



Time Series Usage in Security Applications

Khalid Alfalqi

CIISE, Concordia University, Canada

Faculty of Computer Science, Umm Al-Qura University, Saudi Arabia

Abstract—Time series modeling has a significant impact on different fields of study. In fact, researchers have developed different models that can improve our understanding of many fields that rely on time series. As a result, a number of models have developed to improve the efficiency of time series modeling and forecasting. In general, this paper sheds light on some time series definitions, as well as objectives. In addition, the paper describes time series components and models. Moreover, this paper examines time series in real world applications such as security.

Keywords— Time series , Autoregressive Process , Autoregressive moving average (ARMA) , Fraud Detection and Malware Detection .

I. INTRODUCTION

1.1 Background

There are numerous definitions of time series, based on a study's purpose, as well as the nature of specialization. Time series, broadly speaking, is a set of observations for a particular phenomenon that had been recorded at different times, usually occurring at uniformed intervals, such as: hours, days, months, or years.

A second definition of time series is a series of numbers and values recorded according to a particular time (seasons, or months, or days, or any unit of time), so it is a sequential historical record, which is prepared in order to build future forecasting. [1]

There are multiple purposes for time series analyses; for instance, time series is used in statistics, signal processing, finance, weather forecasting, earthquake prediction, and astronomy"[2]. With these, time series data can be analyzed to gain a better understanding of the implicit structure and functions of observations. In fact, understanding the techniques of a time series allows one to develop a model with which explain the data in the forms of prediction, monitoring, or control.

This ability has real-world implications. One of the most important duties of governments, institutions, and corporations, is planning for the future so as to achieve the objectives, provide stability to constituents, and predict potential events before they occur. In this, time series analyses represent one of the most important methods of prediction, and are especially useful with respect to generating economic indicators, annual company sales, and population trends.

There are multiple objectives for using time series in various domains; some time series objectives include:

- Description: define the design in correlated data.
- Demonstration: understanding the data.
- Forecasting: predicting trends from past samples.
- Intervention analysis: how can a single event alter the whole time series?
- Quality monitoring: Defining a problem by monitoring the variation of a size.[3]

1.2 Time Series Components

A time series can be divided into four parts, called components, which can be determined by analyzing and decomposing the time series. The four main components describe the variations of data over time.

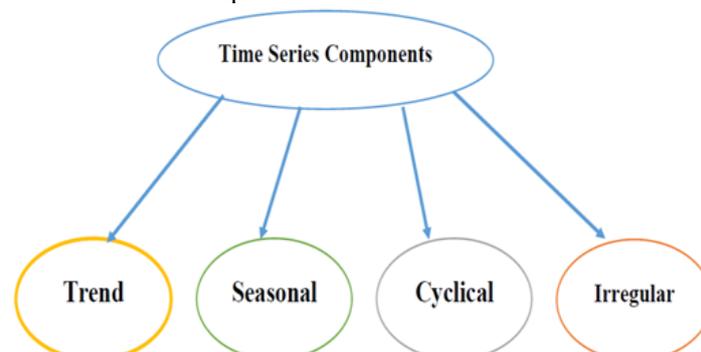


Fig1: Four time series components.

Any time series can have some or all of the following components:

- A. Long-term of secular trend.
- B. Seasonal variations.
- C. Cyclical movement.
- D. Irregular variation.\

A. Long-Term of Secular Trend (T).

This element describes the regular movement of a series across a relatively long period of time, and is usually the most important component of a time series. Also, this component describes the evolution of the series over the long term. Moreover, it may be a positive trend, if the value of the series is increasing over time, or it may be negative trend if the value of the series decreases over time. These trends can be seen in population growth, price inflation, and general economic changes. [4]

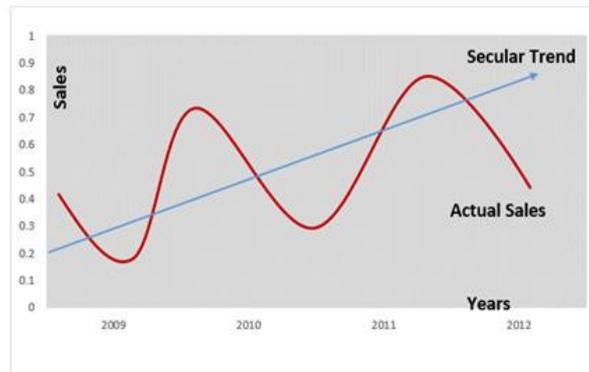


Fig2:Sales trendsover four years

B. Seasonal Variations (S).

This component describes the short-term movements over a fixed period of time due to seasonal factors such as weather, vacation, customs, and holidays. This movement can be seasonal or quarterly, and result from changes in patterns as a result of external factors, which are often systematic way. This component is often categorized by periods no longer than a year (daily, weekly, monthly, or quarterly).[4]

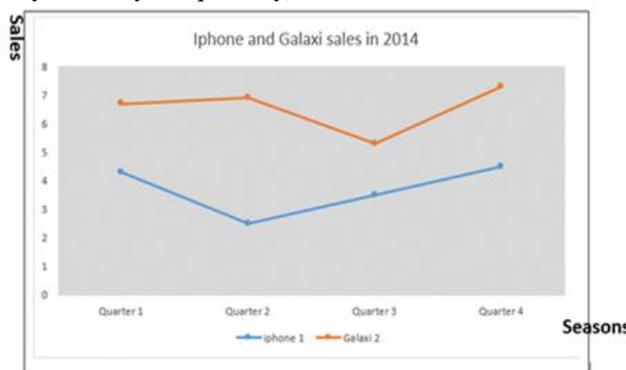


Fig3:Sales of cellular devices by quarters, in 2014.

C. Cyclical Movement (C).

This third component tracks changes in time series values that are not of a fixed period, with duration being more than the seasonal variation period; this usually represents at least two years.

This component consists of four stages in the full cycle: prosperity, recession, depression, and recovery. One popular example of cyclical movement is the business cycle.[4][1]

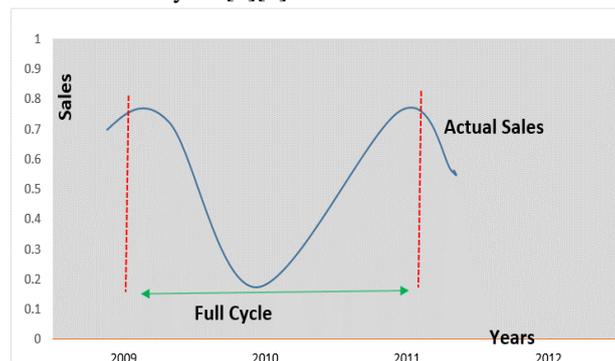


Fig 4: Full cycle of a company over several years

D. Irregular variation (I).

This component graphs sudden changes in the time series caused by random factors, such as earthquakes, volcanoes, epidemics, wars, and labor strikes, whose impact cannot be predicted due to the random movements. These factors are unlikely to be repeated. For example, the sales volume of a store may increase suddenly as a result of an unplanned local music festival. [4][1]

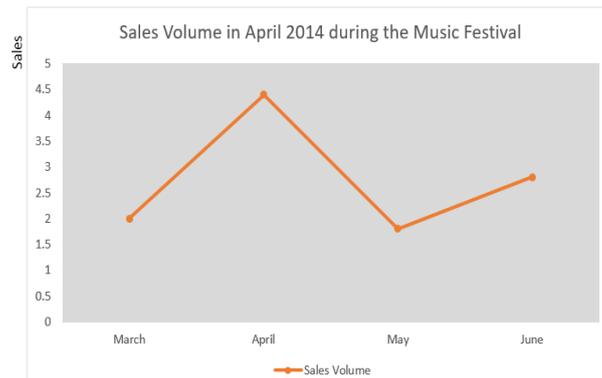


Fig5:Increasing sales volume of a store in April, where an unplanned music festival occurred.

Time series decomposition models mainly possess an additive relationship or multiplicative relationship.

➤ Additive Relationship:

This model assumes that the four time series components are expressed as a sum of the four components. Also, this implies that the value of any component does not affect, and cannot be affected, by the value of other components.

$$y_t = T + C + S + I$$

➤ Multiplicative Relationship:

This model assumes that time series is presented as a product of the four components. Also, this implies that the value of the four components are dependent on each other.[5]

$$y_t = T \times C \times S \times I$$

II. TIME SERIES MODELS

2.1 Autoregressive Process (AR):

The autoregressive process is a random process used in statistical calculations to speculate the future values by weighting past values. An autoregressive process operates under the assumption that past values affect existing, and thus future, values. These autoregressive processes are particularly useful in technical analysis. AR uses past values to predict future events; however, this sort of analysis has a disadvantage, namely that the past value is not guaranteed to forecast to future trends. [6]

The autoregressive process is defined by the equation:

$$X_t = \sum_{r=1}^p \phi_r X_{t-r} + \varepsilon_t$$

AR fitting steps in modeling

The first step in fitting the collected data into the AR model is to speculate the value of p . This can be achieved using the least squares assessment. The second step is to define the approximate distribution for the data. To find the distribution of the data, we need to first find the distribution of the error terms.[7]

2.2 Autoregressive Moving Average (ARMA):

Autoregressive moving average (ARMA) is a mathematical model that relies on autocorrelation in a time series. ARMA is a combination of autoregressive analysis AR and moving average MA methods.

ARMA assumes that the time series is stationary. These ARMA models are very popular in several fields, such as economics and hydrology.

Modeling ARMA to data helps us to understand some physical systems. Also, it provides researchers with the ability to predict the behavior of some phenomenon by recording past patterns. In addition, ARMA has practical benefit in developing and testing simulations. ARMA is defined by:

$$X_t - \sum_{r=1}^p \phi_r X_{t-r} = \sum_{s=0}^q \theta_s \varepsilon_{t-s}$$

ARMA fitting steps in modeling

ARMA modeling has a very clear steps. The first step is to define the model, and appoint the appropriate structure (AR, MA or ARMA). The second step is to assess the coefficients of the model. Ultimately, this step is unknown to the user, and is carried out automatically by a computer. The third step is to inspect the model, which is called diagnostic inspection. In this, there are two very important elements when checking the model. The first element is to assure that the superfluity of the model is random, and the second is to ensure that the speculated parameters are statistically significant.[8]

III. TIME SERIES APPLICATION IN SECURITY

3.1 Credit Card Fraud Detection using Time Series

Nowadays, credit card fraud detection has become a significant challenge to financial corporations. Credit card fraud is a type of identity theft in which an unauthorized party steals the credit card information of another party, and uses it for a variety of purposes, such as purchasing goods or withdrawing funds. This fraud can occur online through the Internet, or offline in a variety of ways. The perpetrator can obtain credit card information by using lost or stolen credit cards, stealing cards from mailboxes, by “shoulder surfing” during transactions, by using social engineering tools, or by looking at personal records. Fraud detection is a quick response used to identify potential cases, and warn customers if fraud occurs; this also allows the security corporations to act quickly and prohibit the intruder from unlawfully using the credit cards.

There are several new technologies designed to prevent such fraud, one of which is data mining. Data mining is very efficient and play an important role in fraud detection, as it takes the fraud data as an input, and “results methods” or patterns as an output. Specifically, this paper proposes two methods of security. The first method is a module that is implemented by an outlier detection approach using distance-based methods, as in data mining. The outlier is detected in the transaction that occurs after the fraud took place. The second method is a module using time series analysis to predict the next incoming transactions. The fraud then can be confirmed by asking a series of security questions to the card holder. By observing the perpetrator’s spending behavior, we can detect potential activity by using the two proposed modules: [9]

Module 1 - Detect the anomaly of an incoming transaction.

Module 2 - Predict the next transaction, thus detecting an anomaly

Module 1 – Distance Based Method

After observing and analyzing a user’s past spending behavior, and obtaining the threshold value, we can then detect whether an incoming transaction is a fraud by comparing the amount of the new transaction with the user’s analyzed spending behavior. If the transaction amount exceeds the threshold value, then the transaction is deemed suspicious, and is more likely to be a fraudulent transaction. In these cases, we ask the user to verify the transaction by allowing him/her to answer the established security questions. [9]

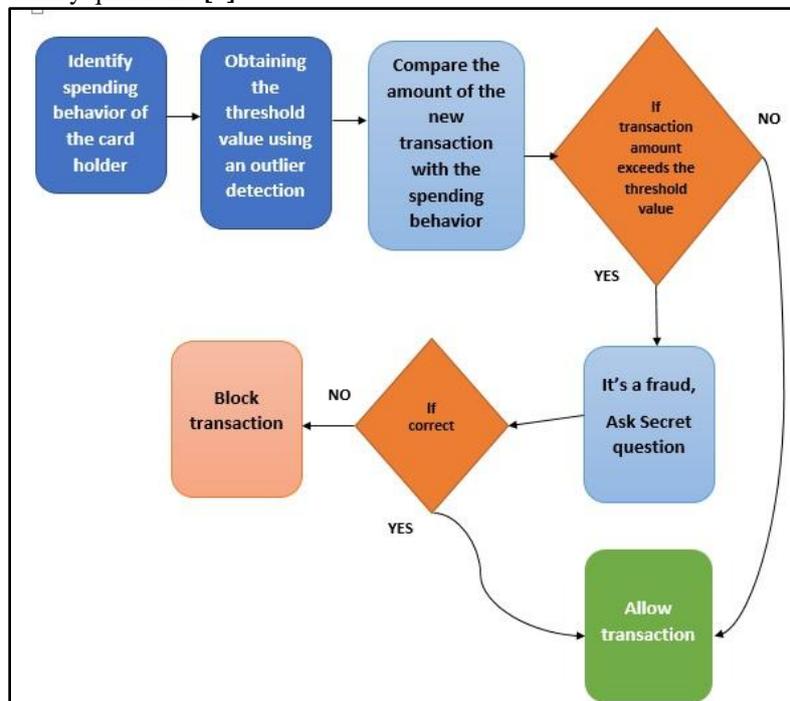


Fig 6: Module 1 – Data flow diagram

Algorithm:

1. Identify past spending behavior and the threshold value.
2. Find the (x,y), where x is the mean of the transaction number, and y is the mean of the transaction amount.
3. Find the distance between each point using Euclidean distance formula

$$\text{Euclidean distance} = \sqrt{(xi - x)^2 + (yi - y)^2}$$

Where $i=1, 2, 3, \dots$
4. Fix the maximum distance as the threshold.
5. If an incoming transaction occurs, repeat steps 2 and 3.
6. If the distance of the new transaction is less than the threshold, then the transaction is not fraudulent.
7. Once a suspicious transaction is detected, ask the user to verify the transaction by allowing him/her to answer the security questions.
8. Update the threshold value after each transaction. [10]

Module 2 – Label Prediction Method

In this module, a data pattern is observed and recorded at a regular periods using time series methods. Using a label prediction methodology to assign a label for each transaction. These labels can be low, medium, or high. In addition, this method calculates the amount of each transaction, and the amounts are then clustered into low, medium, or high categories. This module allows for two levels of outlier detection. First, each transaction amount is clustered using a k-means clustering algorithm. Second, an outlier detection can occur by comparing the distance of the new incoming transaction with the predicted range of transaction; if there is any deflection, then the transaction is suspicious, and potentially represents a fraudulent transaction. If this takes place, we can ask the user to verify the transaction by allowing him/her to answer the previously established security questions.

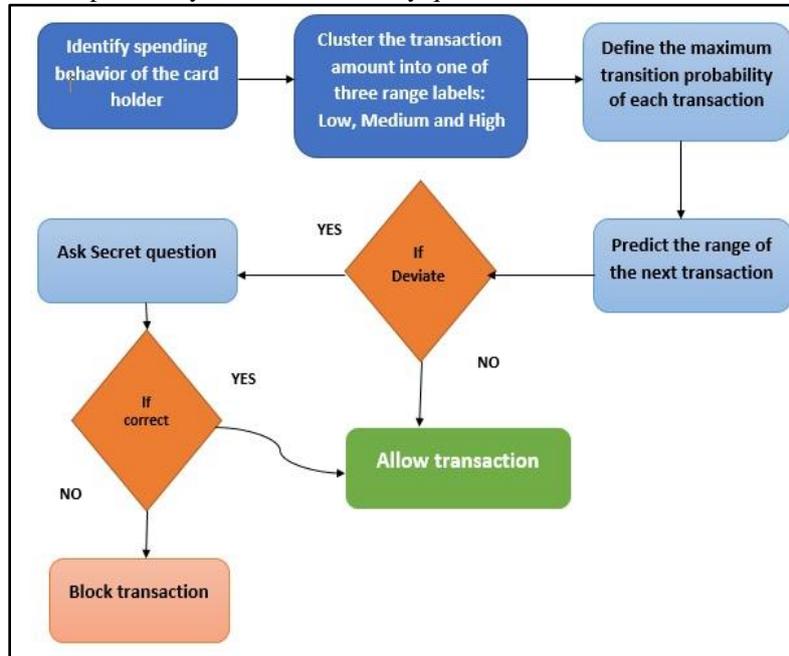


Fig 7: Module 2 – Data flow diagram

Algorithm:

1. Identify the spending behavior.
2. Categorize each transaction into one of three label clusters: Low (L), Medium (M) and High (H).
3. Obtain the range of each transaction for every transaction amount.
4. Define the maximum transition probability of each transaction.
5. Compare the range of the incoming transaction with the maximum transition probability.
6. If the range of the new transaction is beyond the scope of the predicted range then the transaction is suspicious.
7. Once a suspicious transaction is detected, ask the user to verify the transaction by allowing him/her to answer the security questions.
8. Update the transition probability after each transaction.[10]

3.2 Android Malware Detection using Multivariate Time-Series Technique

Nowadays, smartphones has become popular around the world. As a 2014 Gallup Report indicated, by July 2014, ~80% of Korean adults reported using smartphone. Essentially, four out of five Korean adults are using smartphones each day. Also, with the gradual growth of smartphones usage, the services and applications have increased at a similar pace. The user friendly nature smartphones and smartphone services has resulted in a decrease in the number of people who use PCs. As a result, smartphones are becoming a prime target for hackers as a platform on which to launch malicious code attacks. Given this reality, it is important to consider methods with which to protect smartphones. [11]

Among existing operating systems (OS) currently used on smartphones, the android OS is the most targeted by malware users and attackers. The reason for this is that Android uses an open source platform, so users can develop their own applications. In fact, there are several studies that consider methods with which to prevent malicious codes from targeting smartphones. These studies are mainly divided into three categories: signature-based analyses, behavior-based analyses and dynamic-based analyses.

This paper focuses on the dynamic-based analysis, which is a technique used to detect malicious codes by analyzing the sensitive data. Also, by using a multivariate time-series technology (ARMA), malicious codes can be detected. [11]

3.2.1 Malware Detection Framework

The structure of the malware detection system consists of four stages. These stages include:

- A. Resource information extraction.
- B. Data modeling.
- C. Data sampling with Z-score.
- D. Malicious code detection.

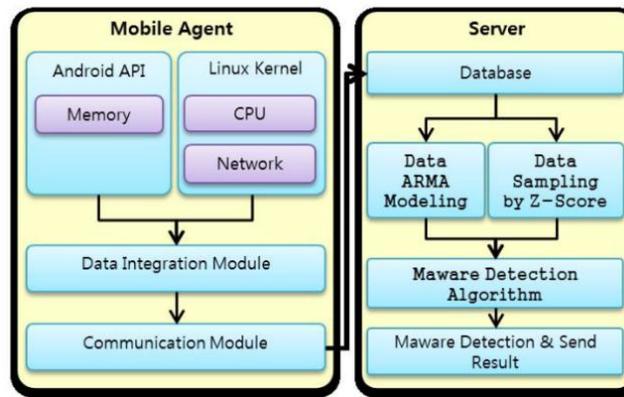


Fig 8: Structure of Malware detection system

A. The Resource Information Extraction

There are five methods used in this stage: Rxbyte;Txbyte for information of networks; information of the memory (usage memory); the device’s total CPU use rate (TotalCPU); and the CPU use rate based on the user’s account (User CPU).

AsFigure 8 depicted, information about Android’s hardware and operating system can be extracted from the Android Linux kernel. Moreover, information regarding the software can be extracted from the Android’s API. Then, the data integration module integrates the extracted information into data. Finally, by using the data communication module, the data is transferred to the server.

B. Data Modeling with Multivariate Time-series Techniques :

Once the data is transferred to the server, the server analyzes the data using the autoregressive moving average model (ARMA), and calculates the data sample’s mean. In addition, this method assigns a weight to the data by calculating the correlation of the respective resource information, and assigning the higher weight. [11]

To use the ARMA model, one must find the values of the parameters p, d, and q first. These values can be obtained by calculating the Akaike Information Criterion (AIC) value, as shown in Fig 8.

```

Procedure Parameter_estimate()
Input Data : The time-series data of normal status
Output R : The set of estimated parameter
(AR, MA, Integrated)
(1) min_AIC=100;
(2) for p:=0 to 8 do
(3) for q:=0 to 2 do
(4) for d:=0 to 2 do
(5) TEMP=sarima(Data, p, d, q);
(6) if(TEMP$AIC < min_AIC) do
(7) min_AIC=TEMP$AIC;
(8) AR=p; MA=q; Integrated=d;
(9) end-if
(10) end-for
(11) end-for
(12) end-for
(13) R=(AR, Integrated, MA)
    
```

Fig9: Algorithm for calculating optimal parameter of ARMA

C. Data Sampling with Z-score :

The main reason for using the data sampling with Z-score is to reduce the time required for analysis. This expediency results from the fact that Z-score is able to detect malicious codes quickly with no need to check the entire dataset or if the extracted data deviates from the original data.

D. Malicious Code Detection :

This process detects malicious codes by comparing the data modelled in the ARMA model with the data sampled through Z-score. This process uses Normal_Data, Zscore_Data, p, d, and q as inputs in order to calculate the threshold

values as an output. After defining the threshold values, the malicious code is detected by comparing the threshold value with the actual value, based on the assumption that there is a significant difference between the actual value and the predicted value if a malicious code is executed.[11].

IV. CONCLUSION

Time series analysis plays an important role in our lives. It accelerates the wheel of development in different fields. There are several models in time series analyses, which allow researchers to develop applications in security field. This paper discusses the benefits of using time series models to detect malicious code in Android smartphones. By applying the multivariate time-series technique and ARMA models, the integrated data is very helpful, and allows one to detect the malicious code. In addition, time series models work extremely effectively in detecting credit card fraud. Ultimately, time series models allow for the identification of transactional fraud by analyzing the cardholder's spending behavior.

REFERENCES

- [1] Peter J. Brockwell. Richard A. Davis, "Introduction to Time Series and Forecasting, Second Edition, New York, Springer" 2002, pp 1-7 .
- [2] Shumway, Robert H., Stoffer, David S "Characteristics of Time Series" ,Time Series Analysis and Its Applications 3rd edition, New York, Springer ,2011, pp 1-11 .
- [3] Chatfield, C., "The analysis of time series – an introduction. Chapman and Hall, London, UK. Sixth Edition," 2004, pp 3-7.
- [4] Ratnadip Adhikari , R. K. Agrawal, "An Introductory Study on Time Series Modeling and Forecasting," LAMBERT Academic Publishing ,2013
- [5] University of Cambridge, statistical laboratory <http://www.statslab.cam.ac.uk/~rrw1/timeseries/t.pdf> 2014.
- [6] Xiaoguo Wang, Yuejing Liu, "ARIMA Time Series Application to Employment Forecasting," Proceedings of 2009 4th International Conference on Computer Science & Education, IEEE 2009 .
- [7] Zsuzsanna Horvath, Ryan Johnston, "AR(1) TIME SERIES PROCESS "University of Utah: 2006, <http://www.math.utah.edu/~zhorvath/ar1.pdf> .
- [8] John H. Cochrane "Time Series for Macroeconomics and Finance," spring 1997; Pictures added Jan 2005, pp 31-38 .
- [9] R.Devaki, V.Kathiresan ,S.Gunasekaran, "Credit Card Fraud Detection using Time Series Analysis, "International Conference on Simulations in Computing Nexus, ICSCN-2014, International Journal of Computer Applications® (IJCA) 2014.
- [10] Abhinav Srivastava, Amlan Kundu, Shamik Sural and Arun K. Majumdar, "Credit Card Fraud Detection Using Hidden Markov Model," IEEE Transactions on Dependable and Secure Computing, vol. 5 No. 1, 2008.
- [11] Ki-Hyeon Kim, and Mi-Jung Choi, "Android malware detection using multivariate time-series technique," APNOMS, IEEE, 2015