A NOVEL APPROACH TO REDUCE THE NOISE OF TIME SERIES DATA USING PREDICTIVE WEIGHTED MOVING AVERAGE

T.Rajesh[1], Dr. K.V.G Rao[2]

Abstract— With the growing demand of the long term and short term planning based on predictions of time series data, this is the demand of the modern research to provide significant and promising methods for time series data analysis. Nevertheless the time series are prone to fluctuations and thus makes it difficult to analysis and perform any kind of predictive operations for making decisions. Time series clustering is an important solution to various problems in numerous fields of research, including business, medical science, and finance. However, conventional clustering algorithms are not practical for time series data because they are essentially designed for static data. This impracticality results in poor clustering accuracy in several systems. Thus smoothing of the time series is the most significant part of the research. Hence this work analyses the performance of various types of smoothing methods like moving average, weighted moving average and identifies the pitfalls of these methods. The final outcome of this work is proposing and evaluating the performance of a predictive weighted moving average method for smoothing the data. The novelty of the work includes reduction of the difference between the actual data points from the time series and calculated average time series. This contribution will help this work to contribute in an algorithm for calculating the piecewise aggregate approximation of a time series and also proposing a clustering algorithm in future.

Keywords—— Time Series, Pitfalls, Secular Trend, Cyclic variation, Seasonal variations, Moving Average, Weighted moving average, MLP

I. INTRODUCTION

A time series is eventually a collection of periodic recording of data. The data series based on time can be a collection from various sources with symmetric or dissymmetric nature depending on the nature of the data [1]. For example in case of random data sources, the weightage average of the data cannot produce any significant point to make any prediction, thus making it more difficult to use it for any statistical analysis [2] [16]. In order to get rid of the problem, the use of moving average and considering the moving average as new time series data for predictive analysis is the most popular way [3]. Nevertheless, the moving average or MA method is prone to error for multiple reasons. In this work, the problems of the MA methods are been highlighted and suitable solutions are also been recommended.

With the focus of time series data [4], this work defines the basic nomenclature of the time series. The use of time series for making long term decision and also the use in making just in time decisions makes it highly popular. The long term analysis of the time series can range from 5 years to 20 years for any given organization giving the opportunity for the other department of the organization to make plans for financials and human capital related decisions [5]. In the other hand, Just in time decisions are ranging from 1 week to 1 month time to enable less mission critical decisions for the organization [6]. The time series can show high rate of unpunctuality for a longer duration. Hence it is to be understood that, the errors or the noises will affect the long term analysis rather than the Just in time analysis.

[1] *Research Scholar,CSE Department, JNTUH, Hyderabad,India*
[2] *Professor, CSE Department G.Narayanamma Institute of Technology &Science, Hyderabad,India*

More and more organizations are tend to plan for long term strategies and depend on the time series, thus making the time series noise free is the major demand of any research.

Conversely, the facet understanding of the time series is must for any type of operations on time series and a very less number of literatures address that issue. Consequently, in order to proceed further, in this domain of research, this work produces a major understanding of the time series and its basic properties.

II. TIME SERIES COMPONENTS

Depending on the nature of the variation in the data, the time series are categorised in four different categories. In this section, the components of the time series are been examined.

A. Secular Trend

Secular Trends are the time series which demonstrates steady upwards or steady downwards or steady over the time [7] [8] [9].

A secular trend can be represented by the following formula:

$$T = +\bigcup_{i=1}^{n} X_i \qquad \text{(Eq. 1)}$$

Or

$$T = -\bigcup_{i=1}^{n} X_i \qquad \text{(Eq. 2)}$$

Or

$$T = \bigcup_{i=1}^{n} X_i \qquad \text{(Eq. 3)}$$

Here the Eq. 1 represents a steady upwards trend, Eq. 2 represents a steady downwards trend and Eq. 3 represents the steady trend.

The secular trend can be understood with the example [Table – 1].

TABLE I: ANNUAL PRICES OF AN APPLIANCE WITH INDEX NUMBERS FOR PRODUCT X, Y AND Z

Year	Product X	Product Y	Product Z
2000	$100	$101.7	$101
2001	$100.2	$101.5	$101
2002	$100.3	$101.4	$101
2003	$100.5	$101.2	$101.1
2004	$100.7	$100.9	$101.1
2005	$100.9	$100.7	$101
2006	$101.2	$100.5	$101
2007	$101.4	$100.3	$101.1
2008	$101.5	$100.2	$101.1
2009	$101.7	$100	$101

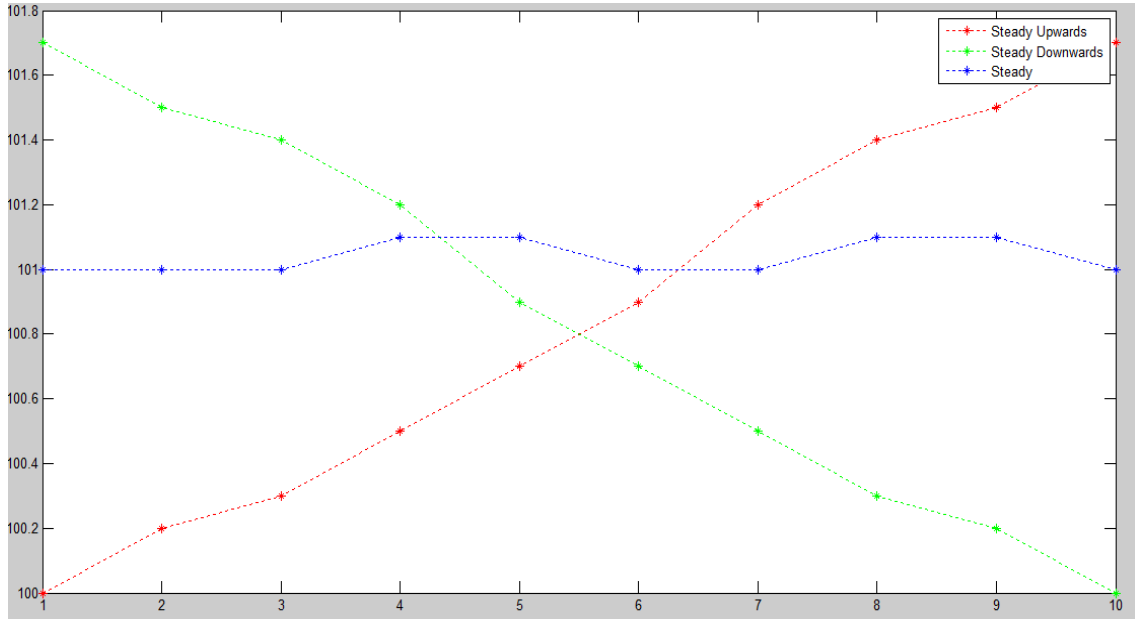The secular trend also can be understood graphically.

Fig. 1 Product Variance in Secular Trend

B. Cyclical Variation

The Cyclical Variations are the time series which demonstrates the prosperity, depression, recovery, recession and secular trends [10] [11].

A cyclical variation can be represented by the following formula:

$$T = +\bigcup_{i=1}^{n-1} X_i \oplus -\bigcup_{i=n}^{n+1} X_i \oplus \bigcup_{i=n+2}^{n+n} X_i \qquad \text{(Eq. 4)}$$

The cyclical variation can be understood with the example [Table – 2].

TABLE II: ANNUAL PRICES OF AN APPLIANCE WITH INDEX NUMBERS FOR ORGANIZATION ABC CORP.

Year	Annual Turn Over for ABC Corp.
2000	$1584
2001	$1272
2002	$1951
2003	$1391
2004	$1702
2005	$1391
2006	$1539
2007	$1647
2008	$1182
2009	$1539

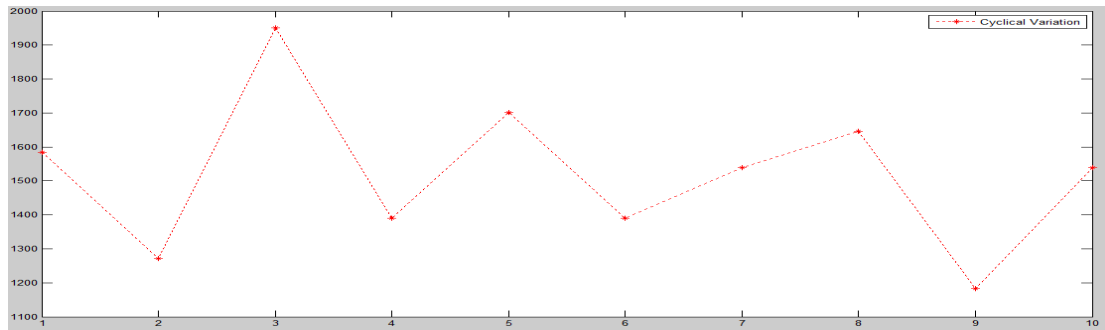The cyclical variation also can be understood graphically.

Fig. 2 Organization Annual Turn Over in Cyclical Variation

The majority of the time series are identified to be cyclical variation, where the average is to be calculated. Due to the noisy nature of the data, it is difficult to calculate the moving average and the weightage average known not be to highly dependable for making predictions. In the next section of this work, the problems with moving average is been highlighted.

C. Seasonal Variation

The Seasonal Variations are the time series which demonstrates the prosperity, depression, recovery, recession and secular trends in a discrete time series [13] [14]. The discreteness of the time series makes it more difficult to use it for long term predictions. However, many time series are captured from the seasonal business data or seasonal healthcare data tend to show the nature of a seasonal variation.

A seasonal variation can be represented by the following formula

$$T = +\bigcup_{i=1}^{n-1} X_i \oplus -\bigcup_{i=n+k}^{n+j} X_i \oplus \bigcup_{i=n+d}^{n+n} X_i \qquad \text{(Eq. 5)}$$

Here it is clear to understand the discontinuity of the time series.

The seasonal variation can be understood with the example [Table – 3].

TABLE III: ANNUAL PRICES OF AN APPLIANCE WITH INDEX NUMBERS FOR FOUR QUARTER

Year	Quarter – 1	Quarter – 2	Quarter – 3	Quarter – 4
2000	$100	$101.7	$101	$101
2001	$100.2	$101.5	$101	$101
2002	$100.3	$101.4	$101.1	$101
2003	$100.5	$101.2	$101.1	$101.1
2004	$100.7	$100.9	$101	$101.1

III. PITFALLS OF MOVING AVERAGE

The major reasons for popularity of the moving average method are not only for smoothing the series rather to calculate the long term fluctuation of the data. In order to apply the moving average method, the time series must have a rhythmic pattern and should not contain a very high fluctuation over a long time. The result of moving average can be best utilized and the fluctuations can mostly be removed if the time series is having regular intervals.

Nevertheless, the moving average of the time series is not prone to errors. This work highlights [15] some of the errors in moving average method:

- The moving average method takes the influences from the past trends, however the method fails to accommodate the influences from other factors like change in demand in case

of production industry or change in temperature in case of bacterial growth or change in medical improvements in case of health care time series.

- The method of calculating the moving average fails to understand the influence of external and random events, due to which a certain point of data varied. This might influence the prediction.

- The time series can consider a long term or a short term series over time. The short term time series can demonstrate a high fluctuation rate and the long term time series demonstrates low fluctuation rates. Thus choosing the period is the key to correct prediction in case of moving average method.

- The moving average method includes few sub methods like simple moving average, exponential moving average. Thus choosing the most suitable method is very crucial for the chosen data.

- The trend analysis for any time series is said to be having no previous records. Thus the technically of the moving average method is arguable.

- The moving average method is fairly good for smoothing the time series; nevertheless the use of this method for prediction is uncertain.

- The moving average method always demands a supporting method for correction to make it useable for prediction without giving much scope for argue.

Henceforth it is to be realized that, the moving average method can be used effectively to smooth the time series and the method is to be accompanied by some special influential method in order to increase the effectiveness.

Thus in the next section of this work, focuses on predictive weightage average method and demonstrates the improvement over the weighted average and moving average methods.

IV. PROPOSED PREDICTIVE WEIGHTED MOVING AVERAGE METHOD

Considering the need for smoothing the time series for better predictive analysis, the recent demand of improved moving average method cannot be ignored. Thus in this part of the work, the improved predictive moving average method is been proposed and evaluated.

Firstly, we analyse the moving average method and weighted moving average method using the following lemmas and theorems.

Lemma 1: The moving average of any given time series reduces the fluctuation of the time series thus creates a new time series.
Where,

$$\{n_i, n_{i+1}, n_{i+2}, \ldots\ldots\ldots n_{i+k}, n_{i+k+1}, \ldots\ldots n_n\} \quad \text{(Eq. 6)}$$

Denotes the time series, where the indexes denote the items in the series in regular intervals.

$$\{n_j, n_{j+1}, \ldots\ldots\ldots n_{j+k}, n_{j+k+1}, \ldots\ldots n_{n-j}\} \quad \text{(Eq. 7)}$$

Denotes the moving average series, where the indexes denote the interval of the weighted average for the series.

Proof: The intervals of the moving average series are higher than the intervals of the actual time series. Thus the relation between the two intervals can be formulated as:

$$i << j \qquad \text{(Eq. 8)}$$

Henceforth the relation between two successive elements in the time series and weighted average series can be compared as:

$$\left| n_i - n_{i+1} \right| >> \left| n_j - n_{j+1} \right| \qquad \text{(Eq. 9)}$$

Thus the fluctuations can be reduced further for a long term time series.

Lemma 2: The weighted average compared to the moving average reduces the data point values, thus provides a verification references to the moving average series.
Where,

$$\left\{ n_i, n_{i+1}, n_{i+2}, \ldots\ldots\ldots n_{i+k}, n_{i+k+1}, \ldots\ldots n_n \right\} \quad \text{(Eq. 10)}$$

Denotes the time series, where the indexes denote the items in the series in regular intervals.

$$\left\{ w_i, w_{i+1}, w_{i+2}, \ldots\ldots\ldots w_{i+k}, w_{i+k+1}, \ldots\ldots w_n \right\} \quad \text{(Eq. 11)}$$

Denotes, the time series weights for each data points

$$\left\{ n_j, n_{j+1}, \ldots\ldots\ldots n_{j+k}, n_{j+k+1}, \ldots\ldots n_{n-j} \right\} \quad \text{(Eq. 12)}$$

Denotes the weighted moving average series, where the indexes denote the interval of the weighted average for the series.

Proof: The weighted moving average is calculated with the bound of the weight, which increase the value of the data points. Thus,

$$n_i < n_i \bullet w_i \qquad \text{(Eq. 13)}$$

However, the weighted moving average series data points are calculated also with the bounds of weights as divided. Thus,

$$n_i < n_j / w_i \qquad \text{(Eq. 14)}$$

Henceforth with the influence of Lemma 1, it is proven that the weighted moving average method produces a modified series with much lesser data point values.

There further, this work proposes the novel predictive weighted moving average method. The proposed method calculates the moving average of the time series and produces the new series.

However, due to the limitations of the moving average method, the data points may tend to avoid the external events influence and making it arguable to be used for predictive analysis for short terms. Hence, this work deploys a novel multi-layer perceptron model and uses the weighted moving averages as weight of the MLP. The details of the proposed algorithm are demonstrated here:

Step-1. Accumulate the time series data points into a natural collection.

Step-2. Calculate the moving average of the time series and accumulate the new series into another natural collection independent of the nature of the of the data point values.

Step-3. Calculate the moving weighted average of the original time series. The new series will be considered as weights for the MLP.

Step-4. Lastly, calculate the final time series using MLP where the input is the moving average time series and the weights are the weighted moving average values.

Here the algorithm is realized mathematically,

The time series is represented as:
$$N = \{n_i, n_{i+1}, ...n_{i+k}, n_{i+k+1}, ...n_n\} \text{ (Eq. 15)}$$

The moving average is represented as:
$$K = \{n_j, n_{j+1}, ...n_{j+k}, ...n_{n-j}\} \quad \text{(Eq. 16)}$$

The weighted average is represented as:
$$W = \{n_j, n_{j+1}, ...n_{j+k}, n_{j+k+1}, ...n_{n-j}\} \text{ (Eq. 17)}$$

Thus, the proposed novel predictive weighted moving average is calculated as:
$$K_i' = \sum_{i=j, l=j}^{n-j, n-j} n_j \bullet w_l \qquad \text{(Eq. 18)}$$

The final data points are proven to more accurate for the time series and proven to more prediction ready proven using Lemma – 1 and Lemma – 2.

The next section demonstrates the configuration details of the neural network schema or the MLP configurations.

V. PROPOSED NEURAL NETWORK SCHEMA

The proposed multilayer perceptron model is made with the sole purpose to reduce the fluctuations and differences between the actual time series and after the smoothing. Henceforth here the work proposes the multilayer perceptron model [Figure – 3].
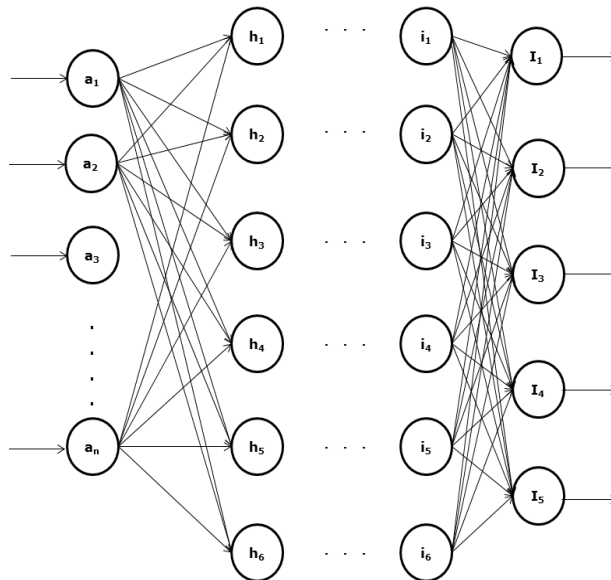

Fig. 3 Multi-Layer Proposed Multilayer Perceptron Model

The proposed MLP is arranged as the input layer is responsible for processing the inputs during the training; the hidden layers are available for considering the weight. The detail of the MLP is discussed in this section of the work [Table – 4].

TABLE IV: MLP CHARACTERESTICS

Attribute of MLP	Detail Description
Back Propagation Learning Rule, variable number of hidden layers	1 to 5 Layers
Random Number Seed	0
Learning Rate	0.1
Learning Rate Function	Static learning rate
Constant Bias Input	1.0
Training Iterations	500
Training Mode	Batch Training - weight changes are applied at the end of each epoch
Transfer Function	Sigmoid (Logistic), □S-shape function between +1 and 0
Momentum	0.2
Weight Decay	0.1
Bias Input Value	1.0
Inputs	36
Output Layer	5
Total Neurons	5
Total Nodes	226

VI. RESULTS AND DISCUSSION

Firstly, the comparative study of the moving average, weighted average and predictive weighted moving average is been carried out with a smaller data point values and with smaller term [Table – 5].

TABLE V: Evaluation - 1

Data Point Values	Weights	Average	Moving Average	Weighted Moving Average	Novel Method
1.0	1	3.7	1.0	1	1
4.5	2	3.7	2.8	1	2.75
6.0	3	3.7	3.8	1.833333333	3.833333
4.0	2	3.7	3.9	1.916666667	3.875
3.0	1	3.7	3.7	1.9375	4.375
4.0	2	3.7	3.8	2.055555556	4.25
5.0	3	3.7	3.9	2.045454545	4
4.0	2	3.7	3.9	1.964285714	4
3.0	1	3.7	3.8	1.96875	4
2.0	1	3.7	3.7	2.029411765	3.5

The assignments of weights are as per the following rule:
- Range 0.0 to 3.0   : Weight = 1
- Range 3.1 to 4.0   : Weight = 2
- Range 4.1 to 7.0   : Weight = 3

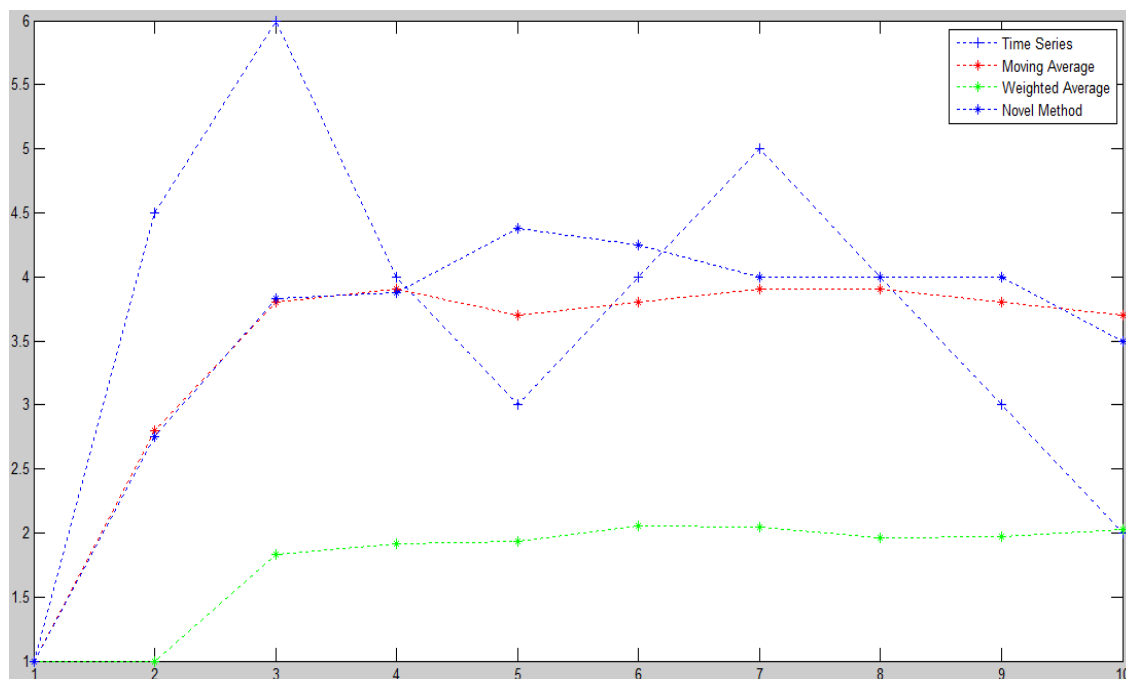The improvement can be visualized graphically [Figure – 4].

Fig. 4 Evaluation – 1 Comparisons

Secondly, the comparative study of the moving average, weighted average and predictive weighted moving average is been carried out with a higher data point values and with larger term [Table – 6]. The data set used in this evaluation is annual nonfatal disabling mine injuries 1930 – 1978 [15].

TABLE VI: Evaluation – 2

Data Point Values	Weights	Average	Moving Average	Weighted Moving Average	Novel Method
14160	3	11584.5	14160.0	1	14160
12163	2	11584.5	13161.5	4720	13161.5
11902	1	11584.5	12741.7	5264.6	12741.66667
11197	1	11584.5	12355.5	6370.8333	12355.5
10944	1	11584.5	12073.2	7060.2857	11551.5
11133	1	11584.5	11916.5	7545.75	11294
11070	1	11584.5	11795.6	7944.3333	11086
11183	1	11584.5	11719.0	8256.9	11082.5
10446	1	11584.5	11577.6	8522.9090	10958
10115	1	11584.5	11431.3	8683.1666	10703.5
9639	1	11584.52	11268.4	8793.3076	10345.75
9917	1	11584.52	11155.8	8853.7142	10029.25
11552	1	11584.52	11186.2	8924.6	10305.75
11916	1	11584.52	11238.4	9088.8125	10756
12329	2	11584.52	11311.1	9255.1176	11428.5
11220	1	11584.52	11305.4	8929.7894	11754.25
8545	1	11584.52	11143.0	9044.3	11002.5
11107	1	11584.52	11141.0	9020.5238	10800.25
14389	3	11584.52	11311.9	9115.3636	11315.25

14794	3	11584.52	11486.1	8597.08	12208.75
13554	3	11584.52	11584.5	8204.3214	13461

The assignments of weights are as per the following rule:
- Range 00000.0 to 12000.0   : Weight = 1
- Range 12000.1 to 13000.0   : Weight = 2
- Range 13000.1 to 16000.0   : Weight = 3

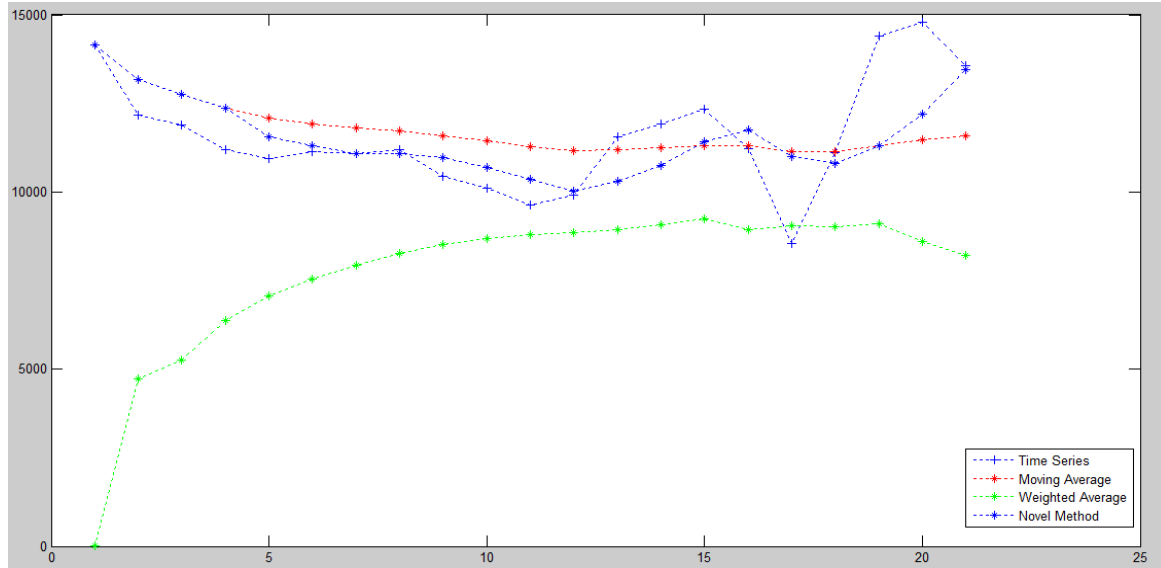The improvement can be visualized graphically [Figure – 5].

Fig. 5 Evaluation – 2 Comparisons

VII. CONCLUSIONS

This work analyses the time series and the properties of any time series which might influence the applied mathematical convolutions. With the carried out study in the space of time series, the nature of time series is been understood and it is also understood that due to the fluctuation nature of the series, it is difficult to analyze the series. Thus this work proposes the methods for smoothing the fluctuations. However with the light of the work it is understood that the standard methods like moving average and weighted average methods are prone to errors for various reasons. Thus this work proposes a novel MLP based method called predictive weighted moving average method for smoothing the data points or the fluctuations. The proposed method in this work demonstrates significant improvement in smoothing fluctuations and also considers the closeness of the data points with the original data points. Henceforth, this work contributes to the time series analysis in a significant way and makes the time series highly acceptable for further analysis in the space of predictive analysis for long and short term planning.

REFERENCES

[1]   R. Kalman "A new approach to linear filtering and prediction problems" Trans. ASME J. Basic Eng. vol. 82 pp. 34-45 1960.
[2]   A. Doucet S. Godsill C. Andrieu "On sequential Monte Carlo sampling methods for Bayesian filtering" Statist. Comput. vol. 10 no. 3 pp. 197-208 2000.
[3]   J. B. Tenenbaum V. de Silva J. C. Langford "A global geometric framework for nonlinear dimensionality reduction" Science vol. 260 pp. 2319-2323 2000.

[4]   M. Belkin P. Niyogi "Laplacian eigenmaps for dimensionality reduction and data representation" Neural Comput. vol. 15 pp. 1373-1396 2003.

[5]   D. L. Donoho C. Grimes "Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data" Proc. Nat. Acad. Sci. vol. 100 pp. 5591-5596 2003.

[6]   R. Coifman S. Lafon "Diffusion maps" Appl. Comput. Harmon. Anal. vol. 21 pp. 5-30 Jul. 2006.

[7]   R. Talmon R. Coifman "Empirical intrinsic geometry for nonlinear modeling and time series filtering" Proc. Nat. Acad. Sci. vol. 110 no. 31 pp. 12 535-12 540 2013.

[8]   T. Berry J. Harlim "Semiparametric forecasting and filtering: correcting low-dimensional model error in parametric models" J. Comput. Phys. vol. 308 pp. 305-321 2016.

[9]   T. Berry D. Giannakis J. Harlim "Nonparametric forecasting of low-dimensional dynamical systems" Phys. Rev. E vol. 91 no. 3 2015.

[10]   T. Berry J. Harlim "Nonparametric uncertainty quantification for stochastic gradient flows" SIAM/ASA J. Uncertainty Quantification vol. 3 no. 1 pp. 484-508 2015.

[11]   T. Berry J. Harlim "Forecasting turbulent modes with nonparametric diffusion models: Learning from noisy data" Phys. D: Nonlinear Phenom. vol. 320 pp. 57-76 2016.

[12]   R. Talmon R. R. Coifman "Intrinsic modeling of stochastic dynamical systems using empirical geometry" Appl. Comput. Harmon. Anal. vol. 39 no. 1 pp. 138-160 2015.

[13]   B. Nadler S. Lafon R. Coifman I. G. Kevrekidis "Diffusion maps spectral clustering and eigenfunctions of Fokker-Planck operators" Proc. Neural Inf. Process. Syst. vol. 18 pp. 955-962 2005.

[14]   R. Coifman et al. "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps" Proc. Nat. Acad. Sci. vol. 102 no. 21 pp. 7426-7431 May 2005.

[15]   Data Set Source: Time Series Data Library (citing: McCleary & Hay (1980)).

[16]   Amjan Shaik et al, "WCS: A Metric to predict the fault proneness of OO Applications", IJARCS, ISSN : 0976 - 5697, 2012.