International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-5, Issue-6)

Research Article | **June 2016**

Strong Correlation Technique for Detecting Relationships in Time Series Applications for Univariate Analysis

P. Sridevi, M. Bhanu Sridhar
Department of CSE, GVP College of Engineering for Women,
Visakhapatnam, India

Abstract–

Checking for association between attributes of data objects in large machine learning projects generally results in wrong analytical consequences. Since traditional correlation is used in predicting models, it may lead to undesired outputs. Many time series datasets exhibit time interdependency among their values. To improve the prediction quality of the model, it is important to detect the real relationships and correlations among the variables. Strong correlation is when the auto-dependencies of both variables exhibit a correlogram pattern. The bumpiness metric, which is considered as a strong technique, is simple and involves only one parameter and can be viewed as an improved perception over the traditional regression. Bumpiness coefficient is also much helpful in Time Series datasets where data objects are ordered over a period of time. In this paper it is proposed to apply the bumpiness metric to capture the association between two variables for any auto-dependencies between them. The paper, through its results, points out the importance of the metric in applications of Machine Learning in Big Data.

Keywords– Autocorrelation, bumpiness metric, Univariate data, relationships, time series.

I. INTRODUCTION

Science and society are often interested in the relationship between two or more variables [1]. As an example, it is fine to what the typical weight is, when we are interested in typical height. When relationships between two variables are considered, it is better to conclude if the 'relationship' is positive or negative, linear or quadratic or something else. It should also be noted that autocorrelation has its own say in the relationships between variables.

Correlation in Statistics may refer to any extensive class of statistical relationships between two random variables, involving dependence. Autocorrelation can be broadly stated as the similarity between observations as a function of the time lag between them. In whatever type of time series data we utilize, it can depict substantial autocorrelation [2]. Autocorrelation helps in detecting non-randomness of the data and is also useful for the purpose of identifying an appropriate time series model(s) for computations, deductions and predictions.

Regression is used to find the relationship between the variables X and Y in a data set [4]. Generally, researchers use mean and variance as metrics to find the centrality and volatility of the data objects. This is the traditional approach which can also be a weak approach. Instead of this, many other metrics can be used to compare the relationship between two variables. Here in this paper along with regression a synthetic metric is introduced, which is called Bumpiness metric that is designed to avoid overfitting and is also outlier resistant because extreme values of bumpiness are not considered.

In this article the bumpiness metric is used to detect the autocorrelation between two considered attributes of the time series data. The time variable X is not used in the computation and is assumed that observations are equi-spaced. The paper is organized as follows: Section 1 of the paper deals with introduction; in Section 2, the time series data is presented and discussed; the bumpiness metric is presented in Section 3 along with its advantages over similar other metrics along with autocorrelation; Section 4 deals with the data together with experimental results and finally the conclusion is presented in Section 5.

Our future work is concerned with application of bumpiness metric on multi-variant medical data for finding spurious relationships between the concerned attributes. This is currently at research stage and would be very useful in aiding the ailing patients to become an important part in the general usage of the Doctors to keep the lamp of life burning.

II. TIME SERIES DATA

Time Series data accounts for the data points taken over a period of time [5]. This data can be observed for any internal structure such as autocorrelation. Time series data is used for identifying patterns in the observed data.

Univariate distribution is a distribution of one variable and a multivariate distribution is distribution of several variables. Univariate refers to a function or a polynomial of only one variable. Objects involving more than one variable may be called multivariate. Univariate time series refers to a set of values over time of a single attribute. Multivariate time series refers to the changing values over time of several attributes. In univariate analysis the response variable is influenced by single factor whereas in multivariate analysis the response variable in influenced by multiple factors.

The analysis of time series is based on the pretext that successive values in the data represent measurements taken at equally spaced time intervals. In the rapidly growing field of time series modeling and analysis a scarcely used metric called Bumpiness factor is used to calculate the correlation within the same variable. Time series data set will contain the time as *x* and the variable as *y* though *x* is not considered but assumed to be equispaced.

For example, the time series data for Carbon dioxide (CO_2) concentration in air noted over a period of one year may look like

Table 1: Time Series Data for CO_2

YEAR	MONTH(X)	CO_2 (Y)
1995	Jan	330.62
1995	Feb	331.40
1995	Mar	331.87
1995	Apr	333.18
1995	May	333.92
1995	Jun	333.43
1995	Jul	331.85
1995	Aug	330.01
1995	Sep	328.51
1995	Oct	328.41
1995	Nov	329.25
1995	Dec	330.97

Time series analysis (TSA) is done for identifying the nature of the sequence of observations and predicting the future values of the time series variable. It is used for the determination of patterns and the presence of noise in the temporal ordering among the variable. Business Intelligence Tools are developed for TSA because of the availability of huge marketing data sources and software packages like SPSS.

III. BUMPINESS METRIC AND AUTOCORRELATION

3.1 Bumpiness Metric

Median, Mean and Variance are the metrics used in statistics to find the correlation between variables. Centrality (mean, median) and Volatility (variance) are widely used statistic measure to find the correlation between variables. These measures ignore the order of the data in time series analysis where order plays an important role. It is assumed that the time ordered observations are recorded for time series analysis. Bumpiness metric which is another class of scarcely used metric can be useful in time series data sets where order plays an important role. Bumpiness also supports the notion of dependence among the data points.

3.2 Autocorrelation & Bumpiness

Autocorrelation of lag one can be considered as Bumpiness. Autocorrelation can be used to detect non-randomness in data and also to identify appropriate time series model for non random data [6].

Bumpiness can be denoted as r. and can be calculated by [7]

$$r_k = \frac{\sum_{i=1}^{N-k}(Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

Autocorrelation assumes the time variable is equi-spaced. Instead of correlation between two different variables correlation is found between two values of the same variable at times X_i and X_{i+k}. where k is the lag. Bumpiness is autocorrelation lag k where k =1. Auto correlation plotted for lag 1 to k (k=1,2,3,4,5,6,7,8) can show time series model.

3.3 Applications of autocorrelation

- Autocorrelation is used to measure steel-concrete beam deflections in construction industries.
- Autocorrelation analysis for filter transmittance
- In signal processing information about repeating events like music beats.
- spatial autocorrelation between sample locations

IV. RESULTS

The results provided below have been obtained by applying the Bumpiness Metric on certain stock data and beam deflection data. In each of the figures a clear picture of the autocorrelation is obtained by deducing the autocorrelation function (ACF) value and checking whether it is nearer to 0 or 1. A stock data sample is also provided to make the picture clearer.

Table 2: Stock Data

S. no.	Date (X)	Price (Y)
1	5-Jan-09	0.035754
2	6-Jan-09	0.025426
3	7-Jan-09	-0.02886
4	8-Jan-09	-0.06221
5	9-Jan-09	0.00986

6	12-Jan-09	-0.02919
7	13-Jan-09	0.015445
8	14-Jan-09	-0.04117
9	15-Jan-09	0.000662
10	16-Jan-09	0.022037
11	19-Jan-09	-0.02269
12	20-Jan-09	-0.01371
13	21-Jan-09	0.000865
14	22-Jan-09	-0.00382
15	23-Jan-09	0.005661
16	26-Jan-09	0.046831
17	27-Jan-09	-0.00663
18	28-Jan-09	0.034567
19	29-Jan-09	-0.02053
20	30-Jan-09	-0.00878
21	2-Feb-09	-0.02592
22	3-Feb-09	0.015279
23	4-Feb-09	0.018578
24	5-Feb-09	-0.01413
25	6-Feb-09	0.036607
26	9-Feb-09	0.011353
27	10-Feb-09	-0.04054
28	11-Feb-09	-0.02211
29	12-Feb-09	-0.01489
30	13-Feb-09	0.007027
31	16-Feb-09	-0.01149

CASE-1 Stock Data

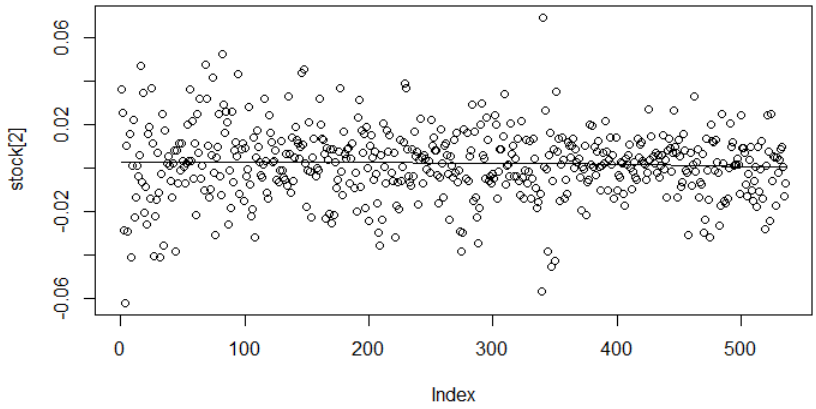

Fig.1: Scatter plot showing randomness and outliers

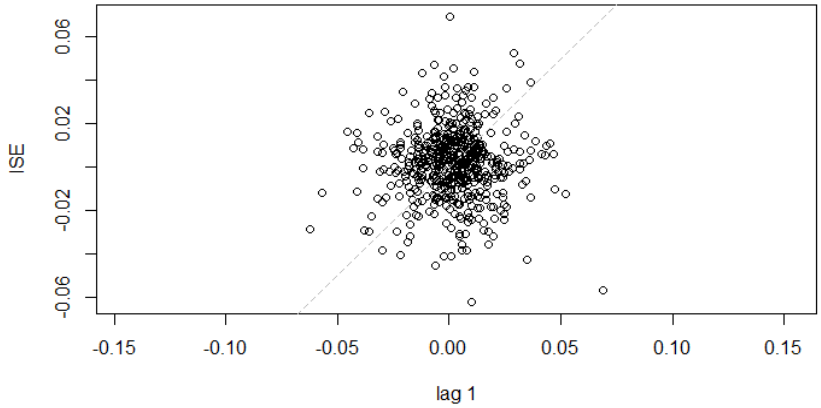

Fig. 2: Random data with no underlying pattern 1

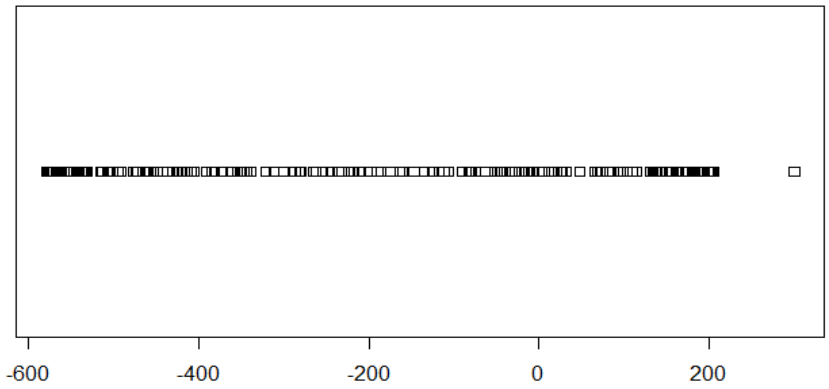

Fig. 3: Fixed variation

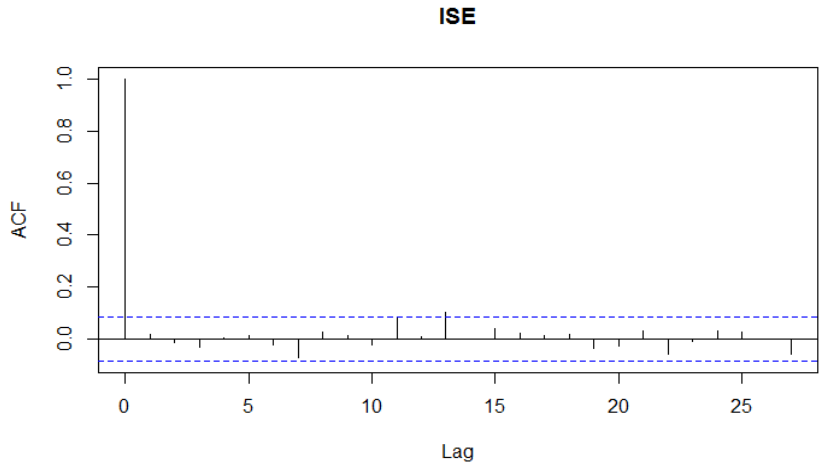

Fig. 4: Data exhibits no auto correlation

CASE-II: Beam Deflections

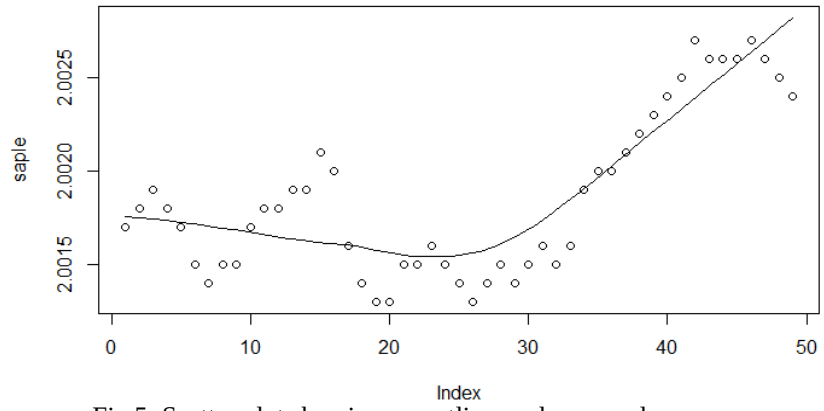

Fig.5: Scatter plot showing no outliers and non randomness

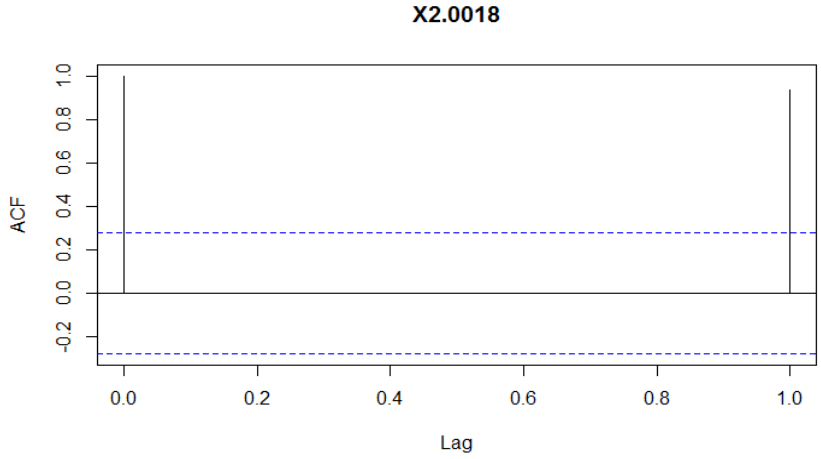

Fig. 6: Autocorrelation plot of non-randomness showing an observed pattern in data

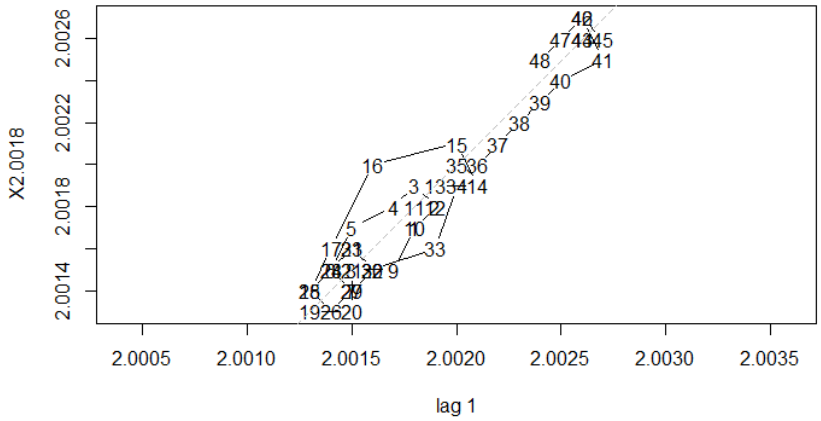

Fig. 7: Correlogram Identifying a Pattern

V. CONCLUSION

In this paper, two data sets are utilized and it has been observed that using bumpy metric we can detect the patterns in the data by correlogram images (figure 2 and figure 7 autocorrelation plots). If the data is random the autocorrelation value will be close to zero. Here it can be observed that in the first data set ACF Lag1 is 0.019 (close to zero and randomness, Figure 4) and in the second dataset the ACF is 0.937 (close to one as shown in figure 6) which detects non-randomness. Using bumpy metric, meaningful patterns can be derived which can provide the insights into Univariate data analysis.

Though traditional metrics like mean and variance can be used, by means of bumpiness metric, we can derive meaningful patterns from time series data where the observations are recorded in an orderly manner. The potential use of bumpiness metric is felt in time series applications apparently and in natural language processing where order is of much concern. Lag one auto correlation is the highest used of all autocorrelation formulae. It is the single best indicator of the autocorrelation structure of time series. It is close to 1 for very smooth time series, close to 0 for pure noise and very negative for periodic time series. This can also be considered the greatest advantage by going through the results.

The next stage of this research plans to apply this methodology on multi-variant data to identify the spurious relationships between the variables of the dataset. Further, it is planned to carryon this work onto medical data.

REFERENCES

[1] www.stat.ucla.edu/~rgould/m12s01/relations.pdf
[2] Friston, Karl J., Peter Jezzard, and Robert Turner. "Analysis of functional MRI time-series." *Human brain mapping* 1.2 (1994): 153-171.
[3] Singh, Rohit, et al. "Active learning for sampling in time-series experiments with application to gene expression analysis." *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005.
[4] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, 2001.
[5] Liao, T. Warren. "Clustering of time series data—a survey." *Pattern recognition* 38.11 (2005): 1857-1874.
[6] Lichstein, Jeremy W., et al. "Spatial autocorrelation and autoregressive models in ecology." *Ecological monographs* 72.3 (2002): 445-463.
[7] Sridhar, M. Bhanu, Y. Srinivas, and M. H. M. Prasad. "Software reuse in cardiology related medical database using K-means clustering technique."*arXiv preprint arXiv:1311.1197* (2013).

