Approach for Retrieving Similar Stock Price Patterns Using Dynamic Programming Method

Yoshihisa Udagawa
Faculty of Engineering,
Tokyo Polytechnic University
1583 Iiyama, Atsugi-city, Kanagawa
243-0297 Japan

udagawa@cs.t-kougei.ac.jp

ABSTRACT

With the widespread use of the internet, online stock trading becomes popular. A huge amount of stock trading data are accumulated in the internet. Now, stock price prediction is a challenging research subject of the data mining techniques. Because stock price can vary according to uncontrollable factors such as interest ratios, investors' sentiment, or political actions, the fluctuation of stock prices moves seemingly random. However, technical analyses of stock prices recognize that there are chart patterns occurred repeatedly known as "Japanese candlestick chart patterns" for examples. In this paper, we propose a dynamic programming approach to retrieve similar stock price patterns. The longest common substring (LCS) algorithm is improved to deal with similar numeric sequences. The proposed LCS algorithm is compared with the Dynamic Time Warping (DTW) measure through experiments using the Nikkei stock average. Results on a morning star pattern being known as a powerful reversal pattern show that the proposed LCS algorithm finds the results as expected. However, from the viewpoint of investors, the proposed algorithm has room for improvements.

CCS CONCEPTS

• **Information systems** → Information retrieval → Evaluation of retrieval results • **Theory of computation** → Theory and algorithms for application domains → Database theory → Incomplete, inconsistent, and uncertain databases

KEYWORDS

Dynamic programming, Longest common substring, Dynamic time warping, Nikkei stock average, Technical analysis, Prediction of future stock price

iiWAS '17, December 4–6, 2017, Salzburg, Austria

ACM ISBN 978-1-4503-5299-4/17/12…$15.00
https://doi.org/10.1145/3151759.3151820

1 INTRODUCTION

With the recent development of internet technology and digitization of financial transactions, a huge amount of data concerning finance are daily generated and accessible on the internet. There is an increasing demand for predicting financial indicators such as stock prices using the data. The successful prediction of a future stock price could yield not only significant profit for an investor but also bring more investors to stock market stabilizing the market and leading the growth of country's economy.

Fundamental analysis assumes that the price of a stock depends on its intrinsic value, company's past performance, investor sentiment, government actions, and the global economy. Fundamental analysis generally takes rather long time before having a prediction. Because of this analyzing processes, most people believe that fundamental analysis is not likely suitable for short- and medium-term prediction.

Technical analysis tries to predict the immediate future stock price through studying on the trends of past stock data. It is based on the assumption that investors collectively repeat the investment behavior because they participate the market sharing the same objectives, i.e., getting profit. Because investors' objectives remain the same, technical analysts believe that recognizable price patterns will repetitively develop on a chart. History of technical analysis began in the 17th century in Europe [1]. In Asia, technical analysis is developed by a rice trader, Munehisa Homma, during the 18th century. His works evolved the use of candlestick techniques, and led to the "Japanese candlestick chart patterns [1,2]" or "candlestick charting" for short. Though the candlestick charting cannot predict the future stock price in a strict sense, it can at least help to identify trading opportunities. Candle stick patterns are now widely used by technical analysts along with statistical techniques such as the moving average [1]. The candlestick charting proposes several chart patterns for bullish/bearish trends, top/bottom reversals, and continuation of price trends. The major shortcoming of the chart patterns is subjective or lack of measured predictive capabilities. Not to mention, the capabilities of the chart patterns to predict the future market trend are crucial for providing adequate guideline for investors.

In this paper, we propose a systematic method to estimate capability of candle stick patterns in the sense of predictability of the future stock price. The proposed method is applied to the Nikkei stock average to evaluate the effectiveness. The contributions of this paper are as follows:

(I) The longest common substring(LCS) algorithm [3] which is one of dynamic programming methods [4] is improved to deal with similar numeric sequences;

(II) The proposed longest common substring (LCS) algorithm is compared with the Dynamic Time Warping (DTW) [5] measure through experiments using the Nikkei stock average.
(III) A visualized evaluation method is devised for evaluating the degree of the predictability of the future stock prices.

The remainder of the paper is organized as follows. Section 2 gives the review of the relevant literature using data mining techniques to predict the price and the trend of stocks. Section 3 provides overviews of the proposed mining algorithm. Section 4 describes experimental results using the Nikkei stock average. Section 5 concludes the paper with our plans for future work.

2 RELATED WORK

Technical analysis [1] is a process of predicting the future direction of prices through the study on historical data. Technical analysis mainly focuses on the chart and quantitative data with the belief that the history tends to repeat itself.

Several studies have been conducted to find out the way to predict the future stock price and the market trends.

Paliyawan, P. [6] proposes an approach based on chart patterns recognition by using data mining classification. Time series forecasting is conducted to examine a suitable span of time for the stock market data. Important chart patterns to support decision making in the Stock Exchange of Thailand are found out.

Al-Radaideh Q. A et al. [7] present an approach to help investors decide the better timing for buying and selling stocks based on knowledge extracted from the historical data. The decision is made based on decision tree classifier which is one of the data mining techniques. They perform experiments using historical stock data listed in Amman Stock Exchange (ASE).

Alkhatib, K. et al. [8] apply k-nearest neighbor(kNN) algorithm and non-linear regression approach in order to predict stock prices. The kNN algorithm selects k recodes of training data set that are closest to the unknown records. Using the sample data on the Jordanian Stock Market with k=5, the experimental results show that kNN is viable for stock predictions with small error ratio.

Hadavandi, E., et al. [9] present an integrated approach based on genetic fuzzy systems(GFS) and artificial neural networks (ANN) for constructing a stock price forecasting expert system. Results show that the proposed approach outperforms previous fuzzy-based neural network methods.

Enke, D. et al. [10] discuss an approach that uses data mining methods and neural networks for forecasting stock market returns. The authors examine the effectiveness of the neural network models used for level estimation and classification.

3 SIMILARITIES OF NUMERIC SEQUENCES

3.1 Dynamic Programming

Dynamic programming (DP) is a method for solving a complex problem by breaking it down into a collection of simpler subproblems. Given a problem to solve, the key idea behind dynamic programming is to solve different parts of the subproblems, then combine the solutions of the subproblems to reach an overall solution.

Dynamic programming can be classified into the top-down approach and the bottom-up one. The longest common subsequence (LCS) and the dynamic time warping (DTW) algorithms are categorized as the bottom-up approach.

3.2 The longest common subsequence(LCS)

Finding the longest common subsequence (LCS) problem is defined as follows. Given two sequences, find the length of longest subsequence present in both of them. A subsequence is a sequence that appears in the same relative order, but not necessarily contiguous. For example, the LCS for sequences "ABCGBDAZEB" and "AXGAEQ" is "AGAE."

LCS algorithm: Let the input sequences be X[1 … m] of length m and Y[1 … n] of length n. X[0] and Y[0] mean the empty sequences. Let D[i, j] denote the length of the longest common subsequence of X[i] and Y[j].

A) If either sequence is empty, then the LCS is empty, i.e., D[i, 0] = 0 and D[0, j] = 0 for $0 \leqq i \leqq m$ and $0 \leqq j \leqq n$.
B) If X[i] and Y[j] match (X[i] = Y[j]), then the LCS is become longer than the previous sequences by one, i.e., D[i, j] = D[i−1, j−1] + 1.
C) If X[i] and Y[j] do not match (X[i] ≠ Y[j]), then the LCS is the maximum of the previous sequences, i.e., max(D[i − 1, j], D[i, j − 1]).

3.3 LCS for numerical subsequences (nLCS)

The LCS algorithm is originally developed for the sequence of characters or strings. In this study, we improved the LCS algorithm in order to deal with longest common *numerical* subsequences using a tolerance given by a user. If the difference of two numbers of two comparing subsequences is not greater than the tolerance, the two numbers are regarded as the same. The LCS algorithm for strings is easily improved the algorithm for numerical sequences by replacing (X[i] = Y[j]) with $|X[i] - Y[j]| \leqq diff$ where *diff* is a tolerance given by a user.

3.4 Dynamic Time Warping (DTW)

The dynamic time warping (DTW) is a dynamic programming algorithm originally developed for speech recognition. It features to align two sequences by warping the time axis iteratively by minimizing the distance between the two. DTW is introduced into data mining community as a utility for various tasks for series arranged by time. The DTW is the well-known solution for time series problems in a variety of domains, including real-time speech recognition and online signature recognition. Kinlay, J. [5] proposes to apply the DTW to historic stock price sequence analysis.

DTW algorithm: Let the input numerical sequences be C[1 … m] of length m and Q[1 ... n] of lengths n. C[0] and Q[0] mean the empty sequences. Let D[i, j] denote the minimum distance of the subsequences C[i] and Q[j]. Let dist(C[i], Q[j]) denote the distance between elements of C[i] and Q[j].

A) If either sequence is empty, then the DTW is infinity because no sequences match the empty sequence, i.e., D[i, 0] = ∞ and D[0, j] = ∞ for $0 \leqq i \leqq m$ and $0 \leqq j \leqq n$.
B) If both sequences are empty, then D[0, 0] is zero since they are match completely.
C) For $1 \leqq i \leqq m$ and $1 \leqq j \leqq n$, D[i, j] is calculated using dynamic programming to evaluate the following recurrence.
D[i, j]= dist(C[i], Q[j]) + minimum{ D(C[i−1], Q[j−1]), D(C[i−1], Q[j]) , D((C[i], Q[j−1]) }

dist(C[i], Q[j]) defines a distance between C[i] and Q[j]. In this study, we define dist(C[i], Q[j]) = |C[i] − Q[j]|, i.e., absolute difference. D[m, n] gives the DTW of the two numeric sequences.

4 EXPERIMENTS USING THE NIKKEI STOCK AVARAGE

4.1 Data Conversion

Since stock prices change dynamically, the prices are converted to the ratio of the prices. The conversion is rational for calculating relative price. The formula below is used in the conversion.

$$R_i = (CP_i - CP_{i+1})*100 / CP_{i+1} \quad (1 \leqq i \leqq n) \qquad (1)$$

CP_i indicates the closing price of the *i*-th business date. CP_1 means the closing price of the current date. R_1 is the ratio in percentage of the closing price of the current date and the day before the current date. Since stock prices are converted into percentage, the conversion reduce the influence of the highness or lowness of the prices. n is 1,816 in the experiment performed in the study.

4.2 Experiments Only Using DP-Similarity Measures

At the first stage, experiments only using the longest common numerical subsequence (nLCS) and the dynamic time warping (DTW) measures are conducted. These experiments purely correspond to the conditions of the "Japanese candlestick chart patterns." The results are discouraging. Although retrieved stock price sequences are similar before the specified period of the reference date, trends of the price sequences after the reference date are seemed to be random. The results mean poor prediction capabilities of the nLCS and DTW.

Analyses of the results show that the randomness occurs due to the relative position among the stock price of the reference date, the 5-day moving average and the 25-day moving average. In the following, experimental results using moving averages are discussed.

The 5-day moving average is calculated by the latest five days' closing prices. Because these prices are just a sample of larger population of closing prices, the estimated (or sample) standard deviation [11] is taken for a measure of threshold to decide whether a given 5-day moving average is within an expected distribution.

Let *diff, diff_5av* and *diff_25av* denote tolerances for a closing price, the 5-day moving average and the 25-day moving average of a given reference date, respectively. *diff_5av* and *diff_25av* are calculated by the following formulas that are in accordance with Bessel's correction [11].

$$diff\_5av = diff / \mathrm{SQRT}(4) = diff / 2 \qquad (2)$$

$$diff\_25av = diff / \mathrm{SQRT}(24) = diff / 4.89898 \qquad (3)$$

4.3 Experiments Using DTW and Moving Averages

The following experiments are conducted concerning the reference date of April 21, 2017. Fig. 1 shows the candlestick chart of the Nikkei stock average before and after ten days of the reference date. The candlestick chart patterns around April 21, 2017 are considered as the "morning star" pattern. When found in a period of lower trading, this pattern can indicate a strong reversal signal to uptrend [1,2].

Fig. 2 summarizes the result of experiences using the DTW measure and moving averages. The periods of stock price sequences range from three to ten. Because the DTW measure can be calculated for all price sequences of specified periods, the DTW algorithm returns 1,791 (= 1,816 − 25) values as the result of the DTW calculation. The results are sorted in ascending order. X-axis of Fig. 2 indicates the number of the DTW measures selected from the smallest DTW. Y-axis indicates the average DTW measure that is calculated by dividing the sum of the DTW measures by the number of DTW measures, i.e., the value of X-axis.

Length of the stock price sequence that begins on April 21, 2017 varies three to ten as indicated "Period 3" and "Period 4", etc. Fig. 2 shows that the DTW measures per business day increase slowly as a whole. This means the DTW measures are not good at identifying similar stock price sequences. We find that several significant stock price sequences are ranked in insignificant positions.

Figure 1: Candlestick chart involving morning star pattern.

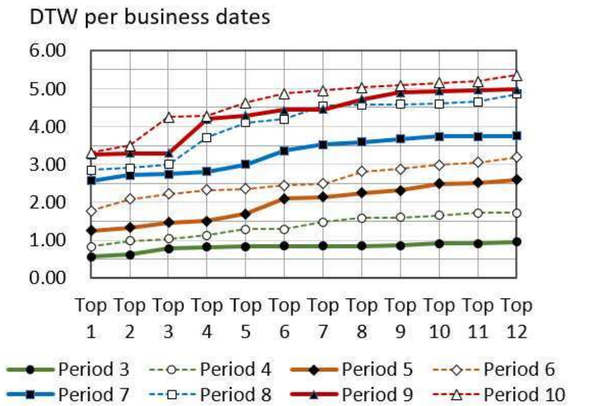

Figure 2: Summary of experiences using the DTW measure and moving averages.

Fig. 3 shows the retrieved top-five business dates that are judged similar to April 21, 2017 by the DTW measure and moving averages. We find eight "20170830" (Aug. 30, 2017) instances, seven "20120719" (July. 19, 2012), five "20170831" (Aug. 31, 2017), four "20170911" (Nov. 11, 2017) and four "20110614" (June 14, 2011), etc.

	Period 3	Period 4	Period 5	Period 6	Period 7	Period 8	Period 9	Period 10
Top 1	20120529	20120529	*20160412*	**20170831**	***20120719***	**20170830**	**20170830**	**20170830**
Top 2	20131010	**20170830**	20110614	20170911	**20170831**	***20120719***	**20170831**	20170911
Top 3	20121004	*20160412*	20110901	**20170830**	**20170830**	20110614	***20120719***	20110614
Top 4	**20170830**	***20120719***	**20170830**	***20120719***	20170911	**20170831**	20110518	***20120719***
Top 5	***20120719***	20110901	**20170831**	20151229	20110614	20170410	20170911	20170815

Figure 3: Retrieved top-five business dates.

Fig. 4 and Fig. 5 show the candlestick charts on Aug. 31, 2017 and July 19, 2012, respectively. Both charts show downtrends instead of a bullish uptrend as shown in Fig. 1.

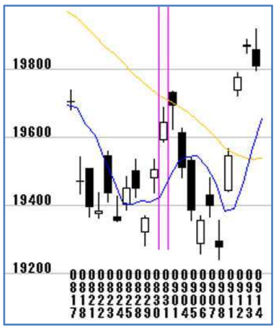

Figure 4: Candlestick chart on Aug. 31, 2017.

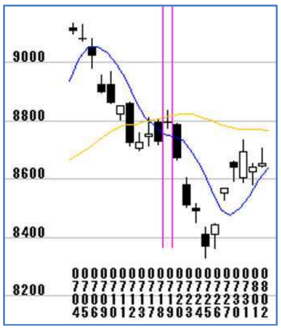

Figure 5: Candlestick chart on July 19, 2012.

Fig. 6 shows the candlestick chart before and after ten days on Nov. 11, 2017, while Fig. 7 shows the candlestick chart on Apr. 12, 2016. Both candlestick charts show strong bullish uptrends as shown in Fig. 1. However, Nov. 11, 2017 (Fig. 6) is retrieved four times and is ranked fifth place in the number of retrieved instances. As for Apr. 12, 2016 (Fig. 7), it is retrieved only twice and is ranked seventh place.

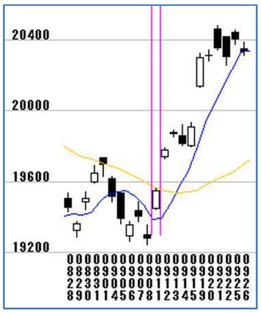

Figure 6: Candlestick chart on Nov. 11, 2017.

Figure 7: Candlestick chart on Apr. 12, 2016.

Charts shown in Fig. 4 to Fig. 7 suggest that the DTW algorithm is of limited capability to identify charts with the same trend when the length of the stock price sequence is smaller than 10.

4.3 Experiments Using nLCS and Moving Averages

4.3.1 Four days to April 21, 2017. Four candlesticks enclosed by a dashed rectangle in Fig. 1 shows an approximate morning star pattern. The typical morning star pattern consists of three candlesticks, i.e. a large black body candlestick followed by a small body and then a large white body candlestick. The third and the forth candlesticks can be combined to form a large white body candlestick. We take the combination of candlesticks and interpret the four candlesticks in Fig. 1 as a morning star pattern.

Fig. 8 summarizes the result of experiences using the nLCS measure and moving averages on April 21, 2017 as the reference date of a period of four days. The ratios of the number of partial matches to the total number of matches increase as the tolerances for a closing price, i.e., *diff*, decrease. The fact shows that partial matching using the nLCS works effectively.

diff	Match	Full Match	Partial Match	Ratio of P.M.
1.8	21	14	7	0.333
1.7	16	10	6	0.375
1.6	10	5	5	0.500
1.5	8	3	5	0.625
1.4	7	3	4	0.571
1.3	6	2	4	0.667
1.2	3	1	2	0.667

Figure 8: Result of retrieval using nLCS on four candlesticks.

Fig. 9 shows the retrieved business dates, the DTW measure and the nLCS, i.e., the number of matches, for *diff* of 1.3% in Fig. 8. It lists the six retrieved business dates with four of those are partially matched. The ratio of nLCS is calculated by dividing the nLCS by the period, i.e., four. The list is sorted in ascending order by the DTW measure. Though the ratio of nLCS is lower in accuracy than the DTW measures, the reciprocals of the nLCS measures are roughly related to the DTW measures.

Date	DTW	nLCS	Ratio of nLCS
20170421	0.000	4	1
20110901	1.290	4	1
20120907	2.766	4	1
20130902	3.554	3	0.75
20101104	3.886	3	0.75
20110915	5.428	3	0.75
20141218	5.701	2	0.5

Figure 9: Result of retrieval for *diff* of 1.3%.

Fig. 10 shows overlapped closing prices in percentage that are retrieved using the nLCS measure and moving averages. All reference dates are aligned on the origin to make the comparison easy.

The thick black line correspond to the reference date, i.e., April 21, 2017. Thin solid lines indicate the stock prices of retrieved dates in percentage. These lines correspond to closing price sequences of business dates listed in Fig. 9 except for the reference date. The thick light blue line indicates the average of the candlestick charts plotted by thin solid lines.

Three out of five closing price sequences suggest an uptrend, while two out of five price sequences suggest downtrend. It can be reasonable to say that no trade judgments, i.e., indecision, should be made based on the retrieval.

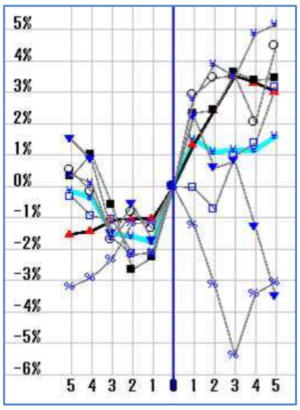

Figure 10: Prediction of future stock trend.

4.3.2 Five days to April 24, 2017. Five candlesticks enclosed by a solid rectangle in Fig. 1 shows an approximate morning star pattern plus one confirmation day. Fig. 11 summarizes the result of experiences on five days to April 24, 2017.

diff	Match	Full match	Partial Match	Ratio of P.M.
1.8	16	7	9	0.563
1.7	12	3	9	0.750
1.6	8	2	6	0.750
1.5	7	1	6	0.857
1.4	5	1	4	0.800
1.3	5	1	4	0.800
1.2	3	1	2	0.667
1.1	2	0	2	1.000

Figure 11: Result of retrieval using nLCS on five candlesticks.

Fig. 12 shows the retrieved business dates, DTW measure and the number of matches for *diff* of 1.3%. It lists the five retrieved business days with four of those days are partially matched. The reciprocals of the nLCS measures are almost related to the DTW measures, which is also shown in Fig. 9.

Date	DTW	nLCS	Ratio of nLCS
20170424	0.000	5	1
20131015	1.865	5	1
20150513	3.621	4	0.8
20151006	4.805	3	0.6
20140522	4.883	4	0.8
20140814	6.937	3	0.6

Figure 12: Result of retrieval for *diff* of 1.3%.

Fig. 13 shows overlapped closing prices whose dates are listed in Fig. 12. Four out of five closing price sequences suggest uptrend that coincides with the bullish trend of the price sequence to April 24, 2017 in Fig. 1.

As for retrieving similar price sequences and predicting the future trends, the proposed algorithm finds the results as expected. However, from the viewpoint of investors, the result is moderate since the uptrend is almost over.

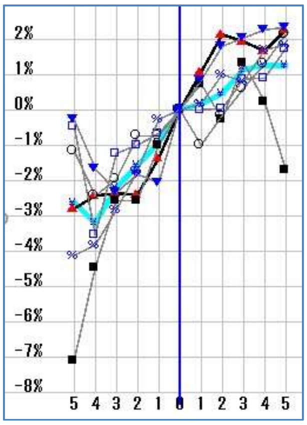

Figure 13: Prediction of future stock trend.

5 CONCLUSIONS

Predicting the stock prices in the near future is of great interest among investors because successful prediction of stock prices may promise attractive benefits. We propose an improved longest common substring (LCS) algorithm for numeric sequences, named nLCS, to retrieve similar stock price patterns for the prediction. The results for the stock market prediction is evaluated through overlapped charts that are retrieved using the proposed algorithm. The comparative analyses are carried out on the other dynamic programming algorithm named the dynamic time warping (DTW) measure. The obtained results show that the proposed approach produces better results than the DTW measure in terms of prediction of future stock prices.

As for the future work, we are planning to improve the proposed algorithm by applying it to various candlestick patterns known as bullish/bearish ones and continuation ones. Comparative experiences using international stock indexes such as DJIA (Dow Jones Industrial Average), FTSE 100, EURONEXT 100, DAX and ASX 200 will be attractive research themes to identify the characteristics of the markets.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their invaluable feedback. This research is supported by the JSPS KAKENHI grant number 16K00161.

REFERENCES

[1] 2011. Technical Analysis, Cambridge Univ. (Feb. 2011), 1-179, http://www.mrao.cam.ac.uk/~mph/Technical_Analysis.pdf.

[2] Rose, M. 2011. Everything you wanted to know about candlestick charts, Thames Publishing Ltd. (2011), 1-22.

[3] Mount, D. 2010. Dynamic Programming: Longest Common Subsequence, CMSC 451, Maryland Univ. (Sept. 2010), 1-5, https://www.cs.umd.edu/ class/fall2010/ cmsc451/lect451.pdf.

[4] Park, J. 2015. Dynamic Programming, CS 97SI, Stanford Univ. (June 2015), 1-36, https://web.stanford.edu/ class/cs97si/04-dynamic-programing.pdf.

[5] Kinlay, J. 2016. Dynamic Time Warping - QUANTITATIVE RESEARCH AND TRADING (Aug. 2016), 1-11, http://jonathankinlay.com/2016/ 08/dynamic-time-warping/.

[6] Paliyawan, P. 2015. Stock market direction prediction using data mining Classification, Journal of Engineering and Applied Sciences, Asian Research Publishing Network (ARPN), 10, 3 (Feb. 2015), 1302-1310.

[7] Al-Radaideh, Q. A., Assaf, A.A. and Alnagi, E. 2013. Predicting Stock Prices Using Data Mining Techniques, Proceedings of the International Arab Conference on Information Technology (ACIT'2013), 1-8.

[8] Alkhatib, K., Najadat, H., Hmeidi, I., and Ali Shatnawi K. M. 2013. Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm, International Journal of Business, Humanities and Technology, 3, 3 (March 2013), 32-44.

[9] Hadavandi, E., Shavandi, H., and Ghanbari, A. 2010.Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting, Knowledge-Based Systems, 23, 8 (Dec. 2010), 800-808.

[10] Enke, D., Thawornwong, S. 2005. The use of data mining and neural networks for forecasting stock market returns, Expert Systems with Applications, 29, 4 (Nov. 2005), 927-940.

[11] Smith, T. G. 2016. Machine Tool Metrology: An Industrial Handbook, ISBN 978-3-319-25109-7, Springer Nature, 2016.