Big Data Time Series Analysis: An Approach with the Size of the Time Series

Jitendra Kumar, Saurabh Kumar
Department of Statistics & Central University of Rajasthan, Ajmer, Rajasthan, India

DOI: 10.23956/ijarcsse/V7I6/0220

Abstract— Data analysis has importance in real-time data. It is collective representation of all the facts of population or variables under interest. As velocity of data generating process (DGP) is third very important identity in Big Data prospective, so this cannot be managed through routine statistical tools of analysis. Big data time series recording is in principle on packet of information in every unit interval of time and increase size of the series. For that reason modelling can be performed after adding the respective information. This gives us challenges to take a decision regarding size of the series for understanding the data generating process. Present paper is proposing an alternative approach of time series analysis in which data generating process is studies based on size of time series. Here, we discuss the convergence of model in term of order. An empirical analysis is carried out on the time series of sectoral indices of National Stock Exchange (India) to show the convergence of the model. This convergence is also advocating the fact that if a series is fitted after certain size, we get a constant model. This supports that true model forecasted rightly the future value and also data is always convergent in nature. Study reveals that the realization of big data time series may be in considered in respect to size of the series, so identified DGP can be best representative intervals for respective study period.

Keywords— Time series; Big Data; Velocity of Big Data; ARIMA model; GARCH model.

I. INTRODUCTION

The process of research into massive amounts of data to disclose hidden patterns and existing secret relationship between them named as big data analytics. Since late 1990’s it became more important after popularization of Big Data. Beyer quotes Gartner’s definition that “Big Data is due to 3V’s framework – high volume, high velocity and high variety of information” [1]. IBM added one more dimension namely veracity to add trust and noise filtering. There are several areas where Big Data came into the picture like medical records, scientific research, government, natural disaster and resource management, private sector, military surveillance, financial services, retail, social networks, mobile phones, sensor networks, telecommunications etc. Russom defined Big Data Analytics (BDA) as a technology and framework for storing, converting, transferring and analyzing massive amounts of constantly updated unstructured data for commercial gain [2]. There are several technical challenges to deal massive dataset, such as data acquisition, data recording, information extraction, query process, representation and analysis due to uncontrolled size of recorded values, continuous generation due to velocity, variation and variety. Labrinidis et.al. had comprehensive discussion on such type of challenges in reference to analysis [3].

A data recorded in chronological order is called Time series which is discussed for the better forecasting. There are sufficient techniques/tools to know the data generating process but if series is real time generated i.e. there is additional on the observation then it cannot be analysed in simple way. A Big data time series is continuous process which may be increasing due to various reasons. This may be termed as real time data, this data not is MB sometimes and it is in terabytes because of the new technology [4]. As there is addition on observation due to velocity which increase the too large and which becomes beyond the capacity of storing software/resource mostly and advocates it to be a Big Data. Therefore modeling of Big Data time series, it is required to fit a large size time series with the provision to manage the velocity which cause of naming this real time generation. Usual time series model now need to introduce for big data time series modelling. The most popular methodology for linear time series model is Box-Jenkins approach [5, 6]. This concept may be useful for the Big Data time series modelling as well as further analysis and interpretation [7]. The time series approaches to convergence for check for the compatibility of difference in (log) output with an in deterministic stationary series [18].

Therefore, we are targeting to establish new approach where time series may be realized on current data and continue the process till sufficient information about the data generating process can be gathered. In the present paper, we are proposing model convergence concept in respect to size of the series, where fitted model is same after at certain size. The model is evaluated based on forecasted band and get that if there is model convergence then current values are within the forecasted range then converge model is best model for forecast the future value otherwise searching another model. The non-convergence of model may be due to inappropriate model selection or information has not achieved the limiting model.

© www.ijarcsse.com, All Rights Reserved
II. CHALLENGES OF BIG DATA TIME SERIES MODELLING

There are/was lots of technologies which can be benefitting and simplifying the analysis in many aspects within the underlined assumption of analytics. Present era is known as information era and all are looking solution through Big Data because it considers from as many as possible sources [8]. The Information collected digitally is increasing rapidly at the rate of 10x approximately in every five year. There are technical solutions and scientist from computer science is working to develop best possible space. Earlier computer was 2.9x10^21 bytes (optimally compressed) and that for communication was 2.0x10^21 bytes only but is 2011it started accommodating 6.4 x10^18 instruction per sec. However computational size is increasing annually more than by 50% rate [9]. There is sufficient work at technical platform but very less work at analytical platform because there is very high rate on increasing the actual data.

Time series is a sequence of numerical data point in chronological order and Big Data time series is realization of such data from all possible associated realization which may be univariate as well as multivariate. The only difference with usual time series and big data is real time generation. As we know that realization is the process to extract the information about the generated data and fitting the model with the help of ACF, PACF and then after verify the assumptions and diagnosis of the residuals of the linear time series models for the particular series [10]. The real time generation increase the time series also may not be managed with the available resources.

Big data bears the properties of real time generation with huge volume. Huge volume creates the problem in respect to software and hardware capacity but real time generations creates problem in statistical framework because statistics deals the fix data. In addition of this most vital challenge is due to volume as well as velocity. Analytics is handling of real time analysis within the limitation of time which must be completed within a specified fixed interval [11]. Main motivation behind the time series is modelling/analysis to understand the data generating process and estimate the true information about the generated data and fitting the model with the help of ACF, PACF and then after verify the assumptions and diagnosis of the residuals of the linear time series models for the particular series [10]. The real time generation increase the time series also may not be managed with the available resources.

Big data bears the properties of real time generation with huge volume. Huge volume creates the problem in respect to software and hardware capacity but real time generations creates problem in statistical framework because statistics deals the fix data. In addition of this most vital challenge is due to volume as well as velocity. Analytics is handling of real time analysis within the limitation of time which must be completed within a specified fixed interval [11]. Main motivation behind the time series is modelling/analysis to understand the data generating process and estimate the true information about the generated data and fitting the model with the help of ACF, PACF and then after verify the assumptions and diagnosis of the residuals of the linear time series models for the particular series [10]. The real time generation increase the time series also may not be managed with the available resources.

III. STATEMENT OF THE PROBLEM AND METHODOLOGY

In the definition of Big Data, velocity is important feature which increases the size of time series. Fan et.al. has commented on dimensionality which creates a practical problem and say that as the dimensionality increases then corresponding heavy computational cost also hike as well as question that it is mandatory to have the analysis of overall data and it is contributing equally [12]. Since the inception of discussion about the big data analytics, researchers are concerned to improve the analytical methods and also minimize the processing time. Here following question are important in respect to analytics:

- Is it sufficient/ necessary to handle complete series?
- Does a large size time series provide more accurate information for a specific model?
- Does the efficiency of the model increase by adding more observation in the sample series?
- Is there any opportunity to get more accurate information about the system/object under study?

All above question are basically talking about the size of time series i.e. the selections of interval which can be best represent the realization of a time series data in a better way. The problem of convergence is a natural phenomenon in case of statistical analysis, so first of all we are interested about the convergence capability of the data series i.e. whether it is converged or not? If yes then what is the limit/time of convergence? In some concept, one may study the impact of size on this DGP. It also gives an opportunity to explore the facts which are responsible for changing the parameters by increasing the sample of size. After identifying the effective factors for parameter changes, we can shift to the process of forecasting. This will also give an idea about the future observation despite the analysis of real time data. The monitoring of analytical results may be carried out till the collection of information based on identified model or as per defined/estimated parameters of the system/object under study.

IV. PROPOSED MECHANISM FOR BIG DATA TIME SERIES

In routine Big Data analysis, we analyse a long series without taking care of size of series without any support that the selected sample is most efficient representation of the whole series. There is always system convergence if there is no change in system within the underlining objects of analysis. Simple, we may say that data is modelled with the principal to identify the factors which are affecting the model. There are two aspect of Big Data time series analysis first size and second how to manage this real time generating data. Major challenges in modeling of big data is to understand the localized patterns and also capture the pattern of the data which is due to real time generation for the analytics. The flow chart of the proposed methodology is given in next page.

There are sufficient literatures to handle a simple as well as complex data and small as well as huge volume of data through technical power. A real time generated data can be simply managed if we get the information through realization of series. If the variation due to additional information is within the limits of variation, defiantly change of model will not be there. If change of the model is not there, then further analysis cannot give any additional information and this is misuse of respective resources. Therefore, present study is proposing methodology in which optimum size of the data for the purpose is identified.

Due to large sample size, understanding the heterogeneity pattern is quite difficult and also allows us to expose hidden patterns associated with small subpopulations and weak commonality across the whole population in current scenario [12]. Heterogeneity appears in big data because it contains several variables under study for multivariate & univariate and change in respect to time also. This is prone to increase the heterogeneity. In classical settings sample size is small or
moderated, data points are taken from small subpopulation, where techniques can easily identify the small variation but in the big data era. In large sample size one may not understand the small variation due to existing heterogeneity.

Fig. 1 Methodology flow chart

V. EMPIRICAL ANALYSIS OF NSE DATA

Market indices play an important role in the ranking of economy of a country and for this purpose each and every country having it’s own administrative setup to declare indices. India declares indices through BSE, NSE, USA through NASDAQ OMX, NYSE EURONEXT, Canada through TMX GROUP, China through Shanghai SE Shenzhen SE, UK through London SE group etc. [13]. Trading in Indian derivatives market is mainly done at National Stock Exchange (NSE) and it is declaring sixteen indices in reference to different sectors and company profile like broad market indices, sectoral indices, thematic indices and strategy indices. Similarly, sectoral indices auto, Bank, Energy, FS, FMGC, IT, Metal and PSU etc.

In the present study, an empirical analysis of proposed mechanism for Big Data time series modelling, we have taken eight indices recorded for the period January 1994 to January 2017 from National Stock Exchange (NSE) website [14]. These are declared daily, so velocity may be considered in respect to daily addition. In this process addition of the data in parent series is less but let’s suppose that addition on in each and every fraction of second. In that case velocity is very
high, so modelling of such type of series is very difficult and classical statistical tools are unable to do the analysis. We are realizing the series individually considering a big data time series by normal computing machines as daily there is addition of one more observation in the series which is equivalent of velocity of Big data. In practical digital data come in this situation in thousands or MBs but here we are interested to train the methodology for the analysis. Here, we may argue that word “Big” is very subjective and relative, however we are analyzing this by a simple i-7 processor and also modeling the series after adding the information in an interval equal to a month.

For numerical illustration we have calculated stock return of sectoral indices using the following formula [15].

\[ y_t = \left( \log P_t - \log P_{t-1} \right) * 100 \]

Where, \( y_t \) is return at time \( t \), \( P_t \) and \( P_{t-1} \) are daily closing price of time \( t \) and \( t-1 \) respectively. For computation purpose, we have added monthly return series in form of bundle (set of monthly observations). In the time series context, assuming that past structure will be preserved in future then to forecast the means. Time series are models which predict future values of a random variable based on the structure defined by the variable i.e. mean past observations are explaining the pattern as well as variation. The autoregressive (AR) and Moving Average (MA) are two basic model considered in time series to explain the correlation structure in stationary series. A series which is depending on its own lagged values such type of process is called an Autoregressive (AR) process. Autoregressive model is represented as

\[ y_t = \mu + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_3 y_{t-3} + \ldots + \alpha_p y_{t-p} + \epsilon_t \]

Autoregressive Moving Average model is mixture of autoregressive and moving average model and it is denoted as ARMA (p, q). The mathematical equation of the ARMA model is

\[ y_t = \mu + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_3 y_{t-3} + \ldots + \alpha_p y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \theta_3 \epsilon_{t-3} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t \]

If series is non stationary in nature then we use the Autoregressive Integrated Moving Average (ARIMA) model. It is a linear non stationary model. In order to get stationary series from non-stationary we use difference operator (d). It is denoted as ARIMA (p, d, q) and model can be represented as

\[ \left( 1 - \sum_{i=1}^{p} \alpha_i B^i \right) \left( 1 - B \right)^d y_t = \left( 1 + \sum_{i=1}^{q} \theta_i B^i \right) \epsilon_t \]

Where, \( B \) is backward operator, \( \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p) \) and \( \theta = (\theta_1, \theta_2, \ldots, \theta_q) \) are coefficient of the autoregressive and moving average model. Due to availability of computer software of time series model and high performance computer, more and more forecasting problem can handled using these models. The best ARIMA model is obtained by Box Jenkins approach [16] and also tested the stationarity by Augmented Dickey Fuller (ADF) test and obtained test statistics values are recorded in Table-1 with descriptive statistics.

<table>
<thead>
<tr>
<th>Series</th>
<th>Mean</th>
<th>SD</th>
<th>Skewness</th>
<th>ADF-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto</td>
<td>0.0685</td>
<td>1.5285</td>
<td>-0.2438</td>
<td>-13.5727**</td>
</tr>
<tr>
<td>BANK</td>
<td>0.0594</td>
<td>2.0013</td>
<td>-0.1274</td>
<td>-13.9191**</td>
</tr>
<tr>
<td>ENERGY</td>
<td>0.0314</td>
<td>1.6860</td>
<td>-0.5905</td>
<td>-13.7322**</td>
</tr>
<tr>
<td>FS</td>
<td>0.062</td>
<td>1.9203</td>
<td>-0.0849</td>
<td>-13.8406**</td>
</tr>
<tr>
<td>FMC</td>
<td>0.0605</td>
<td>1.3401</td>
<td>-0.4305</td>
<td>-14.1703**</td>
</tr>
<tr>
<td>IT</td>
<td>-0.0258</td>
<td>4.4885</td>
<td>-44.8896</td>
<td>-14.3722**</td>
</tr>
<tr>
<td>Metal</td>
<td>0.0302</td>
<td>2.3244</td>
<td>0.1739</td>
<td>-13.3198**</td>
</tr>
<tr>
<td>PSU</td>
<td>0.0339</td>
<td>2.2472</td>
<td>-0.1509</td>
<td>-13.8522**</td>
</tr>
</tbody>
</table>

* **5 % level of significance

In the above Table 1 we observed that the test statistics value is significant at 5% level of significance. The mean and standard deviation of return series of IT sector is negative and high as compare with other sector. All are negatively skewed (except metal sector) and the value of Skewness of the IT sector is not near to zero (-44.8896) it means except IT sector return series all series are approximately normally distributed. In order to testing the stationarity of the return series our null hypothesis is that the series is having unit root or follow non stationarity behaviours. Based on ADF test, we find that all sectorial return series are stationary in nature. In chronological recorded series is becoming large continuously in real time so we are fitting the model after adding the one month data and recorded order of the best fitted model. If there is no change on the order of the series, we say that series is in model convergent.

The important part of the methodology is here that we stop the fitting of series when get that there is no change on order of model. For empirical analysis, we used ARIMA model because it can take into account trends, seasonality, cycles, error and non-stationary aspect. In order to getting best ARIMA model we use forecast package in R-software. A cumulative series will create after adding monthly observations in original series. The best ARIMA model fitted on cumulative series and test ARCH/GARCH effect presents in residual series. We have 151 months of return series observations i.e. 151 bundle. The frequency of best ARIMA model order represented with the help of Bar diagram and shown in below Fig. 2.

© www.ijarcse.com, All Rights Reserved

Page | 770
In order to fitting and testing the ARCH effect at each and every steps of cumulative sectorial indices series. We perform the Ljung-Box test and test statistics is Q-statistics at lag 12 and ARCH test. This test is basically testing the presence of autocorrelation in residual and indicates the inefficiency of the model for forecasting purpose whereas presence of autocorrelation in square residual indicates the existence of ARCH effect. If such type of correlation exist in the residual series then basic assumption of the model not satisfies, and in this situation efficiency of the model is reduces. One may conclude that this model is not suitable for prediction of future values. In generally financial data having ARCH effect, in our case the all return series showed ARCH effect, so here we considered GARCH(1,1) model for controlling the variation in the residuals series and the specification for GARCH(1, 1) can be written as [17]:

\[ \varepsilon_t = \sigma_t z_t, \quad z_t \approx N(0, 1); \quad t = 1, 2, \ldots, T; \]

\[ \sigma_t^2 = \omega + \phi_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \]

Now these days many statistical computing software are used by data scientist to analysing the real time generated data. Among those software R language is broadly used by statistician and data miners for developing the statistical packages. Proposed methodology may be helpful for analysing such type of huge dataset. In R software many packages was developed for modeling the time series data, “forecast” package one of the basic package for analysing univariate time series model and for GARCH modelling we use “rugarch” as well as “FinTs” package. All these are analysis the time series which is containing fix number of observation i.e. size of series and not able to manage the real time generation phenomenon of Big Data. However proposed methodology is giving scop to analyse the Big Data time series considering a check point to record the model only after convergence. The converge order and coefficient of the cumulative series under study is recorded in Table 2 and Table 3.

In Table 2 we found that in the IT sector model is converge very soon as compare with other. After 2892 observation the PSU return series converged with order ARIMA(2, 0, 0). The coefficient of the converged model has been recorded in Table 3.
Table 3 Converge model Coefficient For sectorial return series

<table>
<thead>
<tr>
<th>Series</th>
<th>ARIMA + GARCH(1,1)</th>
<th>$\mu$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
<th>$\alpha_4$</th>
<th>$\alpha_5$</th>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\omega$</th>
<th>$\phi_1$</th>
<th>$\beta_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto</td>
<td>(0,0,1)</td>
<td>0.1145</td>
<td>**</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.1355</td>
<td>0.0849</td>
<td>0.1097</td>
<td>0.8689</td>
<td></td>
</tr>
<tr>
<td>BAN K</td>
<td>(0,0,1)</td>
<td>0.1258</td>
<td>***</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.1343</td>
<td>0.0669</td>
<td>0.0924</td>
<td>0.8949</td>
<td></td>
</tr>
<tr>
<td>ENERGY</td>
<td>(5,0,2)</td>
<td>0.06 **</td>
<td>1.04 28</td>
<td>0.97 91</td>
<td>0.03 98</td>
<td>0.00 25</td>
<td>-0.00 26</td>
<td>-0.9906</td>
<td>0.92 64</td>
<td>0.0437</td>
<td>0.1155</td>
<td>0.8758</td>
</tr>
<tr>
<td>FS</td>
<td>(1,0,1)</td>
<td>0.1506</td>
<td>***</td>
<td>0.83 39</td>
<td>0.97 91</td>
<td>0.03 98</td>
<td>0.00 25</td>
<td>-0.6998</td>
<td>0.12 82</td>
<td>0.0823</td>
<td>0.1028</td>
<td>0.8787</td>
</tr>
<tr>
<td>FMC G</td>
<td>(2,0,2)</td>
<td>0.1118</td>
<td>***</td>
<td>0.38 44</td>
<td>0.99 77</td>
<td>0.03 98</td>
<td>0.00 25</td>
<td>0.3838</td>
<td>0.99 29</td>
<td>0.239</td>
<td>0.2623</td>
<td>0.6564</td>
</tr>
<tr>
<td>IT</td>
<td>(0,0,0)</td>
<td>0.1407</td>
<td>***</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-0.3874</td>
<td>-0.3874</td>
<td>0.3874</td>
<td>0.2286</td>
<td>0.6669</td>
</tr>
<tr>
<td>Metal</td>
<td>(0,0,1)</td>
<td>0.3515</td>
<td>***</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0417</td>
<td>1.4238</td>
<td>0.3963</td>
<td>0.4814</td>
<td></td>
</tr>
<tr>
<td>PSU</td>
<td>(2,0,0)</td>
<td>0.072</td>
<td>*</td>
<td>0.12 64</td>
<td>-</td>
<td>0.00 87</td>
<td>-</td>
<td>-0.1266</td>
<td>0.776</td>
<td>0.8981</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*10 % level of significance; ** 5% level of Significance; *** 1 % level of significance.

We have forecasted confidence interval suing the convergent ARIMA and ARIMA+ GARCH(1, 1) model for 30 days. The system is control or not it is decided with help of forecasted confidence interval and original return series. If the original series is cross the forecasted confidence interval then we say that converge model may be changed. In this situation we again search best converge model. The confidence interval and original series of January 2017 month is plotted in the below Fig. 3.

In the above Fig. 3 one can see that Auto, FS, FMCG, Metal and PSU sectorial indices cross the ARIMA+GARCH upper limits that indicate that the best fitted ARIMA + GARCH model need to be modified and search another best model. The sectorial return series of BANK and ENERGY lies within the ARIMA as well as ARIMA + GARCH confidence interval.
The process is control for these two sectorial indices not need to search best ARIMA+ GARCH model for coming months. Using this model one can easily forecast the future return value. The motive of the present study to find size of the series where model is stable. Here, evaluation of forecasted data is also made to have the confidence that our model is predicting correct parameters as well as forecasting future values.

VI. LIMITATIONS OF PROPOSED MECHANISM

A. More centric on the Size of series
Since the inception of big data, it is very clear that if a data cannot be managed by well-equipped hardware unit as well as improved analytical system for a real time generated time series then understanding the DGP at two dimensions first data generating process and it’s impact on parameters of the process due to real time generation. Most of the analytical tools are in view that a software/hardware can handle huge volume data to provide the solution, however size of handled data may not be representing for whole population. Present paper dealt the importance of size of data in reference of Big data time series model only however some more important dimension may also be equally important like other variables are effecting time series considering all variable as well as significance in respect to model and efficiency of parameters. These are not evaluated in the study.

B. Ignorance of shift of trend:
Present study is aiming to capture the data, which can realize a real time generated series. As long as series is generated, there may be trend shift as well as change on the distributions of the analysis. These can change the parameters of the model as well as generating distribution. This aspect of time series is not taken care at this instance. One can extend this for other most suitable model and then do the forecasting as well as interpretation of Big Data time series.

C. Competetiveness of the proposed model
In the present study, we have modelled the series to understand the data generating process in respect to parameters and impact of size on the series parameter. So we can forecast series parameters as well as series in reference of Big Data time series. As the velocity of Big Data time series can be the managed only through the nature of the parameter. Here, we have used the ARIMA approach and studied the parameter in reference to addition of the information i.e. size of the series.

VII. CONCLUSION

Convergence is natural phenomenon of the system and we may also be interested to find this for time series model. Simply, we can say that if additional information of the data is not able to shift the trend of the series or not contribute anything in respect to model as well as values of parameter(s). However, there may be change on other parameters of the model like standard error, information criterion but this may not add more as per objective of modeling. In practice, Big Data analysis tools are able to analyse huge volume data or too large time series without any evaluation of the model in respect to size. Present paper has investigated the importance of size in the analysis of time series in place of all observation or optimal size. Methodology gets justified for the NSE data. Here we have observed that if we get convergence on the model, only achieve to get the forecasted values within the forecasted range however if not then need to explore for getting the convergent model. If size of the series is not taken care it may cause the extra cost of the analysis in respect to recording of the data, cost on analytics as well as cost of hardware and software tools. The work may be extended for the generalized model considering the other models as well as other associated variables.

REFERENCES


