International Journal of Knowledge Based Computer Systems
7 (1), June 2019, 10-18
http://www.publishingindia.com/ijkbcs/

Crude Oil Consumption Forecasting Using Classical and Machine Learning Methods

Zameer Fatima[1*], Alok Kumar[2], Lakshita Bhargava[3] and Ayushi Saxena[4]
[1]Assistant Professor, Department of Computer Science & Engineering. Maharaja Agrasen Institute of Technology, GGSIPU, New Delhi, India. Email: zameerfatima@mait.ac.in
[2]Student, Department of Computer Science & Engineering. Maharaja Agrasen Institute of Technology, GGSIPU, New Delhi, India. Email: alokrkmv12@gmail.com
[3]Student, Department of Computer Science & Engineering. Maharaja Agrasen Institute of Technology, GGSIPU, New Delhi, India. Email: lakshitabhargava96@gmail.com
[4]Student, Department of Computer Science & Engineering. Maharaja Agrasen Institute of Technology, GGSIPU, New Delhi, India. Email: ayushi.saxena2305@gmail.com
*Corresponding Author

Abstract: The global oil market is the most important of all the world energy markets. Since crude oil is a non-renewable source, its quantity is fixed and limited. To manage the available oil reserves, it will be helpful if we have estimation about the future consumption requirements of this resource beforehand. This paper describes methods to forecast crude oil consumption of next 5 years using past 17 years data (2000-2017). The decision making process comprised of: (1) Preprocessing of dataset, (2) Designing forecasting model, (3) Training model, (4) Testing model on test set, (5) Forecasting results for next 5 years. The proposed methods are divided into two categories: (a) Classical methods, (b) Machine Learning methods. These were applied on global data as well as on three major countries: (a) the USA, (b) China, (c) India. The results showed that the best accuracy was obtained for polynomial regression. An accuracy of 97.8% was obtained.

Keywords: Accuracy, Machine learning, Oil consumption prediction, Time series forecasting.

I. Introduction

Oil is the largest source of primary commercial energy in world. Global share of oil consumption in year 2015 was 32.94% [1]. Fossil fuels and in general, energy carriers play a key in various economic sectors and sub-sectors. Distribution of oil demand worldwide in 2016 in different sectors was road (50.21%), petrochemicals (14.26%), residential/agriculture use (8.94%), aviation (7.45%), electricity generation (2.55%) and other industries (16.59%) is shown in Fig. 1 [2]. The countries that we used for our research are the primary consumers of oil. They account for 1696 Mt of total oil consumption in the world [3]. Therefore, fossil fuels are of great importance for consumers, decision-makers and planners due to their value of income, consumption etc. Out of all the fossil fuels, crude oil is the most important fuel due to its dominant role as an energy source. In the present scenario where consumption of fossil fuels is increasing at a very rapid rate due to advancement of technology and increase in global human population in the 20th Century, it has become a task of utmost importance for every country to have a fairly accurate estimation of crude oil consumption in upcoming years. According to June 2018 BP statistical review of world energy, the oil consumption has constantly increased in past 31 years out of 34. Global oil consumption growth averaged 1.8%, or 1.7 million barrels per day (b/d) in 2017, above its 10 years average of 1.2% for the third consecutive year. China and USA were the single largest contributors to growth. US have 4.28% of total world population but uses 25% of world's oil. It has been estimated that if oil consumption continues at this rate, then the existing oil reserves will not last more than 50.2 years [4]. Considering the recent trends, there is an urgent need to devise new strategies to manage the resources. This can be done judicious if we have a fairly good prediction of consumption patterns of the near future.

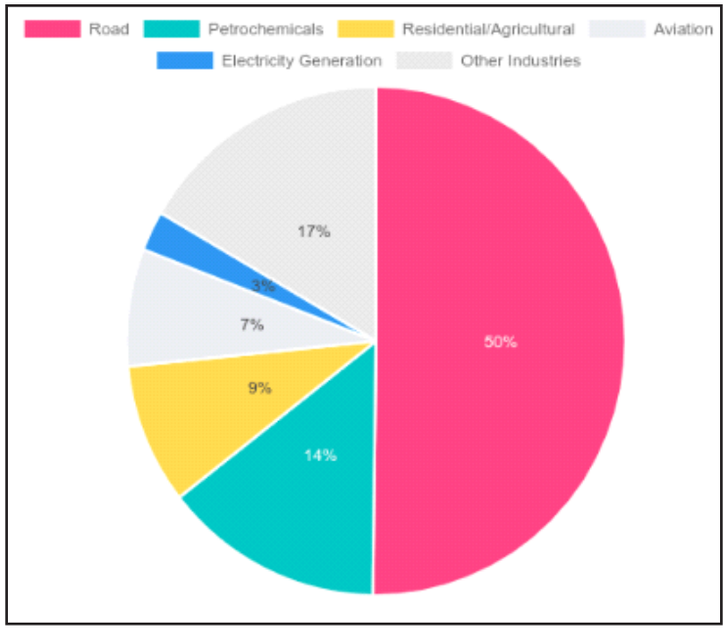

Fig. 1: Global Crude Oil Consumption in Various Sectors

II. Literature Review

In this section we provide an overview of methods used in previous studies for forecasting. In [5], forecasting of fossil fuel production in Turkey has been done using comparative cubic regression and ARIMA model. In their work they have used various forms of ARIMA model like CARISMA and SARIMA to achieve a fairly good prediction result. But as this study was only done for Turkey and we didn't find any such study for global consumption we were inspired to design our own model for predicting global crude oil consumption. They have used different methods for different fossil fuel types. Using the cubic regression model they obtained highest goodness - of - fit value of 0.959 for lignite compound. For natural gas, they used ARIMA model and found that it gave better results as compared to regression analysis. They found that the most successful one was the forecasting of oil production and for this they used the SARIMA model.

Another retrospective was made by Landsberg who re-assessed the demand and supply of US resources from 1960-2000 and published it under resources for the future [6]. After evaluating his work for decades he got interesting insights. One of them was that over a very long time span some unforeseeable changes occur which cannot be predicted with good accuracy for example oil shocks or environmental changes or opposition to use of some new renewable source of energy. So, making predictions for even next 20 years in future is far too long for maintaining the validity of the various assumptions made. Therefore long-range forecasts should be avoided. The same was concluded by Smil, who focused and catalogued failures of long-range energy forecasts [7, 8]. Keeping this in mind, we decided to forecast the oil consumption for next 5 years.

III. Methods and Data

This paper makes use of energy statistical yearbook 2018 dataset for crude oil input to refineries. We compile here results obtained from two different techniques: Classical methods for time series analysis and machine learning techniques. The classical methods used are simple exponential smoothing and ARIMA method. The machine learning methods used are linear regression, polynomial regression and support vector regression. The first step of these methods involves preprocessing data which includes checking for missing values, converting dataset into matrix of independent variables and vector of dependent variables and selecting the train and test set for our models. For dealing with any missing value in our dataset we used the standard method of finding mean of the column and replacing all the missing values in that column by mean value. The second step is to design the forecasting models. The chosen dataset is univariate time series data hence choice of models narrowed down to five models which support univariate data because of the nature of their algorithms. Among the five models that we choose to implement on our dataset two of them are simple exponential smoothing, ARIMA model which are classical model for time series forecasting. We also choose three machine learning models which include simple linear regression, polynomial regression, and support vector regression. The third step is to train the model.

The dataset was split into train and test set according to the ratio 70/30 to prevent overfitting. The final step is to calculate the performance metric for each model which is accuracy here for finding the accuracy of the model we tested our model on test dataset after training them on training dataset and compared the predicted result with the actual value. In this section, we discuss these models.

A. Simple Exponential Smoothing

Exponential smoothing is a rule of thumb method used for smoothing time series data using exponential functions to assign exponentially decreasing weights over time. This technique finds wide application in the time series analysis specifically. This method is perfectly apt for univariate data which lacks trend and seasonality. The trend represents a systematic linear or even a non-linear component that changes over time but does not repeat within the time range captured by our data. Seasonality occurs when the data is affected by seasonal factors. Seasonality is for a fixed or known period of time. It is required to remove trend and seasonality so as to dampen out and remove any random increases or decreases in the data. The simplest form of exponential smoothing is given by the equation:

$$F_t = F_{t-1} + \alpha(A_{t-1} - F_{t-1})$$

Where F_t = new forecast

F_{t-1} = previous period forecast

A_{t-1} = previous period actual value

α = smoothing or weighting constant ($0 < \alpha < 1$)

The terms in this equation are form of geometric progression which is, in fact, an exponential equation. The values predicted using exponential smoothing method are weighted averages of the past values, with the weights decreasing exponentially as the data gets older. The more recent the observation, the more is the associated weight. For forecasting of crude oil consumption, it was logical to give more weight to the recent consumption values. The value of α, if closer to one, has less smoothing effect and gives higher weight to recent data. On the other hand if the value of α is closer to zero, then it has higher smoothing effect. To find the most suitable value of exponential smoothing factor we trained our model for different values of α. This simple form of exponential smoothing can also be said as ARIMA (0, 1, 1) with no constant term. The initial value of F_t plays an important role in computing further forecasts. Since all the values in dataset are considered for this method, it can be set equal to y_1 (used here) or equal to average of first few observations. After testing our model we got best result when the value of exponential smoothing factor was 0.3. So the value of α which we have chosen for our model is 0.3 shown in Table I.

TABLE I: ACCURACY COMPARISON FOR DIFFERENT VALUES OF SIMPLE EXPONENTIAL SMOOTHING

Year	Predicted Value for Different				Actual Value
	0.1		0.3		
2000	3502	3506	3408	3457	3510
2001	3518	3515	3518	3508	3534
2005	3513	3519	3590	3515	3813
2008	3552	3592	3633	3547	3900
Accuracy (%)	95.45	95.68	95.88	95.06	

A comparison of predicted result and actual result for our global test set data is shown below in both Fig. 2 and Table II.

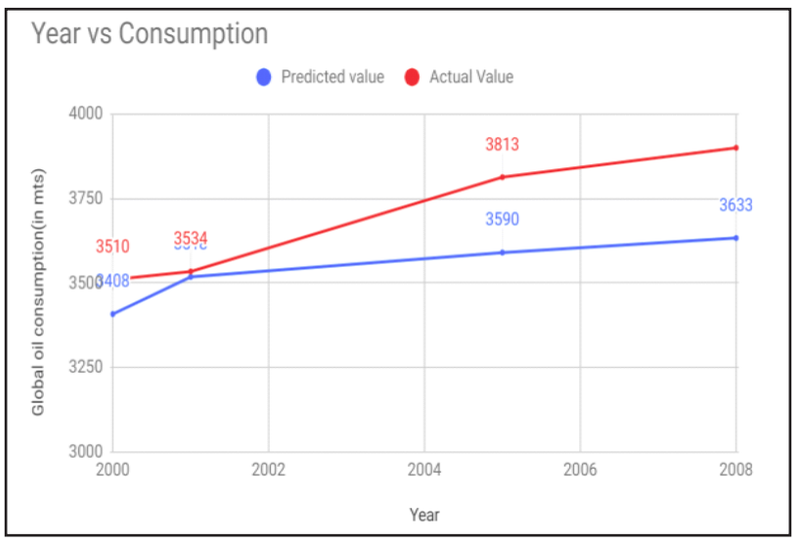

Fig. 2: Actual and Predicted Values of Test Set for Simple Exponential Smoothing

Table II: Actual and Predicted Values of Test Set for Simple Exponential Smoothing

S.No.	Year	Actual Value	Predicted Value
1	2000	3510	3408
2	2001	3534	3518
3	2005	3813	3590
4	2008	3900	3633

B. NonSeasonal ARIMA Model

As opposed to simple exponential smoothing which is based on trends and seasonality of data, ARIMA model aims to study autocorrelations. On combining differencing to make sure stationery in time series with autoregression and moving average we obtain nonseasonal ARIMA model. The model can be written as:

$$y'_t = c + \Phi_1 y'_{t-1} + \cdots + \Phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Where y'_t = differenced time series

y'_{t-p} = lagged time series by p steps

ε_t is white noise

This model is also written as:

ARIMA (p, d, q) where

p = order of autoregressive part

d = order of differencing

q = order of moving average part

Steps involved in this process are:

- Stationeries time series using differencing.
- Studying patterns of autocorrelations and partial correlations to determine if lags of stationeries series and/or lags of forecast errors are to be included.
- Fit the model and check their ACF and PACF plots to see if the entire pattern has been explained and all coefficients are significant.

We plotted the autocorrelation plot for global dataset which showed us positive correlation. Therefore, a good starting point can be an AR parameter set to 7.

For our data, we used ARIMA (7, 1, 0), which gave us a model with stationary time series of difference 1 and positive correlation with 7 lags in Fig. 3.

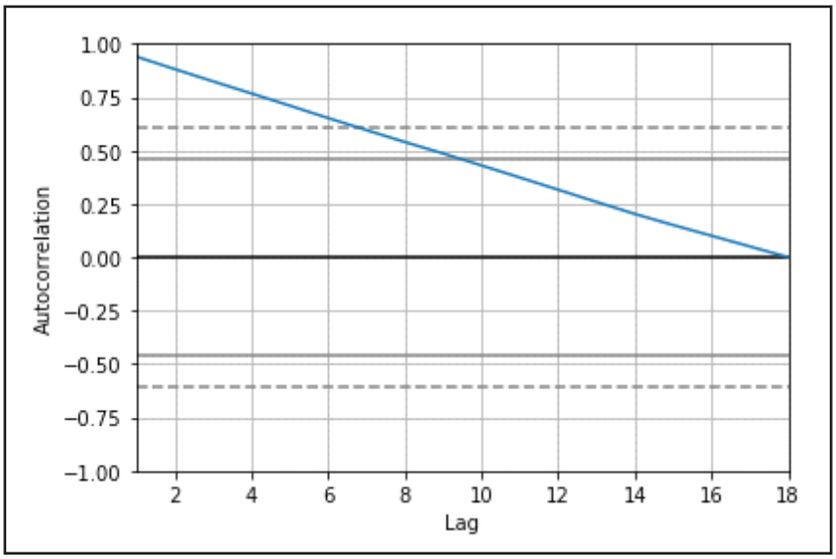

Fig. 3: ACF Plots for Global and Country Wise Dataset Using ARIMA (7, 1, 0)

C. Linear Regression

In a dataset consisting of a single independent variable linear regression can be method of choice. This method is widely used for predicting values of the dependent variable which is related only to a single independent variable. In linear regression we try to fit our dataset on a line. The equation used for fitting the model is:

y = B0 + B1*x

Where y = dependent variable

x = independent variable

B0, B1 are coefficients of the regression

In this model we need to estimate coefficients B0 and B1 which can give us the best fit line.

To estimate the best fit line we use least square method. This method involves the following steps:

- First, we will calculate the mean of x and y values:

$$X = \frac{\sum_{i=1}^{n} xi}{n}$$

$$Y = \frac{\sum_{i=1}^{n} yi}{n}$$

- This step involves finding the slope of the best fit line.

$$m = \frac{\sum_{i=1}^{n} (xi - X)(yi - Y)}{\sum_{i=1}^{n} (xi - X)\text{^}2}$$

- Once we have found the slope of line we can find the y-intercept of the line using the formula:

  b = *Y*-m.x

Once we have found slope and intercept of the line we can form the equation of line using slope, intercept form of a line.

After forming the equation of line our next job was to find best fitting line for our dataset. For finding best-fitting line for our dataset we used the least-squares method. The least square method is a proven method for finding best-fitting regression line in a dataset. In this method we first take the mean of dependent variables and assume that our best-fitting line lies along the mean. After assuming mean as our best-fitting line we find distance of data point from that line. These distances are known as residuals. Residuals above the mean and below the mean should add up to zero. Now we need to find sum of squared errors (SSE). There are two reasons for squaring the errors before finding their sum:

- It makes residuals (errors) positive.
- It emphasizes larger deviations.

Now to minimize SSE, we move our line in the plain and SSE values for different slope (orientation) of the line. Our best fitting line corresponds to the slope where SSE value is minimum.

Real-World Applications of Simple Linear Regression

Simple linear regression is a highly useful and fairly simple model which is often used to forecast values. Some major fields where simple linear regression can find its use are:

Predicting future sales of a product based on past data.

- Used by economist to predict economic growth of country and states.
- Used by management team within an organization to predict the salary of the employees.
- Oil price and consumption prediction.
- Estimating the price and number of house a builder can sell the upcoming year in a particular locality.

In our project we had a time series data of crude oil consumption in last two decades. We used simple linear regression to forecast consumption of crude oil in next five. Although it turned out that this model is not an exact fit for our data set we achieved a fairly good result using this model. We achieved an accuracy of 97.77% using this model. A comparison of predicted result and actual result for our global test set data is shown below in both Fig. 4 and Table III.

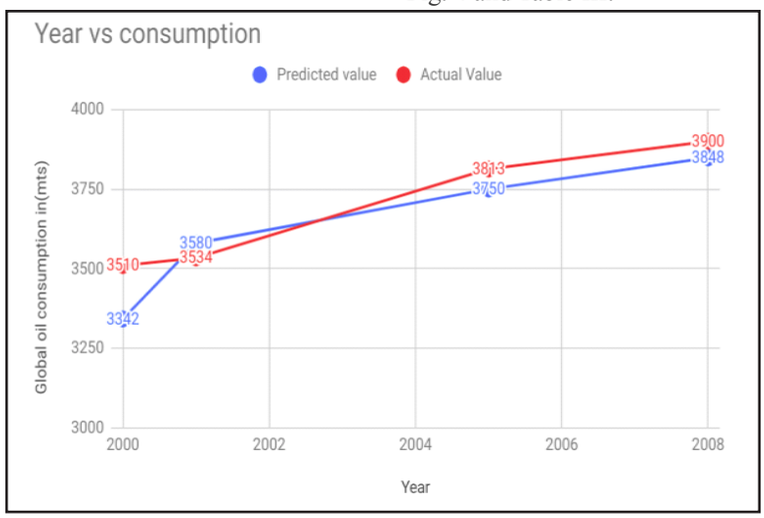

Fig. 4: Actual and Predicted Values of Test Set for Linear Regression

TABLE III: ACTUAL AND PREDICTED VALUES OF TEST SET FOR LINEAR REGRESSION

S. No.	Year	Actual Value	Predicted Value
1	2000	3510	3342
2	2001	3534	3580
3	2005	3813	3750
4	2008	3900	3848

D. Polynomial Regression

After implementing Linear regression we analyzed that the dataset that we have is not much linear rather it is scattered or diversified. In such a case, polynomial regression would be a better choice for obtaining results. In this method, the independent variable y is modelled as an nth degree polynomial in x. The historical data can be represented using an n^{th} degree polynomial. This mathematical technique is applied to

determine the trend influence and to make a demand forecast. The generalized formula for polynomial regression is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots\ldots \beta_n x^n + \varepsilon$$

Just like linear regression model for finding best fit polynomial regression model for our dataset, we have used the method of least squares.

Why We Used Polynomial Regression for Forecasting

When we tried to fit linear regression model on our dataset we found that the model was not an exact fit. So to achieve better accuracy we implemented a polynomial regression model.

Using this method, we obtained a fairly good accuracy.

We first implemented the polynomial regression method using degree = 2. The accuracy obtained was quite low. Then we implemented it using degree = 3. The accuracy obtained was much better. We achieved an accuracy of 97.89% using this model for global data. A comparison of predicted result and actual result for our global test set data is shown below in both Fig. 5 and Table IV.

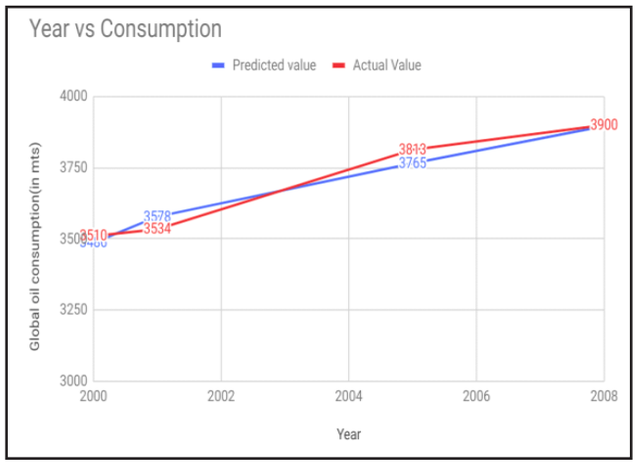

Fig. 5: Actual and Predicted Values of Test Set for Polynomial Regression

TABLE IV: ACTUAL AND PREDICTED VALUES OF TEST SET FOR POLYNOMIAL REGRESSION

S. No.	Year	Actual Value	Predicted Value
1	2000	3510	3328
2	2001	3534	3462
3	2005	3813	3874
4	2008	3900	3781

E. Support Vector Regression

Support vector regression is similar to SVM algorithm which both works well on linear as well as nonlinear datasets. Support vector regression makes use of two boundary lines which are at a distance ε from the hyperplane. We try to decide boundaries at ε distance from hyperplane such that data points closest to hyperplane or support vectors are within those boundary lines. After fitting a hyperplane through the support vectors, we use regression to forecast future values. The main difference between linear regression and support vector regression technique is that linear regression tends to minimize the error between the line of regression and data points whereas in support vector regression the main aim is to get maximum data points within a particular threshold value as specified by the boundaries of the hyperplane. SVR uses kernel functions to transform data into higher dimensions so that it becomes possible to apply linear separation. So, here we used an RBF kernel.

Why We Used SVR Model and How It Differs from Other Regression Models?

In simple regression models our main goal is to minimize the error rate which is generally achieved by using the method of least squares. But in SVR we try to fit the error within certain threshold. For doing this, we assume two boundary lines and a hyperplane, and we consider only those points which lie within the boundary line. Once we have done this we try to find the best fit line for the model. The best fit line is the line hyperplane that has maximum number of points. The main reason we had for choosing SVR was that after fitting our dataset on linear regression and polynomial regression model we found that regression models are working quite well for our dataset. We got fairly good accuracy in both the cases. So in the hope of achieving better accuracy rate we thought of implementing the SVR model.

How We Implemented SVR on Our Dataset

We implemented SVR on our dataset in the following steps:

- For simplicity we assumed that our hyperplane is passing through Y-axis. So on the basis of this assumption we got equation of hyperplane as:

  wx+b=0
- Next we had to find the equation of boundary lines for this we assumed boundary lines to be **e** distance apart from the hyperplane. Thus we got an equation of two boundary lines as:

  wx+b = e & wx+b = -e respectively.
- Once we had equation of boundary lines and equation of hyperplane we moved our hyperplane for different value of w to find the best-fit line.

Using this model, we achieved an accuracy of 96.82%.

A comparison of predicted result and actual result for our global test set data is shown below in both Fig. 6 and Table V.

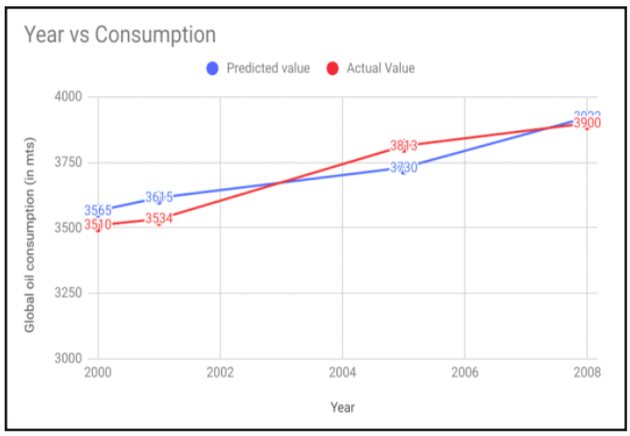

Fig. 6: Actual and Predicted Values of Test Set for Simple Support Vector Regression

TABLE V: ACTUAL AND PREDICTED VALUES OF TEST SET FOR SUPPORT VECTOR REGRESSION MODEL

S. No.	Year	Actual Value	Predicted Value
1	2000	3510	3565
2	2001	3534	3615
3	2005	3813	3730
4	2008	3900	3922

IV. MEASURE OF PERFORMANCE AND RESULTS

We computed accuracy as a measure of performance. After training our models on training set we tested each and every module on randomly generated test set. Test results were compared with actual result to estimate the accuracy of models. With an accuracy of 97.89% polynomial regression seemed a best-fitting model for global data set. Hence we used polynomial regression for forecasting crude oil consumption result for next five years shown in Fig. 7 and Fig. 8.

With an accuracy of 99.97% ARIMA (7, 1, 0) model gave best result for India. So we did ARIMA model for forecasting next five-year crude oil consumption in India. With an accuracy of 99.983% and 99.989% for China and USA we found that polynomial regression worked best in these two countries. So for these two countries we used polynomial regression for crude oil consumption forecasting.

A. Comparison of Accuracy of Various Models

Tables VI to X list the accuracies we received from all the methods, country-wise.

i. For Global Dataset

TABLE VI: ACCURACY COMPARISON OF VARIOUS MODELS FOR GLOBAL DATA SET

S. No.	Model	Accuracy (%)
1	Exponential smoothing	95.68
2	ARIMA	97.21
3	Simple Linear Regression	97.77
4	Polynomial Regression	97.88
5	Support Vector Regression	96.28

ii. For India

TABLE VII: ACCURACY COMPARISON OF VARIOUS MODELS FOR INDIA

S. No.	Model	Accuracy (%)
1	Exponential smoothing	96.48
2	ARIMA	99.97
3	Simple Linear Regression	99.85
4	Polynomial Regression	99.87
5	Support Vector Regression	98.74

iii. For China

TABLE VIII: ACCURACY COMPARISON OF VARIOUS MODELS FOR CHINA

S. No.	Model	Accuracy (%)
1	Exponential smoothing	96.92
2	ARIMA	99.96
3	Simple Linear Regression	99.95
4	Polynomial Regression	99.98
5	Support Vector Regression	99.92

iv. For the USA

TABLE IX: ACCURACY COMPARISON OF VARIOUS MODELS FOR THE USA

S. No.	Model	Accuracy (%)
1	Exponential smoothing	95.492
2	ARIMA	99.974
3	Simple Linear Regression	99.986
4	Polynomial Regression	99.989
5	Support Vector Regression	99.983

B. Results for Global Dataset

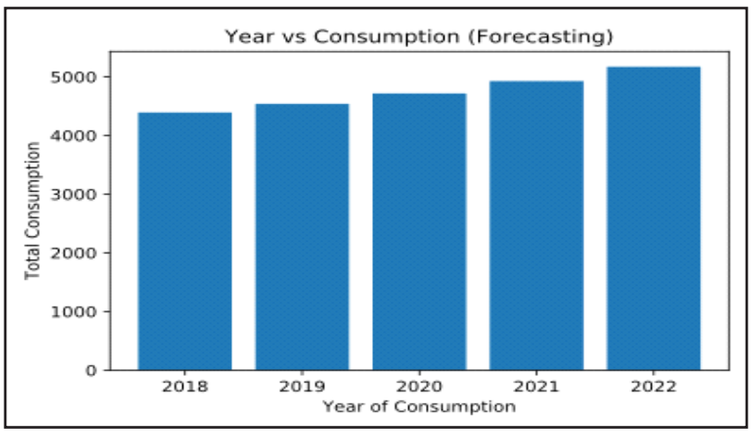

Fig. 7: Estimated Global Crude Oil Consumption in Next Five Years Using Polynomial Regression with Degree 3

C. Results for Country Wise Dataset

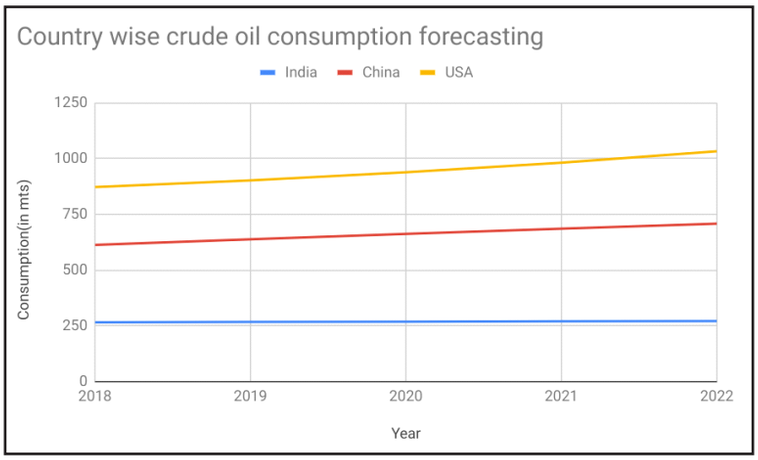

Fig. 8: Estimated Country Wise Crude Oil Consumption in Next Five Years

TABLE X: ESTIMATED CRUDE OIL CONSUMPTION IN NEXT FIVE YEARS

S. No.	Year	Crude Oil Consumption Estimation (mts)			
		Global	India	China	USA
1	2018	4394.19	266	613.06	872.66
2	2019	4541.86	267	638.10	902.38
3	2020	4719.05	268	662.45	938.67
4	2021	4928.77	269	685.95	982.09
5	2022	5174.08	270	708.46	1033.22

V. CONCLUSION

The aim of this paper was to propose the best model which can be used to make predictions of crude oil consumption in next 5 years globally as well as country-wise, as compared to individual forecasts for specific countries. After implementing various classical methods as well as machine learning methods results with varying accuracy were obtained. Out of the five methods which we implemented: Simple exponential smoothing, ARIMA model, linear regression, polynomial regression, and support vector regression; we found that maximum accuracy was obtained using polynomial regression (97.89%) for global

dataset whereas polynomial regression (99.99%), polynomial regression (99.98%) and ARIMA (99.97%) for USA, China and India respectively. With these predicted values the governments and environmentalists can plan the future course of action as to how the resources should be managed sustainably. These values can act as the benchmarks for the assessments that will be done in the future for further planning and decision making for proper and judicious use of this valuable resource.

VI. Future Work

Future research can be carried out by considering various factors contributing to oil consumption in each country like the worldwide gradual shift to electrical energy in major sectors like transportation and Industries. This can be done using neural networks, deep learning models such as LSTMs and combined models as used in [9].

References

1. "World energy resources," www.worldenergy.org, 2016. [Online]. Available: https://www.worldenergy.org/wp-content/uploads/2016/10/World-Energy-Resources-Full-report-2016.10.03.pdf
2. "World oil outlook 2040," www.opec.org, 2017. [Online]. Available: https://www.opec.org/opec_web/flipbook/WOO2017/WOO2017/assets/common/downloads/WOO%202017.pdf
3. "BP statistical review of world energy," www.bp.com, 2018. Available: https://www.bp.com/content/dam/bp/en/corporate/pdf/energy-economics/statistical-review/bp-stats-review-2018-full-report.pdf
4. "Global energy statistical yearbook," Enerdata, 2018. Available: https://yearbook.enerdata.net/crude-oil/world-refineries-data.html
5. V. S. Ediger, S. Akar, and B. Uğurlu, "Forecasting production of fossil fuel sources in Turkey using a comparative regression and ARIMA model," *Energy Policy,* vol. 34, no. 18, pp. 3836-3846, 2006.
6. H. H. Landsberg, L. L. Fischman, and J. L. Fisher, "Resources in America's future: Patterns of requirements and availabilities 1960-2000," Johns Hopkins Press for Resources for the Future, Baltimore, 1963.
7. V. Smil, "Perils of long-range energy forecasting: Reflections on looking far ahead," *Technological Forecasting and Social Change*, vol. 65, no. 3, pp. 251-264, 2000.
8. V. Smil, *Energy at the Crossroads: Global Perspectives and Uncertainties*, The MIT Press, 2003.
9. J. Li, R. Wang, J. Wang, and Y. Li, "Analysis and forecasting of the oil consumption in China based on combination models optimized by artificial intelligence algorithms," *Energy*, vol. 144, pp. 243-264, 2018.