC, 1994.

[14] Johnson Ihyeh Agbinya. "Discrete wavelet transform techniques in speech processing". IEEE TENCON - Digital Signal Processing Applications, pages 514–519, 1996.

[15] Keogh, E., Chakrabarti, K., Pazzani, M., & Mehrotra, S. (2001). "Dimensionality reduction for fast similarity search in large time series databases". Knowledge and information Systems, 3(3), 263-286.

[16] Kim, K. J. (2003). "Financial time series forecasting using support vector machines". Neurocomputing, 55(1-2), 307-319.

[17] Krollner, B., Vanstone, B. J., & Finnie, G. R. (2010, April). "Financial time series forecasting with machine learning techniques: a survey". In Esann.

[18] Lin, J., Keogh, E., Patel, P., Lonardi, S.: "Finding Motifs in Time Series". In: Proceedings of the 2nd Workshop on Temporal Data Mining, at the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2002).

[19] Linardi, M., Zhu, Y., Palpanas, T., & Keogh, E. (2018, May). "Matrix profile X: VALMOD-scalable discovery of variable-length motifs in data series". In Proceedings of the 2018 International Conference on Management of Data (pp. 1053-1066).

[20] Mueen, A., & Chavoshi, N. (2015). "Enumeration of time series motifs of all lengths". Knowledge and Information Systems, 45(1), 105-132.

[21] Mueen, A., & Keogh, E. (2010, July). "Online discovery and maintenance of time series motifs". In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1089-1098).

[22] Pathinarupothi, R. K., & Rangan, E. (2016, September). "Discovering vital trends for personalized healthcare delivery". In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (pp. 1106-1109).

[23] Qader, N. N., & Al-Khafaji, H. K. (2014). "Motif discovery and data mining in bioinformatics". Int. J. Comput. Technol, 13(1), 4082-4095.

[24] Sankoff, D., & Kruskal, J. (1983). Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Cambridge University Press, 2000.

[25] Tanaka, Y., Iwamoto, K., & Uehara, K. (2005). "Discovery of time-series motif from multi-dimensional data based on MDL principle". Machine Learning, 58(2-3), 269-300.

[26] Yeh, Chin-Chia Michael, et al. "Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile." Data Mining and Knowledge Discovery 32.1 (2018): 83-123.

[27] Yi, B.-K., Faloutsos Ch., (2000): "Fast time sequence indexing for arbitrary Lp norms". In The VLDB Journal, pages 385–394.