# Processing of Kuala Lumpur Stock Exchange Resident on Hadoop MapReduce

**H. Law**[1], **S. Aghabozorgi**[1], **S. Lim**[1], **Y. Teh**[1] **and T. Herawan**[1]
[1]Department of Information System, Faculty of Computer Science & Information Technology, University of Malaya, Kuala Lumpur, Malaysia

**Abstract -** *The Kuala Lumpur Stock Exchange (KLSE) is the big data that need to be stored, processed and analyzed as it trade day-to-day. Analyzing and finding similar components (stock market price) may assist investor. However, it is not easy to find the similar components in the KLSE. This is because the components in the KLSE are changing everyday in the market. This paper focus is on using the Hadoop MapReduce to store and process the KLSE big data, use the k-means algorithm to perform the calculation, then find the companies that had similar KLSE closing bids pattern to help the investors to predict a company's next closing bid based on another company that have the similar trends. To facilitate the investors, the similar trend among companies will be shown on the Graphical User Interface (GUI). All the storing, processing and analyzing will be run automatically behind the scene of the GUI.*

**Keywords:** Kuala Lumpur Stock Exchange (KLSE), Hadoop, MapReduce, closing bids, pattern, Graphic User Interface (GUI).

## 1 Introduction

KLSE is formerly known as Kuala Lumpur Stock Exchange and is established in the year 1964. After that, it has been renamed to Bursa Malaysia in the year 2004. Bursa Malaysia is an exchange holding company approved under Section 15 of the Capital Markets and Services Act 2007. It operates a fully integrated exchange, offering the complete range of exchange-related services including trading, clearing, settlement and depository services that are traded on day-to-day. The Prices of the trade are determined by the market forces. The buyers and sellers quote the bid and ask prices and if prices are matched, in the case of KLSE, by its automated trading. Due to the KLSE trade is carried out every day, so there is a dynamic data for the KLSE day-by-day. This big data need to be stored, processed and analyzed so that investors able to see the trend of the stock exchange, and they able to identify when and what stocks to buy and sell, by aware the track of upswings and downswings over the history of one's company according to the sector.

The BigData is the data sets that are large in volume, high velocity, and is complex with variety information assets. The Big Data is in petabytes that consists of billions to trillions of records of millions of people from the different sources such as sales, social media like Facebook, Twitter, patients' record, mobile data, digital pictures and video and more. The Big Data is simply a matter of size[1].

The MapReduce is the programming paradigm for processing large data sets. It consists of two functions, the map function and the reduce function. The map function responsible to partitioning every request into smaller request which are sent to many server, while the reduce job responsible to processing the smaller request using the algorithm provided, and give the best output result to the user.

However, it is not easy to analyze and finding the similar components (stock market price) in the KLSE. This is because the components in the KLSE are changing every day in the market. Hence, the components are highly dynamic to determine the similar trend of the stock prices. The similar trend example is if one's component goes up, the other company's component will also go up as well, and vice versa. Hence, this paper proposed is to use the k-means clustering to determine the similarities among companies, then by using the companies past history of the stock time-series to predict the future closing bids of a company.

The rest of the paper is organized as follows. In section 2, we review the installation of the Ubuntu operating system, in order to use and run the HadoopMapReduce. Section 3 is to review about the k-means algorithm and how this algorithm work in determining the similarities among companies, the Section 4 review about the Graphical User Interface (GUI) in assisting the investor to see a company trend and the k-best similar companies. Finally, in Section 5, we offer and suggest the direction for future work as conclusion.

## 2 Literature review

The categorization of companies in the stock market is very useful for managers, investors, and policy makers. It can be performed based on several factors, such as the size of the companies, their annual profit, and the industry category. However, these features usually change over the course of time; thus, they are improper for categorization purposes. Industry-based categorization is also not preferable due to

evidence that financial analysts are biased by industry categorization [2]. Identifying homogeneous groups of stocks where the movement in one market affects the stock prices in another market. The literature shows that the similarity of stock market in a country is affected by the movement of other stocks in that country or in other regions [3]–[5]. As a result, numerous studies have been performed on the recognition of co-movements among different countries [6]–[8]. Most of these studies consider the co-movement of the stock market between different reigns or countries but not among different industries or companies in a stock market.

Assessment of the stock market similarity among companies in a stock market (e.g. the Kuala Lumpur stock market) can be very helpful for predicting the stock price, based on the similarity of a company to other companies in the same cluster. Based on the others literature review [9]–[17] on the KLSE stock prices forecasting, it could be notice that most of the existing stock market prediction system is just to forecast the further movement or next stock bids by looking at the company past history, by using the Artificial Neural Network models or the neural network prediction, without referring to any company that have similar trends with it. Therefore, the next bid prediction by the system may forecast inaccurately.

In the time series literature review, [18]–[26] , the author tries to cluster the time series of data efficiently by customer segmentation and developing a novel method for clustering time series incrementally based on its ability to accept new time series and also able to update the underlying clusters. While in the other time series literature review [21], the author stated the significant problem of traditional clustering – defining prototype and explained the benefits of the proposed prototype by customer transaction clustering as well as present a prototype for time series clusters efficiency that can be updated based on a fuzzy concept through iterations.

There are several numbers of literatures that has been published about the BigData and Hadoop as well as the stock market over the Internet. Among these publications, one of the literatures is about Evaluation of Data Processing Using MapReduce Framework in Cloud and Stand-Alone Computing [27]. This article described about the comparison of the data processing speed and time in the cloud computing environment and the stand alone system environment. To establish the experiment, the authors compare and concluded that the Linux environment is more suitable to develop the MapReduce than the windows as the windows had problem connection to a distributed cluster. [27].

# 3 Environment Setup

Firstly, before storing and processing the KLSE big data, the installation and configuration for the Hadoop MapReduce in the personal computer (stand alone system) are needed. From the above literature review [28], it is determined that the Hadoop MapReduce is more suitable to install on the Linux environment than the windows environment. To provide a test platform on the windows environment, it is recommended to install the Ubuntu operating system version 10.4 Long Term Support in the personal computer in order to run the Hadoop MapReduce. This Ubuntu Long Term Support operating system is a complete desktop Linux-based operating system that allows the Linux application to be compiled and run on a windows operating. The installation of the Ubuntu operating system enables the Hadoop MapReduce to run on the windows laptop over the Ubuntu. After installation of the Ubuntu operating system, the Hadoop MapReduce in the Ubuntu operating system needs to be configured before it can be used by executing the command[29]. Then, the KLSE stock market price components can be loaded into the Hadoop MapReduce, and user needs to key in the Java coding to extract the desired data such as company name, date and closing bids of the KLSE as the output.

# 4 K-Means Algorithm

The extracted output from the Hadoop MapReduce will be passing to the k-means clustering for further analysis by performing series of calculation on the closing bids, to determine the similarities among companies. Conventional clustering and similarity measures which are applied to static data are not practical for the time-series datasets because they are essentially not designed for time-series data. Hence, various techniques have been recommended for the clustering of time-series data. Most of them try to customize the existing conventional clustering algorithms such that they become compatible with the nature of time-series data. In this cases, usually the distance measure is modified to be well-matched with the time-series data [30].

K-means clustering is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The main idea is to define the k centroids, one for each cluster. In this paper, the following example illustrates the k-means clustering and forecasting using a simulated data set containing a time series components. The reasons of choosing the k-means clustering and algorithm as big data analytics and decision making is because the KLSE big data analysis is within one year time series and its focus only for the closing bids. It is focused on the closing bid because the closing bids are the most real data of the day and this closing bid will be brought to the next day's open bids. The figure 1 below shows the whole process of clustering.

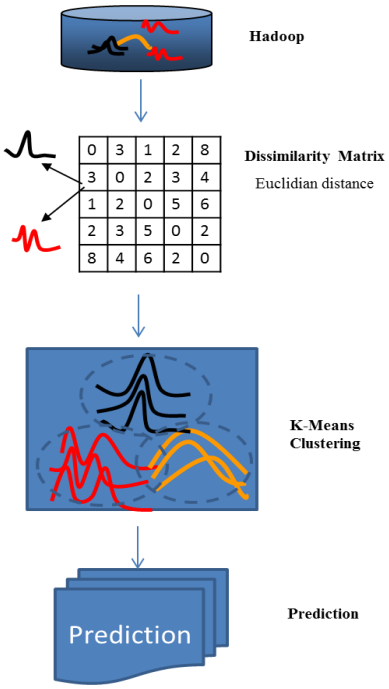


Figure 1. The clustering process

## 4.1 The Identification of clustering stage

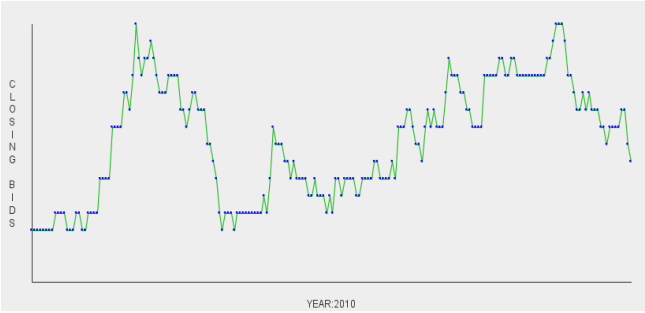


Figure 2. Simulated ARIMA Series KLSE components (From Jan 04, 2010 to Dec 14, 2010).

The starting of the identification stage is to specify the input data set in the k-means clustering. The input data set is the KLSE components. Then use an identify statement to read the KLSE close bids in time series and plot a graph. The graph that has been plotted is shown in the figure 2 above and the table 1 of the data below shows the example of KLSE components data set.

| Ticker | Per | Date | Open | High | Low | Close | Vol | O/I |
|---|---|---|---|---|---|---|---|---|
| AFG | D | 12/14/10 | 0.23 | 0.24 | 0.23 | 0.23 | 0 | 0 |
| AFG | D | 12/13/10 | 0.24 | 0.24 | 0.23 | 0.24 | 512 | 0 |
| AFG | D | 12/10/10 | 0.25 | 0.26 | 0.25 | 0.26 | 3711 | 0 |
| AFG | D | 12/09/10 | 0.26 | 0.26 | 0.26 | 0.26 | 88 | 0 |
| AFG | D | 12/08/10 | 0.25 | 0.25 | 0.25 | 0.25 | 88 | 0 |
| AFG | D | 12/06/10 | 0.25 | 0.25 | 0.25 | 0.25 | 0 | 0 |
| AFG | D | 12/03/10 | 0.25 | 0.25 | 0.25 | 0.25 | 100 | 0 |
| AFG | D | 12/02/10 | 0.25 | 0.25 | 0.25 | 0.25 | 167 | 0 |
| AFG | D | 12/01/10 | 0.24 | 0.25 | 0.24 | 0.24 | 0 | 0 |

Table 1. Company AFG's components data set.

## 4.2 Estimation and diagnosis checking stage

The estimate statement next prints a table of correlations of the parameter wanted, as shown on the table 2 below.

| Ticker | Date | Close |
|---|---|---|
| AFG | 12/14/10 | 0.23 |
| AFG | 12/13/10 | 0.24 |
| AFG | 12/10/10 | 0.26 |
| AFG | 12/09/10 | 0.26 |
| AFG | 12/08/10 | 0.25 |
| AFG | 12/06/10 | 0.25 |
| AFG | 12/03/10 | 0.25 |
| AFG | 12/02/10 | 0.25 |
| AFG | 12/01/10 | 0.24 |

| Ticker | Date | Close |
|---|---|---|
| ASTINO | 12/14/10 | 0.62 |
| ASTINO | 12/13/10 | 0.62 |
| ASTINO | 12/10/10 | 0.62 |
| ASTINO | 12/09/10 | 0.63 |
| ASTINO | 12/08/10 | 0.62 |
| ASTINO | 12/06/10 | 0.62 |
| ASTINO | 12/03/10 | 0.62 |
| ASTINO | 12/02/10 | 0.63 |
| ASTINO | 12/01/10 | 0.62 |

Table 2. Company AFG's close bids and company ASTINO's close bids are extracted.

When the output is extracted from the Hadoop MapReduce, then use formulas to perform the calculation to calculate the entire closing bids distances between companies.

$$\text{Distance } (t_1, t_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$= \sqrt{\begin{matrix}(0.62-0.23)^2+(0.62-0.24)^2\\+(0.62-0.26)^2+(0.63-0.26)^2\\+(0.62-0.25)^2+(0.62-0.25)^2\\+(0.62-0.25)^2+(0.63-0.25)^2\\+(0.62-0.24)^2\end{matrix}}$$

$$= \sqrt{\begin{matrix}0.1521+0.1444+0.1296\\+0.1396+0.1396+0.1396\\+0.1396+0.1444+0.1444\end{matrix}}$$

$$= \sqrt{1.2733}$$

$$= 1.128$$

When the distances between the two companies are known, next is to normalize the distance to the values between 0 and 1 for the standardization purpose.

$$\text{Normalized Distance} = \frac{\text{Distance } (t_1,t_2)}{\sum(x_1+y_1+\ldots+z_1)}$$

$$= \frac{1.128}{\begin{matrix}0.23+0.24+0.26+0.26+0.25\\+0.25+0.25+0.25+0.24\end{matrix}}$$

$$= \frac{1.128}{2.23}$$

$$= 0.506$$

When the distance values between companies had been standardized, the similarities between the companies can be determined.

Similarities ($t_1$, $t_2$) = 1 – Normalized Distance
= 1 – 0.506
= 0.494
= 0.49

After a series of calculation, it can be seen that the similarities between both company AFG and company ASTINO are 0.49. Therefore, it can be concluded that the smaller the similarities ($t_1$, $t_2$) between both companies, the both companies' trends are similar, in contrast, the larger the similarities ($t_1$, $t_2$) between both companies, the both companies' trends are not similar. From the similarities ($t_1$, $t_2$) between company AFG and company ASTINO, it can be concluded that the both companies are neither similar nor not similar. We expect to see the clusters as shown in the figure 3 below.

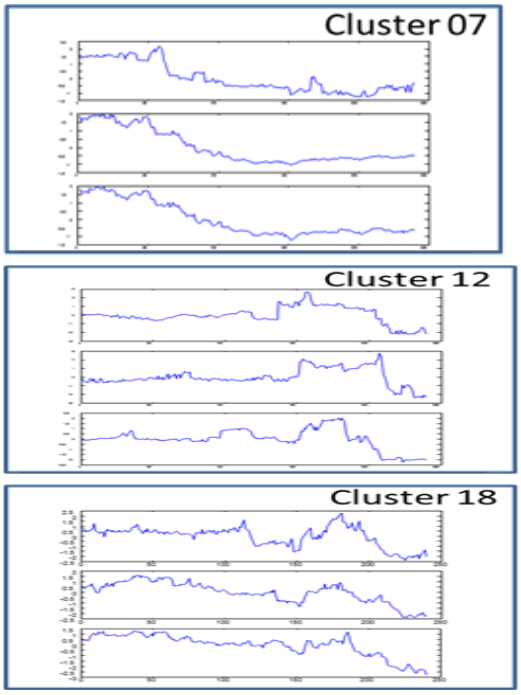


Figure 3. Three sample stock market of three clusters of KLSE datasets

### 4.3 Forecasting of the KLSE stock prices Stage

After get done in finding the similarities($t_1$, $t_2$) between companies, it is suitable to categorized or rearrange the companies that have the most similar trend to less similar trend based on one's company. For an example: the company that has the most similar trend with Maybank are the CIMB bank (most similar trend), followed by the RHB (similar trend), Public bank (similar trend) and Ambank (less similar trend). To produce forecast, company A's next bid will be predicted based on the other company such as company H that has most similar trend with company A because they have the similarity shape of the stock price or they are co-movement that move together in the same trend. The figure 4 below shoes the daily stock price index prediction of KLSE in the graph form.

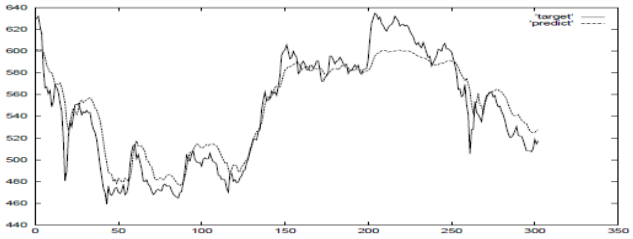


Figure 4. Daily stock Price Index Prediction of KLSE (Out of Sample Data: From July 30, 1990 (horizontal scale 0) to OCT 16, 1991(304))

## 5 GRAPHIC USER INTERFACE (GUI)

In this paper, the user module will be the Java Graphic User Interface (GUI). The purpose of this module is to provide the user selection on their preference company's stock market prices graph and the similar trend of companies with that particular company, then predict the next closing bids accordingly. The GUI performance is shown in the figure 5, 6 and 7 below.


Figure 5. The use interface that allows user to make selection on their desired company and year.

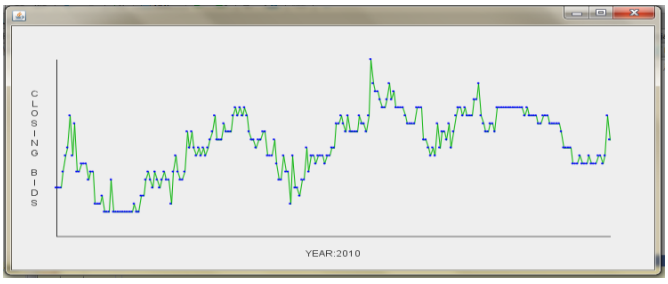

Figure 6. The graph is generated based on the user selection.


Figure 7. The user able to see the k-best similar trend for the selected company.

# 6 Conclusions

The language for the user module is Java. In this paper, for KLSE components as the input of Hadoop MapReduce, the output is the company's closing bids, then passing to the ARIMA model for the series of calculation, and determined the companies similarities. A simple and clear methodology is used to investigate the similar trends of the KLSE for companies. From the calculation, we found out that the series of the calculation should be integrated into one algorithm to facilitate the calculation, and it should be insert it into the Hadoop MapReduce's reducer part, to minimize the time and get the accurate output in the shortest possible of time. This paper's prediction system are useful to the investors in the future as it able to forecast the company's next bid accurately based on the other companies that have similar trend with it.

# 7 Acknowledgment

This work is supported by University of Malaya High Impact Research Grant no vote UM.C/625/HIR/MOHE/SC/13/2 from Ministry of Education Malaysia.

# 8 References

[1] P. Russom, *BIG DATA ANALYTICS (TDWI BEST PRACTICES REPORT)*, FOURTH QUA. TDWI, 2011, p. 40.

[2] P. Krüger, A. Landier, and D. Thesmar, "Categorization Bias in the Stock Market," *Available SSRN 2034204*, 2012.

[3] D. Collins and N. Biekpe, "Contagion and interdependence in African stock markets," *South African J. Econ.*, vol. 71, no. 1, pp. 181–194, 2003.

[4] A. Antoniou, "Modelling international price relationships and interdependencies between the stock index and stock index futures markets of three EU countries: a multivariate," *J. Business, Financ. Account.*, vol. 30, pp. 645–667, 2003.

[5] A. Masih and R. Masih, "Dynamic modeling of stock market interdependencies: an empirical investigation of Australia and the Asian NICs," *Rev. Pacific Basin Financ. Mark. Policies*, vol. 4, no. 2, pp. 1323–9244, 2001.

[6] A. Rua and L. Nunes, "International comovement of stock market returns: A wavelet analysis," *J. Empir. Financ.*, vol. 16, no. 4, pp. 632–639, 2009.

[7] M. Graham and J. Nikkinen, "Co-movement of the Finnish and international stock markets: a wavelet analysis," *Eur. J. Financ.*, vol. 17, no. 5, pp. 409–425, 2011.

[8] L. Norden and M. Weber, "The Coâ€ • movement of Credit Default Swap, Bond and Stock Markets: an Empirical Analysis," *Eur. Financ. Manag.*, vol. 15, no. 3, pp. 529–562, 2009.

[9] J. L. Ford, W. C. Pok, and S. Poshakwale, "The Return Predictability and Market Efficiency of the KLSE CI Stock Index Futures Market," *J. Emerg. Mark. Financ.*, vol. 11, no. 1, pp. 37–60, Mar. 2012.

[10] H. Poh and J. T. Yao, "EQUITY FORECASTING : A CASE STUDY ON THE KLSE INDEX," *Neural Networks Financ. Eng. Proc. 3rd Int. Conf. Neural Networks Cap. Mark.*, pp. 341–353, 1995.

[11] H. Feng and H. Chou, "Evolutionary fuzzy stock prediction system design and its application to the Taiwan stock index," *... J. Innov. Comput. Inf. ...*, vol. 8, no. 9, pp. 6173–6190, 2012.

[12] P. A. Idowu, C. Osakwe, A. A. Kayode, and E. R. Adagunodo, "Prediction of Stock Market in Nigeria Using Artificial Neural Network," *Int. J. Intell. Syst. Appl.*, vol. 4, no. 11, pp. 68–74, Oct. 2012.

[13] B. B. Nair, N. M. Dharini, and V. P. Mohandas, "A Stock Market Trend Prediction System Using a Hybrid Decision Tree-Neuro-Fuzzy System," in *2010 International Conference on Advances in Recent Technologies in Communication and Computing*, 2010, pp. 381–385.

[14] R. P. Schumaker and H. Chen, "A quantitative stock prediction system based on financial news," *Inf.*

*Process. Manag.*, vol. 45, no. 5, pp. 571–583, Sep. 2009.

[15] P.-C. Chang and C.-H. Liu, "A TSK type fuzzy rule based system for stock price prediction," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 135–144, Jan. 2008.

[16] M. B. I. Reaz, S. Z. Islam, M. A. M. Ali, and M. S. Sulaiman, "FPGA realization of backpropagation for stock market prediction," in *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02.*, 2002, vol. 2, pp. 960–964.

[17] P. Gomide and R. L. Milidiu, "Assessing Stock Market Time Series Predictors Quality through a Pairs Trading System," *2010 Elev. Brazilian Symp. Neural Networks*, pp. 133–139, Oct. 2010.

[18] T. W. SR AGHABOZORGI, MR SAYBANI, "Incremental Clustering of Time-Series by Fuzzy Clustering," vol. 688, pp. 671–688, 2012.

[19] S. Aghabozorgi and T. Y. Wah, "Dynamic Modeling by Usage Data for Personalization Systems," *2009 13th Int. Conf. Inf. Vis.*, pp. 450–455, Jul. 2009.

[20] S. Aghabozorgi and T. Y. Wah, "Using Incremental Fuzzy Clustering to Web Usage Mining," in *2009 International Conference of Soft Computing and Pattern Recognition*, 2009, pp. 653–658.

[21] S. Aghabozorgi, T. Y. Wah, A. Amini, and M. R. Saybani, "A new approach to present prototypes in clustering of time series," in *The 7th International Conference of Data Mining*, 2011, vol. 28, no. 4, pp. 214–220.

[22] S. Aghabozorgi and Y. Teh, "Clustering of Large Time-Series Datasets," *J. Intell. Data Anal.*, vol. 18, no. 5, 2014.

[23] S. Aghabozorgi and T. Wah, "Effective Clustering of Time-Series Data Using FCM.," *Int. J. Mach. Learn. Comput.*, vol. 4, no. 2, pp. 170–176, 2014.

[24] S. Aghabozorgi, A. S. Shirkhorshidi, T. Hoda Soltanian, U. Herawan, and T. Y. Wah, "Spatial and Temporal Clustering of Air Pollution in Malaysia: A Review," in *International Conference on Agriculture, Environment and Biological Sciences*, 2014, pp. 213–219.

[25] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. N. C. Ling, "Text Mining for Market Prediction: A Systematic Review," *Expert Syst. Appl.*, 2014.

[26] Saeed Aghabozorgi and T. Y. Wah, "Shape-based Clustering of Time Series Data," *J. Intell. Data Anal.*, vol. 18, no. 5, 2014.

[27] S. Daneshyar and A. Patel, "Evaluation of Data Processing Using MapReduce Framework in Cloud and Stand - Alone Computing," *Int. J. Distrib. Parallel Syst.*, vol. 3, no. 6, pp. 51–63, Nov. 2012.

[28] S. Daneshyar, "Evaluation of Data Processing Using MapReduce Framework in Cloud and Stand - Alone Computing," *Int. J. Distrib. Parallel Syst.*, vol. 3, no. 6, pp. 51–63, Nov. 2012.

[29] T. White, *Hadoop : The Definitive Guide*, 3rd editio. O'Reilly Media / Yahoo Press, 2012, p. 688.

[30] T. Warrenliao, "Clustering of time series data--a survey," *Pattern Recognit.*, vol. 38, no. 11, pp. 1857–1874, Nov. 2005.