Predicting the Best Window Size for GDP Evaluation using TSA and ML Techniques

Anmol L Patil
PES University
01FB14ECS035
anmollp@gmail.com

Bhavik V Shah
PES University
01FB14ECS055
bhavikshah406@gmail.com

Chetan G T
PES University
01FB14ECS059
chetant810@gmail.com

Desanur Naveen Reddy
PES University
01FB14ECS066
dnaveen356@gmail.com

ABSTRACT
— Gross Domestic Product (GDP) is one of the prominent ways used to measure the well-being of countrys economy. In this paper, we aim to explore the computation of the optimal window of past values for the forecasting/predicting GDP. This approach will be tested on the dataset containing GDP of India based on Sector Share from the year 1951-2013. This paper will help us know the accurate window size required for the prediction of GDP which will be done by splitting the dataset into training and testing sets to obtain the best possible results.

I. Introduction

GDP is the gauge for the health of the economy of any country. GDP is one of the most prominently used measures of an economys output or production. GDP enables policymakers and central banks to judge whether the economy is contracting or expanding, whether it needs a boost or restraint, and if a threat such as a recession or inflation looms on the horizon. This paper aims to get insights into the optimum window size required for the prediction of GDP of India based on Sector Share. We look to analysis the following issues :
1) Time Series Analysis
2) Sample window size
3) Prediction Window size / Model accuracy or Correctness
Time series data is collected on the same observational unit at multiple time periods. A good example could be total consumption and GDP for a country, Rupee, Pound and Dollar exchange rates. Time series is used for the estimation of dynamic causal effects that can occur in the future. The probable effect over time on consumption of a commodity and its hike/decrease in the tax. Through this paper we come up with a result as to predict the upcoming years GDP and its correctness, using a window of the past history of GDP. We will also point out which sector contributes most towards the GDP of India. The dataset under use ranges from the year 1951 to 2013 and different sectors include Agriculture and allied services, Agriculture, Industry, Mining and quarrying, manufacturing and services. There are mainly two types of approaches to calculate GDP-Expenditure approach and Income approach. Expenditure-based GDP produces both real (inflation-adjusted) and nominal values, while the calculation of income-based GDP is only carried out in nominal values. The expenditure approach is the more common one and is obtained by summing up total consumption, government spending, investment and net exports.

Thus, GDP = C + I + G + (X M), where
C is private consumption or consumer spending;
I is business spending;
G is government spending;
X is exports, and
M is imports.

II. Recent Work

In the recent past, there have been many approaches in this field of prediction of the GDP using various indicators. Different techniques such as time series analysis, neutral networks, generic programming have been used for prediction.

Tao Wang et al.[2] uses existing research methods such as ARIMA time series to predict GDP of particular city named Shenzhen and tries to improve upon them. Forecasting economic growth for the city and gives a scenario analysis with optimistic and pessimistic insights. Uses ARIMA time series analysis to forecast the economic growth. Limitation is that prediction has been done only for a city and usually GDP is taken for the whole country to get insights about the economy of the country. It gives prediction results in both an optimistic scene and as well as pessimistic scene. The main assumption is that no significant changes happen in policies of the city and the effect of economic fluctuations are not taken into consideration. After the prediction, a scenario analysis is done from which inference about various aspects can be made.

Mohd Zukime hj et.al [6] predicts GDP of Malaysia using knowledge based economy indicators focused on IT infrastructures such as computers, research and development. They aim to show that knowledge based indicators play a major role in contributing to the economy of the country. A comparative study between Neutral networks and Econometric Approaches has also been given. In the neutral networks they use sigmoid function for activation and in the econometric approaches they use regression and multi-layer perceptron approach. Neutral networks perform better than the econometric approaches. For future work they suggest to use more indicators such as microeconomic and macroeconomic variables as well as longer time series for better results.

Meifang Li et.al [8] proposes a new genetic programming method that statistically lowers forecast errors of GDP for the year. This model was tested to forecast the GDP time series of China, United States and Japan from 1980 to 2006. Genetic Programming can be seen as an extension of the Genetic Algorithm approach replacing the fixed length bit-strings with a more dynamic representation structure. In this paper, the time series data were divided into a training set, from 1980 to 2000, and a test set, from 2001 to 2006. The test set is used to compare the proposed GP approach with other modified linear ARIMA model. The model was used to forecast the future GDP growth of China, United States and Japan from 2007 to 2020, and, it was surprisingly found that the GDP of Japan fluctuates periodically, however the GDP of China and United States increases stably in the near future. According to the predicted data we can see that the GDP of China will exceed the GDP of Japan for the first time in 2020 or so.

Periklis Gogas et.al [7] depicts the forecasting ability of the yield curve in terms of US Real GDP curve. Data used is variety of short (treasury bills) and long term interest rates(bonds) from year 1976-2011. Yield curve has been one of proven ways to determine inflation and recession in the future. The proposed model uses Support Vector Machines(SVM) for recession forecasting and also a comparative analysis with logit and probit approaches is done. SVM applied on data which is obtained from the yield curve. The SVM uses two mapping functions which are linear and radial basis function (RBF). RBF performs better than the linear mapping. The conclusion is that SVM gives better accuracy in terms of overall recession prediction compared to logit and probit approaches. They claim to have got 100% accuracy in recession accuracy with 6 false alarms.

H Raymond Joseph et.al [4] shows GDP forecasting through Data mining of seaport Export-Import Records based on the theory of structural risk minimization principle to estimate a function by minimizing an upper bound of generalization. Machine learning approaches are applied on the export and import data for prediction. For the purpose of predicting GDP the factors in the dataset used are weighted. They mention approaches such as SVMs, Fuzzy set based learning algorithms and Artificial neural networks for solving this problem. The classifier classifies the vector of observed data into binary sets of GDP growth forecast greater than or equal to 4% and GDP growth forecast less than 4%, then classified data is once again input to the classifier but with the constraints greater than and equal to 2%,lesser than and equal to 4% as one set and greater than and equal to 4% and lesser than and equal to 6% of another set and so on., until the forecast granularity reaches a desired stage. SVM moves the problem of overfitting from optimizing the parameters to kernel model selection, but this model is quite sensitive to overfitting the model selection criterion.

Yang Liang I et.al [1] establishes a time series model of Per Capita GDP and analyzing its law, based on actual conditions of per capita GDP of Yunnan Province from 1978 to 2007 using ARMA model. Identification and order determination of ARMA model is obtained from observations of autocorrelation and partial correlation functions of samples. The model is tested such that If residual sequence is white noise, the specific fitting could be accepted; if not there is useful information which isn't yet extracted in residual sequence. Based on the above time series model of per capita GDP in Yunnan, a relation between the increase of per capita GDP in Yunnan and one in the last first period was observed. Furthermore, it is also correlated with the ones in the current and last second or last third period. If per capita GDP in last first period increases by 1%, the one in current period will increase by 0.19%.

There are other papers which use the ANNs and ARIMA models for prediction of the GDP. ANNs were used to forecast GDP for Turkey [3]. Similarly, GDP of Kenya was modelled and forecasted using Autoregressive Integrated Moving Average (ARIMA) Models [5].

III. Problem Statement and Approach

There have been many approaches for prediction of GDP for cities and some countries such as China, Kenya and Malaysia using different indicators such as knowledge based, import export data etc. as mentioned in the above section. One of the traditional approaches has been through the yield curve analysis. We aim to choose the optimal window size required for the prediction of GDP for India as these techniques havent been tried for the economy of India. We look to see if we could get some inferences by exploring this approach. Sector Share as an indicator for the prediction has been a novel approach and hasn't been explore before. We aim to choose the best window size using time series analysis and do predictive modeling using ML and forecasting techniques. The dataset has been taken from the Open Government Data (OGD) Platform and it has been released by the Govt of India. The dataset contains 62 years of GDP data of India ranging from the year 1951 to 2013. Total of 7 attributes are used for this study. Each attribute is a Sector Share contribution. The sectors are Agriculture and Allied Services, Agriculture, Industry, Mining and Quarrying, Manufacturing and Services. Also, it has been suggested that having a larger time series data would get better results [6] and thus the dataset used has 62 years of sector share GDP data. Assumptions made while predicting the GDP through this analysis, is that no major change in economic policies by the government is taken and also other factors which affect the GDP are not taken into account.

IV. Proposed System

A. *Summarize Data*

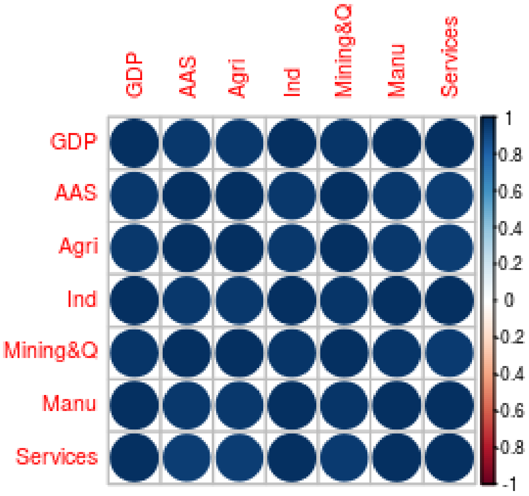

FIGURE 1 : Correlation plot

*1) **Descriptive statistics and Data Visualization** :* Descriptive statistics and data visualization was done to get insights into the data and approach the problem better. Density and histogram plots were plotted to find the nature of the data, this should that all attributes were right skewed and very similar plots. Boxplot was also plotted to check if there were any outliers in the dataset, it was found that GDP, Industry and Manufacturing contained 4 outliers. The Services section contained about

6 outliers. Next we found the correlation of the attributes and this was plotted. It was seen that the attributes were highly correlated with values very close to 1. This was one of the major insights gained from this analysis. We also found that the dataset contained 1 missing value for year 2012-2013.

B. *Preparing Data*

*1) **Data Cleaning and Feature Selection :*** Data was preprocessed to make it suitable for the study and the missing values were added with the technique which gave best result. The original dataset contained 20 columns, but for this study we have only used 7 attributes which were relevant and contained the Sector Share details. To add the missing values, we tried the mean method, linear interpolation and linear regression. Linear regression turned out to give the best results. Mean method did not follow the trend and wasnt suitable. Since the missing value was the last one, linear interpolation wasnt suitable since the value above was known but the value below wasnt present. As the attributes were highly correlated, linear regression gave the best results to add the missing value and it followed the trend.

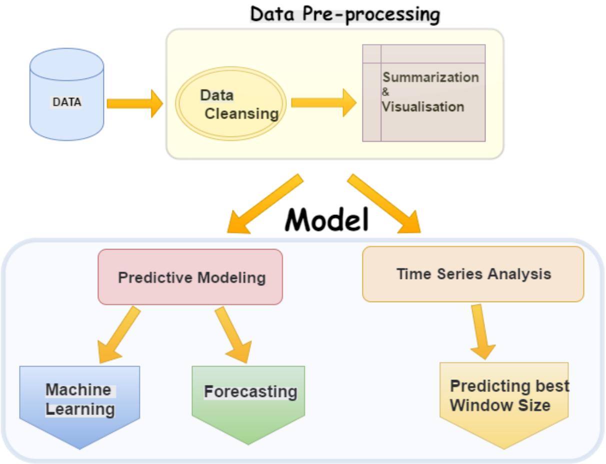

FIGURE 2 : Flowchart

C. *The Model*

For predictive modeling we have used the following techniques :
ML : Regression,SVM, kNN
Time series forecasting methods : Mean, Naive, Drift and Seasonal naive methods.
For selecting the optimal window size we are using the following techniques :
1) ARIMA 2) HOLT WINTERS
The algorithms were evaluated using RMSE and R-squared metrics. RMSE will give a gross idea of how wrong all predictions are (0 is perfect) and R-squared will give an idea of how well the model has fit the data.(1 is perfect, 0 is worst).

V. TECHNIQUES USED

A. *Machine Learning*

The dataset is split in train and test datasets. 20% of the data is set for validation. 10-fold cross validation technique is used for training the algorithms. The ML approaches tested for this dataset are Linear Regression (LR) , Support Vector Machines (SVM), k-Nearest Neighbors (kNN). LR: For linear regression, RMSE value of 1207.367 and Rsquared value of 0.9999985 was obtained. SVM: Kernel used was the Radial Basis Function. The final values used for the model were sigma = 5.701345 and C = 1. RMSE value of 512471.7 and Rsquared value of 0.9630921 was obtained. kNN : The optimal k value which was selected for this model was k = 5. RMSE value of 165545.8 and Rsquared value of 0.9949018 was obtained.

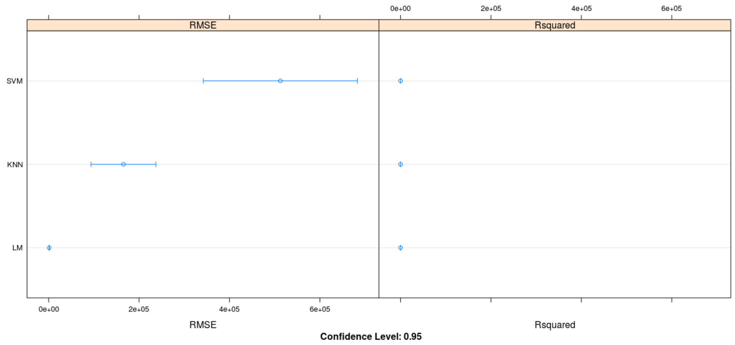

FIGURE 3 : Box plot of knn,svm and lm

These preliminary training indicates that all algorithms have fit the data well which is indicated by a high R-squared value. But Linear Regression performs the best with least RMSE value. The reason why LR performed the best is because the attributes are highly correlated and in linear fashion which can be seen in Scatterplot and Correlation plot. [Data Viz Section] Comparison of the algorithms is depicted in the figure.

B. *Forecasting*

For forecasting, we used the following methods : Mean method, Naive method , Drift method, Seasonal nave method. The data was split into test and train, 20% of the data was used for testing. The metric used to determine the efficacy of the model was Mean Absolute Percentage Error (MAPE). This metric is considered as one of the best ways to determine how good the forecasting is. We also used 2 tests to check if ARIMA model was suitable which will be explained in the next section.

C. *Time Series Analysis*

The dataset is converted into vectors and then window size is selected. Initially we are selecting the window size as 5(last Five years GDP) and predicting the next year GDP .The predicted value is compared with the available original GDP since it shows greater value of RMSE, the selected window size is not enough to predict the accurate value of future GDP.Then we doubled the window size and until the predicted GDP showed a considerable RMSE , which was closer to the original GDP from the dataset. Further a window size of 60 gave us an approximately precise value of the future GDP comparable to the original. Hence, we conclude that larger the window size the more is the precision of predicting correctly. Two methods used are ARIMA and HOLT WINTERS in which Auto ARIMA method has a higher accuracy than Holt Winters(RMSE of Auto ARIMA is less than Holt Winters). Moreover the values predicted using ARIMA fit well under the lm() plot curve.

VI. EXPERIMENTS AND RESULTS WITH A DISCUSSION ON THE RESULTS

A. *Machine Learning*

From the previous section, it was concluded that Regression for the best model to use for this study. Thus, LR was used for final validation of the model as it performed best. Multiple Regression was used to validate the model with GDP being the y variable and the 6 attributes being x1,x2,x3 and so on. Equation :
formula = GDP AAS + Agri + Ind +‘Mining&Q‘ + Manu + Services

The final RMSE value obtained after validation was 1701.28. For training dataset, the RMSE value obtained was 1207.367. This signifies that the model has performed well on the test data.

B. *Forecasting*

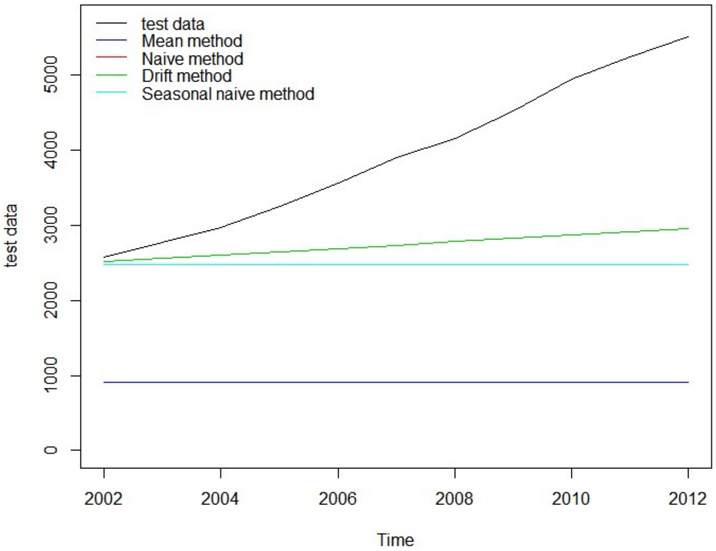

FIGURE 4 : Comparison of various methods

It is rather obvious from the predicted graph that none of these methods produce a good forecast of the series.The mean method and the naive method do not detect the trend or the seasonality in the series. The drift method does detect the trend but not the seasonality, while the seasonal nave method does the reverse. The best method, on the basis of the Mean absolute percentage error (MAPE) is the drift method, which in our suggests that the trend is more important than the seasonality in this series. Next we carried out 2 tests to determine whether ARIMA model is suitable for this study. These tests were Box-Ljung test and Box Pierce test. These tests showed that ARIMA is most suitable for the problem of forecasting.The ARIMA model had the lowest MAPE of all forecasting methods, and it is obvious from the plot that the prediction based on the ARIMA model detects both the trend and the seasonality of the series. The residuals of the model are reasonably good, and the LjungBox test shows that there is no serial correlation.The superiority of ARIMA over the other models of forecasting depends, in part, on ARIMA including a term for differencing.

C. Time Series Method

As we discussed above Auto ARIMA model gives the best prediction compared to Holt Winters method. When we considered the window size of 5 RMSE value of Auto ARIMA model is 11242.11 and for Holt Winters 12568.89 which clearly shows that Auto ARIMA gives the best prediction.

Original GDP	Predicted GDP(Arima)	Year
4937006	333766	2010
5243582	333766	2011
5503476	33376	2012

Table 1 : Window size 5

From the above fit we can infer that the ARIMA model, gives merely same value of 333766, which indicates that this model doesnt hold good for smaller window sizes such as 5.

Now considering a bigger window size, such as 60 , the results found are:

Original GDP	Predicted GDP(Arima)	Year
5243582	5334007	2011
5503476	5740024	2012

Table 2 : Window size 60

From this table, considering 60 as the window size we get a decent prediction of the GDP.

VII. Conclusions

This study shows that for prediction of GDP, predicting modeling techniques such as ML and Forecasting can be used. In case of ML, Linear Regression is best choice when the data is highly correlated. Most forecasting methods fail to perform well as most do not follow the trend and seasonality. But drift method could be used for forecasting if the trend of the data is needed and seasonal method can be used to get the seasonality. ARIMA model is the best choice for forecasting and thus model was used to select the window size. For choosing the best window size, Auto ARIMA clearly is the winner over Holt Winters with excellent predictions. A larger window size produces better results as the trend and seasonality is taken into account. For our data, a window size of 60 performed the best while small window sizes such as 5 performed the worst. Thus the best window size for predicting GDP is a window size which is large enough to take into account the trend and the seasonality of the dataset.

References

[1] Yang Liang and Bin Han. The establishment and analysis of time series model of per capita gdp in yunnan, china. In *Industrial Engineering and Engineering Management (IE&EM), 2010 IEEE 17Th International Conference on*, pages 137–139. IEEE, 2010.

[2] Tao Wang. Forecast of economic growth by time series and scenario planning methoda case study of shenzhen. *Modern Economy*, 7(02):212, 2016.

[3] Meltem Karaatli. Using artificial neural networks to forecast gdp for turkey. 2012.

[4] H Raymond Joseph. Gdp forecasting through data mining of seaport export-import records. In *Proceedings of the International Conference on Data Mining (DMIN)*, page 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2013.

[5] Musundi Sammy Wabomba and Mungai Fredrick Mmukiira Peter Mutwiri. Modeling and forecasting kenyan gdp using autoregressive integrated moving average (arima) models.

[6] Mohd Zukime Hj Mat Junoh. Predicting gdp growth in malaysia using knowledge-based economy indicators: a comparison between neural network and econometric approaches. *Sunway Academic Journal*, 1:39–50, 2004.

[7] Periklis Gogas, Theophilos Papadimitriou, Maria Matthaiou, and Efthymia Chrysanthidou. Yield curve and recession forecasting in a machine learning framework. *Computational Economics*, 45(4):635–645, 2015.

[8] Meifang Li, Guoxin Liu, and Yongxiang Zhao. Forecasting gdp growth using genetic programming. In *Third International Conference on Natural Computation (ICNC 2007)*, volume 4, pages 393–397. IEEE, 2007.

VIII. Contributions

- Missing values,Feature selection and Prediction of window size is done by Anmol and Chetan.
- Time series analysis,Forecasting and Report making is done by Bhavik
- Predictive modeling and Machine learning is done by Naveen
- Summarization and Visualization is done by Bhavik and Naveen

IX. Related Images

1)Data Visualizations
time series
Machine learning
2)Tables
Arima and HoltWinters