Analyzing Stock Markets using Data Warehousing

S. M. Mujtaba* and Mohammed Nadeem
SZABIST
Karachi, Pakistan

Abstract: *Today, almost all the stock markets disseminate data in real time. Real time data is of great importance to day traders who keep track of each and every price movement. However, daily trading data is always important for all investor classes and most importantly, to the technical analysts (people who analyze and predict stocks based on historical data). This paper covers areas of interest in statistics, financial technical analysis and data warehousing. This research work firstly identifies a data warehouse design model which is best suited for stock market technical analysis and secondly provides a framework for technical analysts based on standard tools available in the industry.*

Keywords: *Data Warehousing, Technical Analysis, Time Series Regression, Historical Data.*

1. INTRODUCTION

Traditionally, data has been kept in on-line transaction processing (OLTP) databases or flat files. Applications have been designed to use this raw data and apply programmatic models for analysis. OLTP databases are designed to store (insert, update and delete) the data efficiently, where as for analysis as [1] identifies, the data should be retrieved not only efficiently but effectively as well. On-line Analytical Processing (OLAP) databases and data warehouses help in fast retrieval of information. Data warehouses assist knowledge workers and managers in decision-making and provide information to make informed decisions [2]. Query throughput and response times are more important than transaction throughput [3].

The success of OLAP technology naturally leads to its possible extension from the analysis of static data to that of dynamically changing data, including time-series data, scientific and engineering data, and data produced in other dynamic environments such as network traffic, web click streams, weather or environment monitoring, stock exchange etc [4]. In stock markets, data is stored and disseminated in online and real-time environments, which results in massive amount of data. Such time-ordered data typically can be aggregated with an appropriate time interval, yielding large volume of equally spaced time series data.

A fundamental difference in the analysis of a dynamic environment from that in static one is that the dynamic one relies heavily on regression and trend analysis instead of simple, static aggregates. [5]

Data warehousing and data cube technology has not been seriously applied for storing and analyzing time series data, where as time series analysis has long been one of the most essential data analysis tasks in statistics. It is because the aggregates stored in data cubes are mostly simple aggregates, such as sum, count, max, min, etc. Such aggregates, though good enough for generating summary reports, are not deep or sophisticated enough for time series analysis, including regression analysis etc. [6]

Data warehousing techniques have found very limited application for this business requirement [7]. Though artificial intelligence and neural networks have been used in the analysis of stock market data but have found difficulties in the real world. As both technology and data warehousing techniques continue to improve, it makes it worth while to explore the possibility to analyze and possibly forecast the market data using the available OLAP and data warehousing technology [8, 9].

Traditionally, market analysts have used techniques of fundamental and technical analysis to analyze and predict the behavior of stock prices and stock market indices. Hence we formulate a research question:

"Can we formulate data warehouse models for technical analysis of stock price data with in the paradigm of OLAP technology?"

In this research work, a multi-dimensional star schema OLAP data warehouse is developed to aggregate time series data and data analysis models are applied for predictive time series analysis of stock market data.

2. DATA DEFINITION

In this research work, historical data for KSE100 Index of Karachi Stock Exchange along with five major stocks (HUBC, PTCL, PSO, FAUJI, and MCB) is obtained using raw feed download files from KSE website (www.kse.net.pk) in ZIP format.

During the data cleansing process erroneous data was identified resulting from inconsistencies in the data format of the source files. Erroneous data was corrected after analysis of the data files and modification of initial data extraction process. Figure 1 shows some sample data.

3. BACKGROUND

Use of historical data to predict future performance relies on the assumption that any influences on stock prices are reflected in the price movements. The classical models of

* Doha Stock Exchange

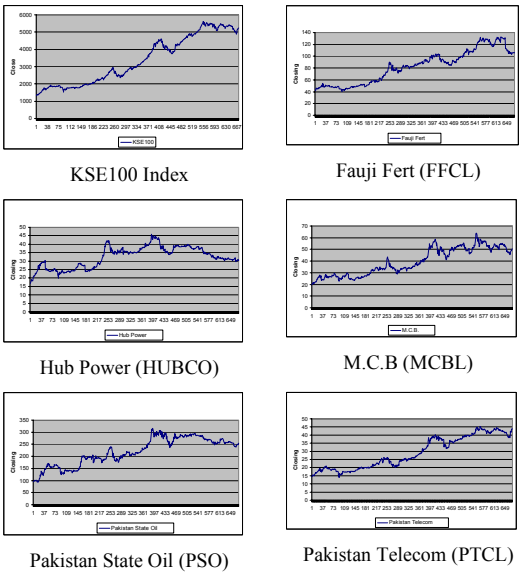

Figure 1: Sample Data

this approach include study of moving averages and regression indicators. They at best are capable of picking out trends in the stock market, but have difficulty in modeling cycles that are by no means repetitive in amplitude, period of shape [10].

The following theories are considered as part of technical analysis for the study.

Moving Average: Moving averages are calculated from historical price information. It shows the average value of a price over a period of time. It represents a smoothing of actual price fluctuations. In flat or consolidating markets, moving averages would closely track the current prices. In trending markets, they can be used in buy and sell decisions.

The general assumption behind all moving averages is that once the stock price moves above the average that it may substantial momentum behind it and is worth buying. The opposite is true if the price of a security moves below the moving average.

The formulas for moving averages are as follows

$$\text{SMA} = \text{Sum of L day's Closing Price} / \text{L}$$

(Source: [11])

Where L is number of days
Valid range for L = 1 to 200 (Usually 20, 30, 50, 100, 200)

$$\text{EMA}_{current} = \{(\text{Price}_{current})\ \text{EMA\%}\} + \{\text{EMA}_{previous}(1 - \text{EMA\%})\}$$

(Source: [12])

Where EMA% is the weight applied to the current day's price

$$\text{EMA\%} = 2 / (n+1)$$

where n is the number of days of EMA's period.

Time Series Regression: There are two main goals of time series regression analysis: (a) identifying the nature of the phenomenon represented by the sequence of observations, and (b) forecasting (predicting future values of the time series variable). Both of these goals require that the pattern of observed time series data is identified and more or less formally described. Once the pattern is established, it can be interpreted and integrated with other data (i.e., use it in theory of the investigated phenomenon, e.g., seasonal commodity prices).

As [13] suggests, the identified pattern can be extrapolated to predict future events that regardless of the depth of understanding and the validity of interpretation (theory) of the phenomenon.

Most time series patterns can be described in terms of two basic classes of components: trend and seasonality [14]. The former represents a general systematic linear or (most often) nonlinear component that changes over time and does not repeat or at least does not repeat within the time range of the data being analyzed. The latter may have a formally similar nature; however, it repeats itself in systematic intervals over time. Those two general classes of time series components may coexist in real-life data.

Linear regression is a mathematical technique used in both technical and fundamental analysis. The technique uses a number of variables to predict some unknown variable. In technical analysis simple regression of price changes over a period of time can also help identify what has been acceptable in terms of valuation levels and project those acceptable levels into the future. Different time periods produce different regression results and can help identify potential price projections when the major long term trends of the market change direction.

Used directly, with an appropriate data set, linear least squares regression can be used to fit the data with any function of the form

$$y = f(x) = a + b.x$$

(Source: [15])

Where 'y' is the variable is to be predicted, 'x' is the independent variable is used to predict 'y', 'a' is the intercept and 'b' is the slope.

Linear least squares regression gets its name from the way the estimates of the unknown parameters are computed. The "method of least squares" that is used to obtain parameter estimates was independently developed in the late 1700's and the early 1800's by the mathematicians Karl Friedrich Gauss, Adrien Marie Legendre and

(possibly) Robert Adrain [16] working in Germany, France and America, respectively.

Multiple regression estimates the outcomes (dependent variables) which may be affected by more than one control parameter (independent variables) or there may be more than one control parameter being changed at the same time.

An example is the two independent variables x and y and one dependent variable z in the linear relationship case:

$$z = a + bx + cy$$

(Source: [17])

The data required for technical analysis is time series data of stock prices, volumes and market indices.

Time series analysis is a special case problem of data warehousing and gives vendors less incentive to develop OLAP based products for technical analysis. However, it is does not eliminate the requirement for efficient solutions to perform technical analysis. It is worthwhile to investigate if OLAP technology can be effectively and efficiently used for technical analysis.

Some research [5, 6] has been done in the past for OLAPing of time series data. However, in [5] the time series data is aggregated and analysis is carried out on the aggregated data. This approach is used to save computing time and memory requirements for cube analysis. However, the trends analyzed using this approach, reflect the data range of aggregated data. This approach does not seem appropriate for technical analysis of stock market data as stock price analysis and prediction require actual values of price and volume in order to determine and predict trading range of a stock.

OLAP stands for Online Analytical Processing. OLAP databases are targeted for decision support. Historical, summarized and consolidated data is required to facilitate complex analysis. Typical OLAP operations include rollup (increasing the level of aggregation) and drill-down (decreasing the level of aggregation or increasing detail) and slice and dice (selection and projection). Dimensional modeling creates individual models to address discrete business processes. Each model captures facts in a fact table and attributes of those facts in dimension tables linked to the fact table. The schemas produced by these arrangements are called star or snowflake schemas, and have been proven effective in data warehouse design.

The central table of the schema is called a fact table. A key characteristic of a fact table is that it contains numerical data (facts) that can be summarized to provide information about the history.

The data in fact tables is aggregated. Aggregation is the process of calculating summary data from detailed records. It is often tempting to reduce the size of fact tables by aggregating data into summary records [18] when the fact tables are created. However, when data is summarized in the fact table, detailed information is no longer directly available to the analyst. If detailed information is needed, the detail rows that were summarized will have to be identified and located in the source system that provided the data. Fact table data should be maintained at the finest granularity possible. Dimension tables contain attributes that describe fact records in the fact table. Dimension tables contain hierarchies of attributes that aid summarization. Dimensional modeling produces dimension tables in which each table contains fact attributes that are independent of those in other dimensions.

4. RESEARCH METHODOLOGY AND FRAMEWORK

This work is a blend of qualitative and quantitative methods of research. An experimental model is formulated which would utilize quantitative analysis techniques of time series data. Instruments developed for these experiments are different data warehouse models. The input data is gathered from the data files downloaded from historical data section of KSE website.

Part of this thesis can be categorized as quantitative research, since the experimental instrument designed is a data warehouse which would analyze the experimental input data using time series analysis techniques mentioned above.

The data warehouse model is fed in with input data and the out put data is compared and graphed with the input data. The question that whether or not a data warehouse can work as a tool for technical analysis is answered by analyzing the results of the experiments. Several data warehouse models are developed to analyze the data.

The experiment design considers Time as the independent variable where as the dependent variables are Closing and Opening Prices. The instrument designed uses these variables to derive certain values based on dependent variables.

The results obtained from the data warehouse when compared with input values show that data warehouse is able to perform time series analysis of stock market data. With such experimental results one may conclude that the data warehousing technology can be employed as a vital tool for technical analysis of stock market data.

More work needs to be done to further fine tune the experimental models, however, [19, 20] provide enough research evidence to safely conclude that the quantitative techniques used in the design of the data warehouse can be employed to use a data warehouse for technical analysis (time series analysis of stock market data).

The analysis model developed for this thesis is built on the foundation of the Microsoft Data Warehousing Framework. Based on SQL Server 2000 and Analysis Services, the framework offers a comprehensive infrastructure for data warehousing.

The Microsoft Data Warehousing Framework includes following components:

- A relational database server (SQL Server) to store the cleansed data
- An OLAP engine (Analysis Services) for multidimensional analysis
- Data Transformation Services (DTS) to extract, transform, and load data from source files
- Multidimensional Expression (MDX)
- Graphical interfaces
- A wide range of supported clients including MS Office 2003

The data warehousing framework for this research consists of a staging database, a subject matter database, an Analysis database, DTS packages, and Excel workbook. Time series data flows into the Excel workbook beginning with the Staging database. The data flows through to the Analysis database where the cleansed data is analyzed with moving averages and regression points. Excel Add-in for SQL Server Analysis connects to the analysis cubes and chart the trends and regression lines.

5. ETL PROCESS

Around 700 data files were downloaded from KSE website by configuring a downloader application to automatically download these files.

A dynamic URL http://www.kse.com.pk /histdata/YYYYMMDD.lis.Z was configured in the downloader application. To download the huge number of historical files automatically. The link was made configurable by selecting the sequential files option and configuring the link as:

http://www.kse.com.pk/histdata/YYYYMM%#%.lis.Z

Once the data files were downloaded into the framework, a DTS Package was designed to extract, transform and load the data into a staging database. Microsoft OLE DB Provider for SQL Server component of DTS package defines the connection properties of the Staging database. The Transform Data Task reads the data from downloaded files and transforms it to match the data types of the target columns in staging database.

6. SCHEMA AND MODELS

StockMart2004 data warehouse is created using SQL Server Analysis Service Manager.

Following cubes were designed for different types of analysis:

- KSE100 MA
- KSE100 ARLR
- Company MA
- Company ARLR

In KSE100 MA cube, calculated members are added using MDX to the cube that analyze the mClose measure.

MDX statements for the calculated members are following:

```
20DaysMA
Avg (LastPeriods (20, [Time].CurrentMember), [Measures].[M Close])

30DaysMA
Avg (LastPeriods (30, [Time].CurrentMember), [Measures].[M Close])

50DaysMA
Avg (LastPeriods (50, [Time].CurrentMember), [Measures].[M Close])

100DaysMA
Avg (LastPeriods (100, [Time].CurrentMember), [Measures].[M Close])

200DaysMA
Avg (LastPeriods (200, [Time].CurrentMember), [Measures].[M Close])
```

KSE100 ARLR Cube calculates the regression points using LinRegPoint function. Linear regression that uses the least-squares method calculates the equation of the best-fit line for a series of points. Let the regression line be given by the equation y = ax+b , where “a” is called the slope and “b” is called the intercept. LinRegPoint uses its last three arguments like the other LinRegxxx functions use them: to calculate the regression line. The function evaluates the first argument and uses the resulting number as the x value in the regression equation (y = ax + b) to calculate the y value.

In Company_MA cube, calculated members are added using MDX to the cube that analyze the Closing measure.

MDX statements for the calculated members are following:

```
20DaysMA
Avg (LastPeriods (20, [Time].CurrentMember), [Measures].[Closing])
```

30DaysMA
Avg (LastPeriods (30, [Time].CurrentMember), [Measures].[Closing])

50DaysMA
Avg (LastPeriods (50, [Time].CurrentMember), [Measures].[Closing])

100DaysMA
Avg (LastPeriods (100, [Time].CurrentMember), [Measures].[Closing])

200DaysMA
Avg (LastPeriods (200, [Time].CurrentMember), [Measures].[Closing])

Company_ARLR Cube calculates the regression points using LinRegPoint function and store these points in calculated members.

A lagged series of Closing is calculated as ([Measures].[Closing], Time.PrevMember) and used a input series for the Autoregressive model. The lagged series is named lagged_Closing. Two autoregressive models are formulated in Company_ARLR cube. pClosing_AR_prevAll which takes all the previous values of Closing in the time series as input to calculate the parameters of regression equation where as pClosing_AR_prev30 uses only previous thirty values of Closing for calculation of "a" and "b" for estimation of "y". Where "y" is the estimation of Closing as per the autoregressive model.

MDX statements that formulate the models are:

pClosing_AR_prevAll
LinRegPoint([Measures].[lagged_Closing],LastPeriods([Time].[Day].Members.Count),[Measures].[Closing],[Measures].[lagged_Closing])

pClosing_AR_prev30
LinRegPoint([Measures].[lagged_Closing],LastPeriods(30),[Measures].[Closing],[Measures].[lagged_Closing])

The second autoregressive model is based on the idea of giving more weight to the recent values [21]; however it is not treated by applying exponential weight. Closing values older than thirty days are ignored to examine stock price response to recent trading cycles. Both the models are then compared for better fit with the real data.

7. RESULTS

The results of the models are visually represented in the charts below:

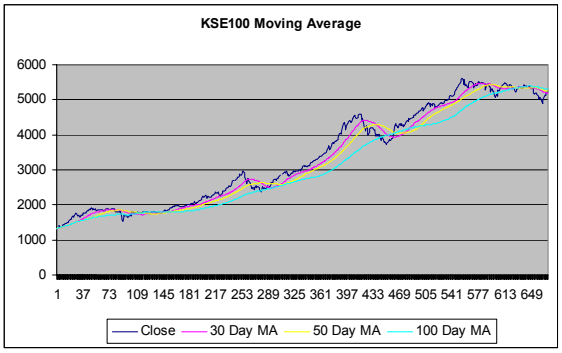

Figure 2: KSE100 MA

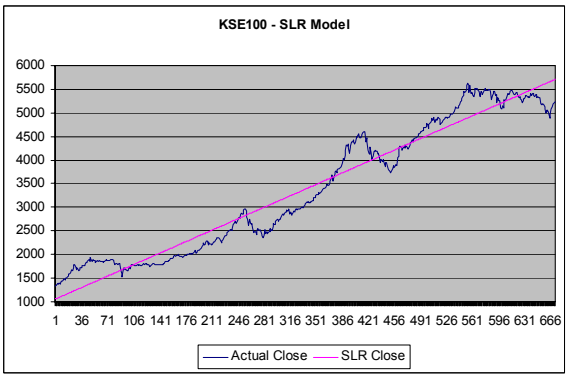

Figure 3: KSE100 SLR

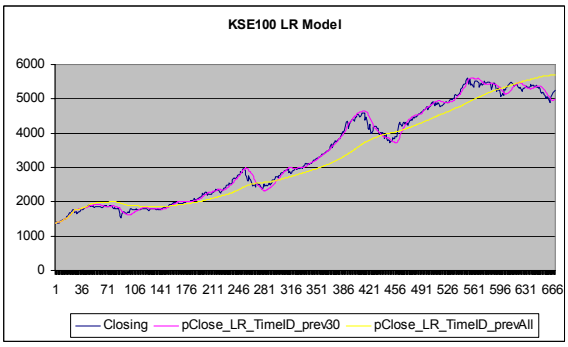

Figure 4:KSE100 LR

The results show that autoregressive techniques are most useful in modeling the market data time series. pClose_AR_prev30 proved to be more closer to the actual data than pClose_AR_prevAll.

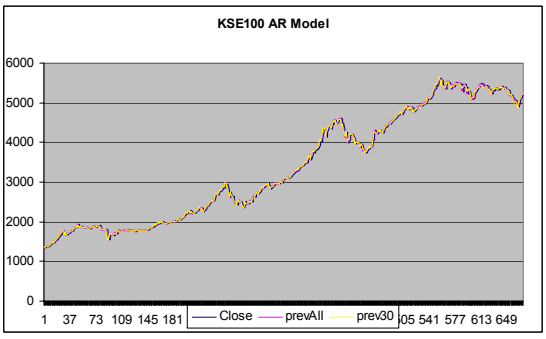

Figure 5: KSE100 AR

Figure 6: Company MA

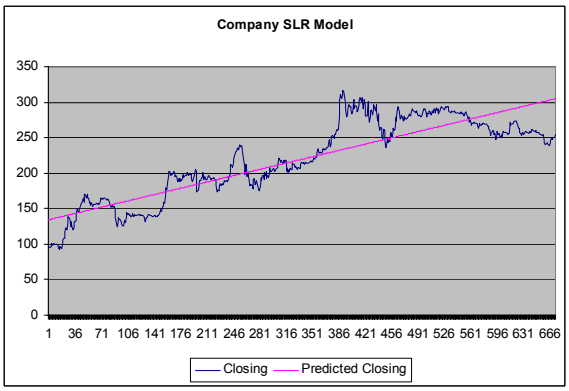

Figure 7: Company SLR

Figure 8: Company LR

Figure 9: Company AR

8. CONCLUSION

After analyzing the results and charts above, it appears that autoregressive models are most useful for analysis of the stock market time series data. An OLAP schema designed to model the time series for autoregressive analysis, gives best fitted series when compared to the actual data. Hence if further work needs to be done to predict and forecast the index and price movements in the future, regression modeling should be considered for data mining as well. The framework used is based on standard industry tools like SQL Server and MS Excel. Excel being the most widely used spread sheet analysis tool in the financial industry would probably provide the technical analysts ease of use, built-in functionality for drawing charts and productivity.

9. FUTURE WORK

As seen above that OLAP and data warehousing technology can be used for technical analysis of stock prices and market index behavior. However, the framework proposed and implemented in this thesis can only help the technical analysts in forecasting the stock price. It might be interesting to find out whether a data warehouse can be designed to work as a technical analyst itself. It would require investigations in the area of OLAP time series data mining and predictive mining. It would also require the researcher to device custom OLAP data mining algorithms (plug-in algorithms) for technical analysis and prediction. Mining patterns would reveal further knowledge about the relationships between the parameters of stock price time series data.

Ralph Nelson Elliott developed the Elliott Wave principle in the late 1920s by discovering that stock markets, thought to behave in a somewhat chaotic manner, in fact, did not. They did, however, trade in what he called repetitive cycles, which he discovered were the emotions of investors as a cause of outside influences, or predominant psychology of the masses at the time. He had stated that the upward and downward swings of the mass psychology always showed up in the same repetitive patterns, which were then divided into patterns he termed Waves. It was understood by the technicians at the time that because of the fractal nature of the markets, Elliot was able to breakdown and analyze the markets in much greater detail.

This allowed him to spot unique characteristics of wave patterns and making detailed market predictions based on the patterns he had identified. Fractals are mathematical structures, which on an ever-smaller scale infinitely repeat themselves. The patterns that Elliott discovered are built in the same way. An impulsive wave, which goes with the main trend, always shows five waves in its pattern. On a smaller scale, within each of the impulsive waves of the before mentioned impulse, again five waves will be found. In this smaller pattern, the same pattern repeats itself ad

infinitum (these ever smaller patterns are labeled as different wave degrees in the Elliott Wave Principle). Fibonacci numbers provide the mathematical foundation for the Elliott Wave Theory of Technical Analysis.

The mining models would need to learn Elliott Wave Patterns and identify such patterns in the target data.

REFERENCES

[1] Widom, J. "Research Problems in Data Warehousing." Proc. 4th Intl. CIKM Conf., 1995

[2] Harinarayan V., Rajaraman A., Ullman J.D. "Implementing Data Cubes Efficiently" Proc. of SIGMOD Conf., 1996

[3] S. Chaudhuri and U. Dayal. "An overview of data warehousing and OLAP technology." SIGMOD Record, 26:65--74, 1997

[4] Agrawal S. et.al. "On the Computation of Multidimensional Aggregates" Proc. of VLDB Conf., 1996

[5] Y. Chen, G. Dong, J. Han, B. W. Wah, and J. Wang. "Multidimensional regression analysis of time-series data streams." In VLDB Conference, 2002

[6] Y. Chen, "Supporting Multi-Dimensional Stream Data and Time Series Analysis in OLAP and Data Cubes", (2002)

[7] Kimball R., Strehlo., "Why decision support fails and how to fix it", reprinted in SIGMOD Record, 24(3), 1995.

[8] Codd, E.F., S.B. Codd, C.T. Salley, "Providing OLAP (On-Line Analytical Processing) to User Analyst: An IT Mandate."

[9] Dewitt D.J., Gray J. "Parallel Database Systems: The Future of High Performance Database Systems" CACM, June 1992.

[10] James Liu and Tommy Leung, "A Web-Based CBR Agent for Financial Forecasting", International Conference on Case-Based Reasoning, 2001 (ICCBR'01)

[11] Moving Averages, Incredible Charts; http://www.incrediblecharts.com/technical/moving_average.htm

[12] Moving Averages, Stock Charts; http://www.stockcharts.com/education/IndicatorAnalysis/indic_movingAvg.html

[13] Javier Contreras, Rosario Espínola, Francisco J. Nogales, and Antonio J. Conejo, "ARIMA Models to Predict Next-Day Electricity Prices", IEEE Transactions on Power Systems, Vol. 18, No. 3, August 2003

[14] E. Weiss, "Forecasting commodity prices using ARIMA," Technical Analysis of Stocks & Commodities, vol. 18, no. 1, pp. 18–19, 2000

[15] The Method of Least Squares, eFunda – Engineering Fundamentals, Mathematica in Engineering, WOLFRAM Research; http://www.efunda.com/math/leastsquares/leastsquares.cfm

[16] Stigler, S.M., "Mathematical Statistics in the Early States," The Annals of Statistics, Vol. 6, pp. 239-265.

[17] Multiple Regression, eFunda – Engineering Fundamentals, Mathematica in Engineering, WOLFRAM Research; http://www.efunda.com/math/leastsquares/lstsqrzrwtxy1d.cfm

[18] Blakeley, J.A., N. Coburn, P. Larson. "Updating Derived Relations: Detecting Irrelevant and Autonomously Computable Updates." ACM TODS, Vol.4, No. 3, 1989

[19] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast subsequence matching in time-series databases." SIGMOD'94

[20] R. Agrawal, K.-I. Lin, H.S. Sawhney, and K. Shim, "Fast similarity search in the presence of noise, scaling, and translation in time-series databases." VLDB'95.

[21] S. Babu and J. Widom. "Continuous queries over data streams." SIGMOD Record, 30:109-120, 2001