Volume 9 Issue 6, 2023
DOI: 10.6919/ICJE.202306_9(6).0029

Study of CO2 Concentration Prediction based on the Multi-class Regression Model

Xiyue Yuan and Haochong Luo

Nanjing No.1 Middle School, Nanjing, Jiangsu 210000, China

Abstract

In this paper, three models, including nonlinear regression, ARIMA, and SVM, are used to investigate CO2 concentration. The data are fitted with a quadratic polynomial and an exponential function by observing the change curve of CO2 concentration. Subsequently, two regression prediction models are developed, and the coefficients in the models are solved by the least squares method. After that, the goodness-of-fit of the regression models is also calculated. In addition, we built prediction models for ARIMA and SVM. By selecting some data as the training set, we compared the accuracy of the quadratic polynomial, ARIMA, and SVM models in predicting known data, and we found that the quadratic polynomial model is the most accurate, which predicted that the CO2 concentration would reach 685 ppm in 2100.

Keywords

Nonlinear Regression; ARIMA Prediction Model; SVM Prediction Model; CO2 Concentration.

1. Introduction

As human industrialization has intensified over the past century, atmospheric CO2 emissions have been increasing. Figure 1 shows the curve of CO2 concentration levels from 1959 to 2021.

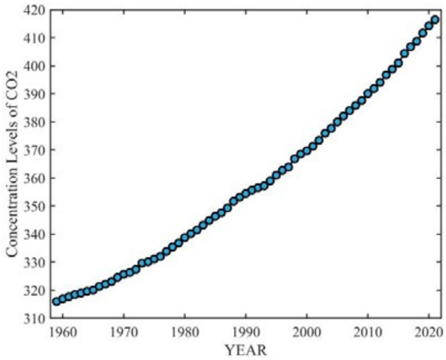


Figure 1. Concentration levels of CO2

As can be seen from Figure 1, the concentration of CO2 increased over time from the initial 315.98 PPM to 416.45 PPM. Since the industrial revolution, the rapid global economic development and the increased burning of fossil fuels have emitted large amounts of CO2 (Hansen, et al. 2012; IPCC, 2014; Solomon, et al. 2009; Le Quéré, et al. 2018; Hansen, et al. 1981; Anderson, 1987; Bonan, 2008; Clark, et al. 2016). Meanwhile, the rapid urbanization has led to the destruction of green plants. Various reasons make the amount of CO2 produced by unreasonable human production activities much higher than the amount absorbed by the ecosystem, which eventually leads to the accumulation of CO2 concentration.

Solar short-wave radiation can reach the ground through the atmosphere, but the large amount of long-wave heat radiation released by the heating of the surface is absorbed by the atmosphere, raising the temperature of the surface and the lower atmosphere. This is known as the greenhouse effect. Since the industrial revolution, the increase of greenhouse gases such as CO2 has increased the greenhouse effect of the atmosphere, disrupting the atmosphere's state of maintaining the Earth's energy balance and causing the overall temperature of the Earth to gradually increase, which is the most important factor causing global warming (Tans, et al. 2007; Brasseur & Solomon, 2005; Hansen, et al. 2006; Solomon, et al. 2007; Ciais, et al. 2005; Karl & Lenny, 1999; Hulme, 2002; Ruddiman & Thompson-Ellis, 2001; Brown, 2000.).

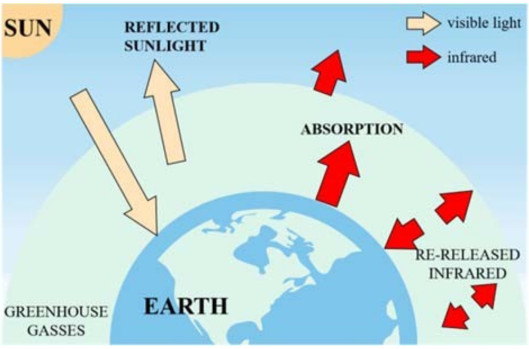


Figure 2. Diagram of the greenhouse effect

In this paper, we develop a mathematical model to fit the atmospheric CO2 concentration level to describe the past and predict the future atmospheric CO2 concentration level. The optimal fitting model is derived by comparison, and provides some data support for future government decisions. In Section 2, this paper first analyzes the CO2 concentration from 1959 to 2021 and builds three different prediction models. Then, we fit and regress the data with quadratic polynomial and exponential functions to establish regression prediction models, which are solved by the least squares method. After analyzing the smoothness of the data, we established a time series prediction ARIMA model. Finally, the SVM method is used to fit the prediction. Section 3 concludes the paper.

2. Establishment and Solution of CO2 Concentration Prediction Model

2.1 Growth Rate of CO2 in Different Years

A comparative analysis is first performed by calculating the growth rate of CO2 concentration for each decade from 1969 to 2021. The formula for this calculation is as follows:

$$R = \frac{M_i - M_{i-10}}{M_{i-10}} \tag{1}$$

where R is the growth rate value in the year i and M_i is the CO2 concentration value in the year i. The calculation results are shown in Figure 3 below.

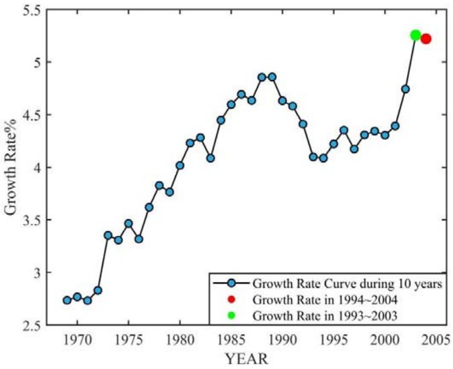


Figure 3. Growth rate of CO2 concentration levels

As can be seen from Figure 3, the growth rate of CO2 concentration is high into the 21st century. This is due to the further industrialization of the world in the 21st century and the massive burning of fossil fuels. Meanwhile, the large increase in the number of cars per capita leads to a large number of tailpipe emissions, and the large-scale destruction of forests by human activities will lead to a rapid increase in CO2 emissions in the new century.

2.2 Modeling of CO2 Concentration Prediction

In order to describe the past and predict the future atmospheric CO2 concentration, this paper uses nonlinear regression, a time series model, and an SVM model for solution analysis.

2.2.1 Establishment and Solution of Nonlinear Regression Model

Here a nonlinear regression model such as quadratic polynomial and exponential function is adopted to fit the CO2 concentration prediction according to the data characteristics. We use least squares to solve for the unknown parameters in the nonlinear regression model and use the nlinfit function in MATLAB to fit the regression to the scatter plot using quadratic polynomial and exponential functions.

Fitting results of the quadratic polynomial

The basic mathematical expression of the quadratic function is

$$y = ax^2 + bx + c, \quad \text{where} \quad a \neq 0 \tag{2}$$

The annual CO2 concentration values from 1959 to 2021 are substituted into the model and the final regression equation is solved. This is shown below.

$$\mathrm{Y} = 0.013\mathrm{x}^2 - 50.2758\mathrm{x} + 48771 \tag{3}$$

Fitting results of the exponential function

The basic mathematical expression of the exponential function is

$$y = ae^x + b, \text{ where } \quad a \neq 0 \tag{4}$$

To obtain the regression equation, we substitute the annual CO2 concentration values from 1959 to 2021 into the model, and the final regression equation is as follows.

$$\mathrm{Y} = 3.7421\mathrm{e}^{0.0026\mathrm{x}} - 271.5663 \tag{5}$$

The results of the regression are shown in the figure below.

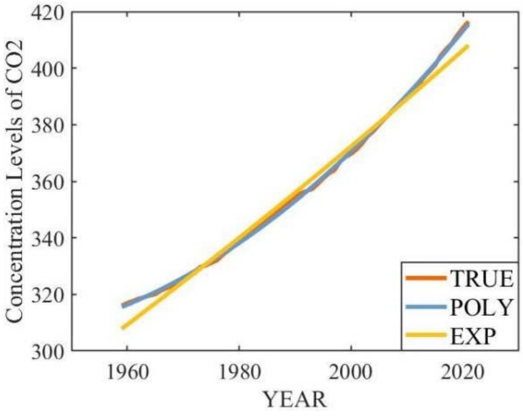


Figure 4. Results of nonlinear regression

From the figure 4, it can be found that the fitted curve roughly overlaps with the original curve. Then the goodness of fit is tested, as shown in the following table.

Table 1. Goodness of fit

method	PLOY	EXP↵
R^2	0.9994	0.9869↵

From Table 1, it can be seen that the goodness of fit of both fitting methods is excellent, and the R^2 of the quadratic polynomial fit is 0.9994, which is greater than the result of the exponential function fit.

2.2.2 Establishment and Solution of the ARIMA Model

Before using the ARIMA model for forecasting, the data need to be analyzed for smoothness. We use the ADF test to determine its smoothness. The ADF test is performed on the sample data using Stata software, and the results are as follows.

Table 2. Test results of ADF for raw data

	Test Statistic	1% Threshold value	5% Threshold value	10% Threshold value
Z(t)	8.683	-3.563	-2.920	-2.595

MacKinnon approximate p-value for Z(t) = 1.0000

The Z_test is greater than the critical value of each test and the p-value is greater than 0.05, so the original series is a non-stationary time series. Next, it is subjected to second-order difference treatment, and after the treatment, the ADF test is performed again. The results are as follows.

Table 3. ADF test for second-order difference data

	Test Statistic	1% Threshold value	5% Threshold value	10% Threshold value
Z(t)	-11.371	-3.566	-2.922	-2.596

MacKinnon approximate p-value for Z(t)= 0.0000

At this point, Z_test is less than the critical value of each test, and the p-value is less than 0.01. This indicates that after the second-order difference, the series becomes a smooth time series, as shown below.

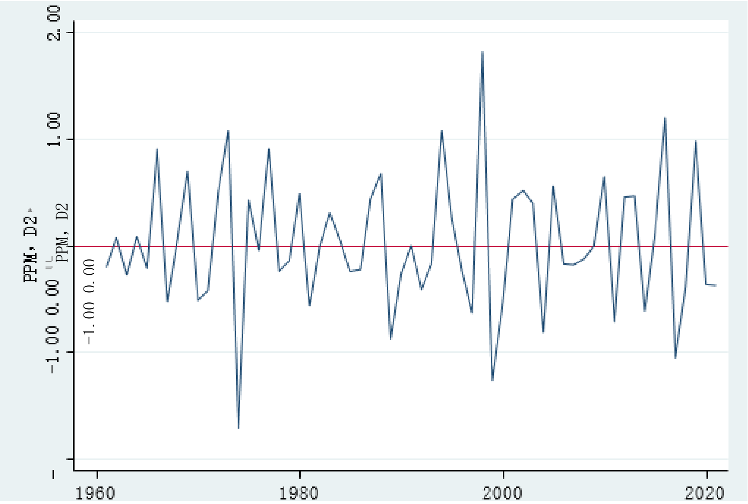


Figure 5. Sequence diagram after second-order differential processing

It can be observed from Figure 5 that the time series after second-order differencing is very close to the smooth time series.

Determination of ARIMA model parameters

In the smoothness analysis, the parameter d in the ARIMA model is determined as 2. Next, the other two parameters p and q need to be determined. The values of AIC and BIC are used to determine the quality of the model. The smaller the value of AIC and BIC, the better the model. The results of the model with multiple parameters are analyzed and judged using Stata, and the ARIMA (0, 2, 1) model is found to have an AIC of 81.63 and a BIC of 85.85, indicating that it is the best model. Therefore, p = 0, q = 1, and d = 2 are the optimal parameters for the ARIMA model.

ARIMA modeling

The ARIMA (0, 2, 1) model is used to predict the future CO2 content, and the model is represented as follows.

$$(1-\mathrm{L})^2 y_t = \propto_0 + (1+\beta_i \mathrm{L})\varepsilon_t$$

$$\rightarrow (1-2\mathrm{L}+\mathrm{L}^2 y_t) = \propto_0 + (1+\beta_{1i}\mathrm{L})\varepsilon_t$$

$$\rightarrow y_t = \alpha_0 + 2y_{t-1} - y_{t-2} + \varepsilon_t + \beta_1\varepsilon_{t-1} \tag{6}$$

The error of the model is calculated using Stata, and the results are as follows.

Table 4. Fitting effect of the ARIMA model

	Stationary R-squared	R-squared	BIC
Value	0.485	1.000	-1.417

Next, a white noise test is performed on the residuals, and the results are shown in the figure below.

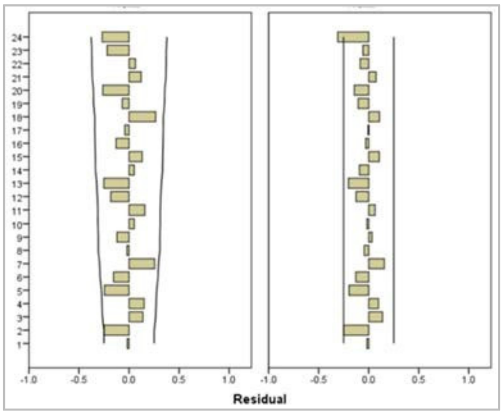


Figure 6. Residual white noise test

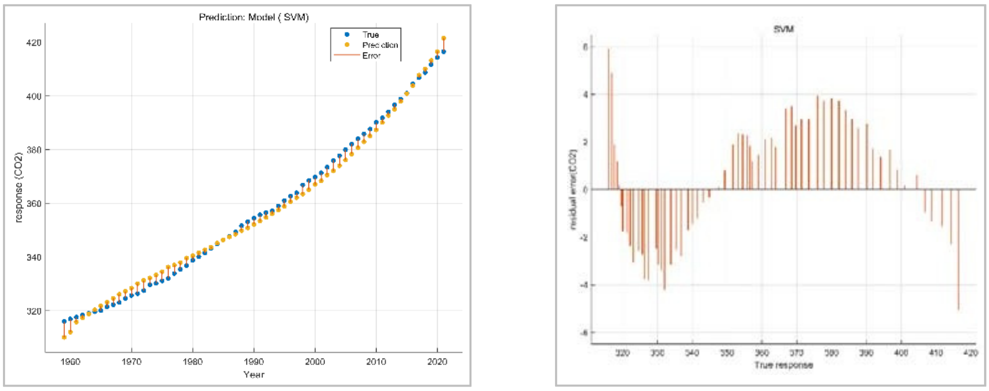
Based on the images of the residual ACF and PACF, the autocorrelation and partial autocorrelation coefficients are not significantly different from 0, so it can be determined that the residuals are white noise.

The coefficients in the ARIMA model are solved by SPSS software, and the obtained model expression is

$$y_t = 0.028 + 2y_{t-1} - y_{t-2} + \varepsilon_t + \varepsilon_{t-1} \tag{7}$$

2.2.3 Establishment and Solution of the SVM Model

Support vector machine (SVM) is a two-class classification model. The main idea is to find a hyperplane that separates the two classes of data points as correctly as possible, while making the separated two classes of data points the farthest away from the classification. The learning strategy is to maximize the interval, which can be formalized as a problem of solving convex quadratic programming. This learning algorithm is the optimal algorithm for solving convex quadratic programming. The SVM model is used to fit the given data and the results and residual plots are shown in the following figure.



a): Fitting result graph **b): Residual plot**

Figure 7. Effect and residual plots of the SVM fitting

2.3 CO2 Concentration Prediction

2.3.1 Regression Model

The obtained regression model is used to predict the CO2 concentration in 2100. In the quadratic polynomial, the concentration will reach 688.33 PPM in 2100, while the exponential function predicts 561.35 PPM. It can be found that both methods predict that the CO2 concentration will not reach 685 PPM in 2050, with the quadratic polynomial predicting 685 PPM in 2100 and the exponential function predicting 685 PPM in 2154.

2.3.2 ARIMA Model

The ARIMA model is used to predict the future CO2 concentration. The results are shown below.

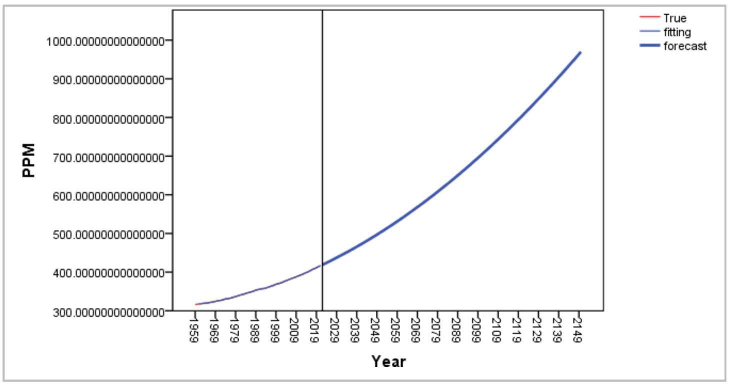


Figure 8. ARIMA prediction results

The CO2 concentration is predicted to reach 685 ppm around 2097 and 700.15 ppm in 2100, which means that the CO2 concentration will not reach 685 ppm in 2050.

2.3.3 SVM Model

Using this prediction model to forecast CO2 concentration from 2022 to 2100, the following results are obtained.

The CO2 concentrations are predicted to reach 1,316 ppm by 2100. CO2 concentrations are expected to reach 685 ppm between 2060 and 2061.

In summary, the three models predict that the CO2 concentration is unlikely to reach 685 ppm in 2050, which is inconsistent with the results of the CO2 concentration projections in the OECD academic report.

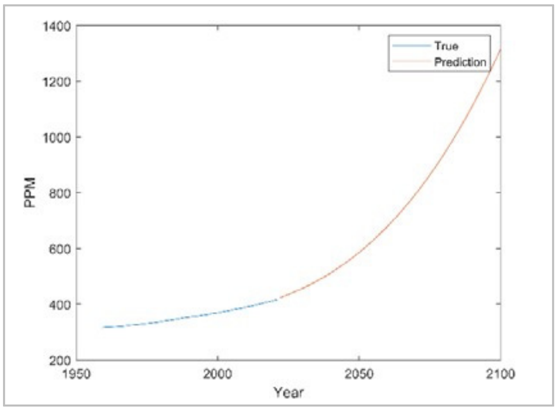


Figure 9. SVM prediction results

2.4 Comparison of Model Accuracy

The data from 1959-2009 are used as the training set and the data from 2010-2021 are used as the validation set to calculate the mean square error of the model. The results are shown in Table 5.

Table 5. Prediction accuracy of different models

Method	MSE
quadratic polynomial	0.0858
ARIMA	0.449
SVM	9.9

In summary, by comparing the accuracy of the three models, the quadratic polynomial model is finally determined as the model with the highest accuracy.

3. Conclusion

In this paper, several models, such as quadratic polynomial, ARIMA, and SVM models, are used for CO2 concentration prediction analysis. In addition, various tests such as smoothness analysis, normality test, and BP test are used for different prediction models. Meanwhile, the goodness-of-fit and mean square error of the models are calculated to ensure the reasonableness and validity of the models. Finally, the quadratic polynomial model is found to be the model with the highest accuracy.

References

[1] Hansen, J., Sato, M., & Ruedy, R. (2012). Perception of climate change. Proceedings of the National Academy of Sciences, 109(37), E2415-E2423.

[2] IPCC. (2014). Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC.

[3] Solomon, S., Plattner, G.K., Knutti, R., & Friedlingstein, P. (2009). Irreversible climate change due to carbon dioxide emissions. Proceedings of the National Academy of Sciences, 106(6), 1704-1709.

[4] Le Quéré, C., Andres, R.J., Boden, T., et al. (2018). Global Carbon Budget 2018. Earth System Science Data Discussions.

[5] Hansen J.E., Johnson D.W., Lacis A.A., Lebedeff S.E., Lee P.F., Rind D.H., Russell G.L.: Climate impact of increasing atmospheric carbon dioxide.Science (1981) 213:957–966.

[6] Anderson K.B.: The Greenhouse Effect and Global Warming.Climatic Change (1987) 10:9–13.

[7] Bonan G.B.: Forests and Climate Change: Forcings Feedbacks and the Climate Benefits of Forests. Science (2008) 320:1444–1449.

[8] Clark P.U., Shakun J.D., Marcott S.A., Mix A.C., Eby M., Kulp S.A., Levermann A. (2016). Consequences of twenty-first-century policy for multi-millennial climate and sea-level change. Nature Climate Change, 6(4), 360-369.

[9] Tans, P., Keeling, R., & Walker, J. (2007). Observational constraints on recent increases in the atmospheric CH4 burden. Geophysical Research Letters, 34(17).

[10] Brasseur, G.P., & Solomon S. (2005). Aeronomy of the Earth's Atmosphere and Ionosphere: IAGA Special Sopron Book Series. Springer.

[11] Hansen J., Sato M., Ruedy R., et al. (2006). Global temperature change. Proceedings of the National Academy of Sciences USA, 103(39), 14288-14293.

[12] Solomon, S. et al. (2007). Climate change 2007: The physical science basis – Summary for policymakers, fourth assessment report of the intergovernmental panel on climate change (IPCC).

[13] Ciais, P., Reichstein M., Viovy N., et al.(2005). Europe-wide reduction in primary productivity caused by the heat and drought in 2003.Nature 442(7101),81–84.

[14] Karl TR & Lenny LL(1999):The US historical climatology network.Monthly WeatherReview 127,1743–1765.

[15] Hulme M.(2002). Recent climatic fluctuations in the high Arctic. Ambio31(4),373–378.

[16] Ruddiman W&Thompson-Ellis J (2001):The Anthropogenic Greenhouse Era Began Thousands Of Years Ago.Climatic Change 49,401–406

[17] Brown TC, Bock AR &Jackson TJ (2000) Ground-based snow measurements in a montane watershed.Part 1:Spatial variability and topographic effects.Water Resources Research 36,3065–3075.