Can Machine Learning be Applied to Carbon Emissions Analysis: An Application to the CO₂ Emissions Analysis Using Gaussian Process Regression

Ning Ma¹, Wai Yan Shum², Tingting Han¹ and Fujun Lai³*

¹School of Financial Management, Hainan College of Economics and Business, Haikou, China, ²Department of Economics and Finance, The Hang Seng University of Hong Kong, Hong Kong, SAR China, ³School of Finance, Yunnan University of Finance and Economics, Kunming, China

In this paper, a nonparametric kernel prediction algorithm in machine learning is applied to predict CO₂ emissions. A literature review has been conducted so that proper independent variables can be identified. Traditional parametric modeling approaches and the Gaussian Process Regression (GPR) algorithms were introduced, and their prediction performance was summarized. The reliability and efficiency of the proposed algorithms were then demonstrated through the comparison of the actual and the predicted results. The results showed that the GPR method can give the most accurate predictions on CO₂ emissions.

Keywords: Gaussian process regression, CO₂ emissions, energy consumption, economics growth, industrialization

INTRODUCTION

As the population of the earth is being exponentially increasing, the exhaustion of carbon dioxide is increasing day by day resulting in the extreme overheating of the environment, which has become a significant reason for climate change. Global efforts to mitigate climate change were focused on the reduction of future days with extreme overheating of the environment. There are many research and surveys that have been conducted by various scientists, students, and other officials which were about the reasons for the high emission of CO₂ in different countries. Most of the empirical studies took the parametric modeling approach to analyze the factors that initiate and support the emission of CO₂. However, the traditional parametric approach optimizes a function to a known form with a set of finite and pre-determined parameters. This rigidity limits the predictive power of the parametric models. In recent years, nonparametric machine learning techniques have played dominantly with the enhancement of the forecast.

In this paper, a Bayesian nonparametric kernel prediction algorithm in machine learning is applied to predict CO₂ emissions. A literature review has been conducted so that the proper independent variables have been identified. Classical least squares, robust least squares, and algorithms of the GPR were introduced and their prediction performance, including the evaluation criteria that are effective in the measurements for model performance, were summarized. The reliability and efficiency of the proposed algorithms were then demonstrated through the comparison between the actual data and the predicted results. It is found that GPR can give the most accurate predictions on CO₂ emissions.
LITERATURE REVIEW

The growth of the economy, energy utilization, and CO₂ emissions are deeply related to each other. Kolstad and Krautkraemer (1993) point out that while the use of resources like the energy has a bright side on growth, it has negative environmental impacts. Traditional growth theories like the Solow growth model failed to consider the environmental impacts of growth (Solow, 1956). More modern growth theories study the interrelationship among energy, the environment, and economic growth (see, for example, Kolstad and Krautkraemer (1993), Jorgenson and Wilcoxen (1993) or Xepapadeas (2005) for a brief review).

Empirical studies depict that the growth of the economy and the ingestion of energy incorporates with the process of CO₂ emission. Recently, Hu et al. (2020) study the dominant reasons for carbon emission among the Belt and Road countries and find that CO₂ emissions have increased significantly due to economic growth. Similarly, Shabaz et al. (2013) found that in Indonesia, the emission of CO₂ increased for the extreme boost of the economic zone, while Shalahbazar et al. (2016) found that economic growth led to CO₂ emissions in Bangladesh and Egypt. Meanwhile, other studies discovered a bilateral causal relationship among the three variables. Munir et al. (2020) prove the fact that there is a relation of aftermaths and economy between GDP and energy ingestion in the major countries of the ASEAN (Association of Southeast Asian Nations), while Liu and Hao (2018) find that in energy-exporting countries, there is a bilateral relationship which may be a full-duplex connection between CO₂ emissions, energy utilization, and GDP per capita. Similarly, a repeating loop effect is observed between energy ingestion, CO₂ emission, and the advancement of the economy by Kahouli (2018). Accordingly, Mohmmand et al. (2020) observed the working principle of the causal relationship among transportation infrastructure, economic growth, and transportation emissions from 1971 to 2017 in Pakistan. The results show short-term causality from transportation infrastructure, economic growth, fuel consumption to CO₂ emissions, and the long-run relationship between economic advancement and infrastructure.

Apart from the growth and energy consumption, industrialization, population growth, and income level also contributed a great share in global carbon emissions. Minx et al. (2011) found that "industrialization" can be taken into consideration for the rapid increase of carbon dioxide emission in China from 2002 to 2007 while Zhang et al. (2014) found that the growth of the tertiary industry can decrease the CO₂ exhalation intensity. Nasir et al. (2021) examined the connection between the factors which are the exhalation of CO₂, industrialization, growth of the economy, energy ingestion, and several connecting factors from 1980 to 2014 in Australia. The observations of those involved say that all variables affect CO₂ emissions. Li et al. (2021) discussed the effect of the growth and structure of the economy on per capita CO₂ emissions in 147 countries from 1990 to 2015. The results show that at the global level, economic growth and economic structure are the most significant positive and positive effects, respectively.

Studies on population have thus far concentrated on the relationship between population growth and emission increase. The effect of population growth on CO₂ emissions can be summarized as follows (Birdsall, 1992): On one side, the energy demand was increased for power generation, industry, and transport. On the contrary, it increased deforestation emissions due to population growth. Empirically, Knapp and Mookerjee (1996) conducted a Granger causality test on annual data from 1880 to 1989 to determine the connecting clauses between global population expansion and carbon dioxide exhaustion. The results show there is a short-term dynamic relationship between the exhaustion of carbon dioxide and population growth. Very recently, Zhang et al. (2020) analyzed the knot between CO₂ emissions, GDP, and fuel ingestion in China and ASEAN countries. It was found that carbon density, energy intensity, GDP, and population are positively correlated with CO₂ emissions. Empirical findings also show that the developing countries are facing the effect of overpopulation and that’s why, they are facing more of a carbon emissions record per year other than the developed countries (Shi, 2003).

In the past decade, the theory and methodology of the Environmental Kuznets Curve (EKC) have been used to analyze the relationship between the net income and exhaustion of carbon of an area (Dinda, 2004; Williams and Rasmussen, 1996). According to the EKC, at relatively low-income levels, emissions increase as income increases. After a certain point, emissions will decline with income. Thus, the emission of CO₂ varies concerning the level of income. Luo et al. (2021) investigated the influencing factors of Shanghai’s CO₂ emissions from 1995 to 2017. They found that personal disposable income is one of the top drivers of CO₂ emissions. Yuan et al. (2014) examined the long-term relationship between China’s per capita income, ingestion of energy, and the emission of CO₂ from 1953 to 2008. They found out, there is a unilateral Granger inter-relation between the gross national income and the emission of CO₂.

Based on the literature above, it concludes that economic upgradation, energy utilization, manpower density, industrialization, and income can be classified as the predominant factors affecting CO₂ emissions. Other factors might also affect CO₂ emissions in China. For example, R&D (Nguyen et al., 2020; Jones, 1995), financial development (Bhattacharya et al., 2017; Zaidi et al., 2019; Wang et al., 2020), the degree of foreign direct investment (Essandoh et al., 2020; Le et al., 2020; Khan and Rana, 2021. etc). This paper limits the focus on how well the different prediction models perform based on the information set which includes only the most predominant driver of CO₂ emissions and excludes those unimportant ones to be captured by the stochastic terms in the models.

METHODOLOGY

Gaussian Process Regression (GPR) method can be introduced as a non-parametric Bayesian regression method (Gershman and Blei, 2012). It captures a wide variety of relations between inputs
and outputs and lets the data determine the complexity of the underlying functions through the means of Bayesian inference (Williams, 1998). Considering the output $y$ of a function $w$ at input $x$ with independent and identically distributed random noise $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. The function accompanied with the distributed random noise can be presented as:

$$y = w(x) + \epsilon$$  \hspace{1cm} (1)

In classical linear regression, $w(x)$ is deterministic whereas the noise term is random. In Gaussian process regression, however, $w(x)$ is assumed to be random and follows a Gaussian process. A Gaussian process is an extension of multivariate Gaussian distribution to infinite dimensions; any finite subset sampled from the Gaussian process follows multiple Gaussian distributions (MacKay, 1998). The distribution over functions can be described with the help of the Gaussian process, where $x$ is applied as the input variable, $m(x)$ is denoted as the mean function, finally, $k(x, x')$ is known as covariance function. These two functions are defined respectively as:

$$m(x) = E[w(x)]$$
$$k(x, x') = cov[w(x), w(x')]$$.

A finite collection of function values sampled from the Gaussian process follows multiple Gaussian distributions:

$$[w(x^{(1)}), w(x^{(2)}), \ldots, w(x^{(n)})]^T \sim \mathcal{N}(\mu, K)$$  \hspace{1cm} (4)

where $K$ is a $n \times n$ (n by n) matrix with the entries $K_{ij} = k(x^{(i)}, x^{(j)})$ and $\mu$ has entries $\mu_i = m(x^{(i)})$. Given a training set that contains observation points $y = [y(x^{(1)}), y(x^{(2)}), \ldots, y(x^{(n)})]^T$ and function values $w = [w(x^{(1)}), w(x^{(2)}), \ldots, w(x^{(n)})]^T$, it follows that the conditional distribution $p(y|w)$ and the Gaussian prior $p(w)$ are $\mathcal{N}(w, \sigma_r^2I)$ and $\mathcal{N}(\mu, K)$, respectively. By definition, the set of observations $y$ and the set of function values $w$ follow a joint multivariate Gaussian distribution. The joint distribution $p(w, y)$, given $w$ is defined as

$$[\begin{array}{c} w(x^r) \\ y \end{array}] \sim \mathcal{N}(\begin{bmatrix} \mu^r \\ \mu \end{bmatrix}, \begin{bmatrix} k(x^r, x^r) & k(x^r, x) \\ k(x, x^r) & K + \sigma_r^2I \end{bmatrix})$$  \hspace{1cm} (5)

where $I$ will be considered as the identity matrix, $\sigma_r^2$ is the unknown variance of the random noise and $(k^i)_{ij} = k(x^i, x^j)$ for $i, j = 1, 2, \ldots, N$. Using the Bayesian rule, the predictive posterior, $p(w(x^r)|y) \sim \mathcal{N}(\overline{w}, \Sigma^r)$, can be obtained, and the mean $\overline{w}$ and variance $\Sigma^r$ are defined by

$$\overline{w} = \mu + (k^r)^T[K + \sigma_r^2I]^{-1}(y - \mu)$$
$$\Sigma^r = k(x^r, x^r) - (k^r)^T[K + \sigma^2_rI]^{-1}(k^r)$$  \hspace{1cm} (6)

The covariance function determines the characteristics of the Gaussian method that can be expressed as $k(x^{(i)}, x^{(j)})$. The covariance function models the dependence between the function values at different input points $x^{(i)}$ and $x^{(j)}$. The covariance function is often called the kernel of the Gaussian process. There are many possible options for the prior covariance function. A popular kernel is the exponential covariance function which allows the model to general a non-negative definite covariance matrix for any set of input points (Williams and Rasmussen, 1996). The exponential covariance function is defined as

$$k(x^{(i)}, x^{(j)}) = \sigma_f^2 \exp \left( -\frac{\|x^{(i)} - x^{(j)}\|^2}{2l^2} \right) + \sigma_n^2$$  \hspace{1cm} (7)

where $l$ is the characteristic length scale, $\sigma_f^2$ is the signal variance, and $\delta_{ij}$ is a Kronecker delta. The Gaussian process regression employs a set of hyperparameters $\theta$ including $l$, $\sigma_f^2$ and $\sigma_n^2$ to increase or reduce the priority correlation between points and consequently the variability of the resulting function. The hyperparameters $\theta$ can be optimized based on the log-likelihood framework:

$$L = \log p(y|\theta) = -\frac{1}{2} y^T C^{-1} y - \frac{1}{2} \log |C| - \frac{n}{2} \log 2 \pi, C = K + \sigma_r^2 I$$  \hspace{1cm} (8)

More details about the regression process of Gaussian can be researched and acknowledged in the book of Williams and Rasmussen (2006), available free online and is accessible via the link: www.GaussianProcess.org/gpml.

**EMPIRICAL RESULTS**

A literature review has been conducted so that five independent variables; namely: economic growth, energy consumption, population, industrialization, and income, have been identified. In this study, the GPR method and the other proposed algorithms are applied to study carbon emissions in China. Economic growth is approximated by GDP (100 million RMB), energy consumption is approximated by per capita energy consumption (tons of standard coal), the population is approximated by population size (10,000 people), industrialization is weighted by the percentage of secondary industry in China, and income is measure by the average annual salary (RMB).

The data of GDP, population size, energy consumption, percentage of secondary industry, and average annual salary are collected from the China City Statistical Yearbook. CO2 emissions data come from four main sources of energy consumption. These are electricity, fuel, heating, and transportation. Those data can be obtained and calculated through the China Urban Construction Statistical Yearbook, the China City Statistical Yearbook, and the submerged government Panel on the change of weather and climate. Since some of those data is not available after 2014, the data in this paper range from the year 2002–2014.

**Statistical Analysis of Prediction Results**

The commonly used criteria in prediction performance are used in this study to evaluate the validity of the fitting. In Table 1, the root means squared error (RMSE), the mean squared error (MSE), the R-square, and the mean absolute error (MAE) are shown, where a well-fitted model should have R-square close to 1, whereas the RMSE, the MSE, and the MAE should be as small as possible. As per the observation from
Table 1, Exponential GPR provides the best fit data as it has the smallest RMSE, MSE, and MAE, and an R-square closest to 1.

Data Visualization
Since the data set is large, which made it difficult to demonstrate and view the whole set of data, visualization methods are typically needed especially for representative scenarios. The prediction results were analyzed at the model level to see the all-over authenticity of the three models and at the individual component level to get a picture of the estimates produced by the three models over the range of some particular variable.

At the overall level, the comparison and deviation of the actual value and the predicted dimension of the emission of carbon dioxide are determined. Figure 1 demonstrates the comparison of actual value and prediction of CO₂ emissions predicted by the three models; for each model, the predicted value is plotted against the actual value. To have a good fit, each plot should resemble a straight line at 45°. However, compare with the exponential GPR model, for the classical least squares model and the robust least-squares model, the predicted values are larger than the actual values over the range of 3.5–4 logarithm units of CO₂ emissions. This means that the classical least squares model and the robust least-squares model are overestimating CO₂ emissions over a particular range compare with the exponential GPR model. The same issue can be observed from Figure 2 which shows the deviation of actual value and prediction of CO₂ emissions for the three models. Figure 2 shows that, compare with the other two models, the deviations for the exponential GPR model cluster more closely around the horizontal line which represents no deviations. It suggests that the exponential GPR model provides a much better fit than the other two models.

Apart from analyzing the prediction results at the overall model level, the all over performance of the three models is also be evaluated at an individual component level. At the individual component level, the estimates produced by the selected models are analyzed over the extended range of some particular variables. Figures 3–7 below plot the actual and predicted values of CO₂ emissions against each of the most predominant factors of the models.

Table 1 | Statistical Analysis of Prediction Results.

|                          | Classical least squares | Robust least squares | Exponential GPR |
|--------------------------|-------------------------|----------------------|-----------------|
| RMSE                     | 0.10964                 | 0.11045              | 0.083734        |
| R-Squared                | 0.96                    | 0.94                 | 0.97            |
| MSE                      | 0.012022                | 0.012196             | 0.0070114       |
| MAE                      | 0.08049                 | 0.079101             | 0.0589303       |

Figure 1 | Comparison of actual value and prediction of CO₂ emissions between the selected models. (A) Classical Least Squares. (B) Robust Least Squares. (C) Exponential GPR. Notes: 1) The horizontal axis represents actual CO₂ emissions in logarithm, and the vertical axis represents predicted CO₂ emissions in logarithm. 2) CO₂ emissions are measured in ten thousand tons of standard coal.
least-squares model are used, extreme deviations between the actual value and predicted value can be observed. In Figures 4A,B, it is determined that the classical least squares and the robust least-squares model overestimate the CO₂ emissions over the range of 2.7–2.8 logarithm units of population size. Similarly, in Figures 5A,B, CO₂ emissions are overestimated by the classical least squares and the robust least squares models over the range of 0.5–1 logarithm units of per capita energy consumption. In Figure 6, although not obvious, CO₂ emissions are overestimated by the classical least squares and the robust least squares models over the range of 40–50% of secondary industry in China.

Figure 7 shows how the predicted values deviate from the actual values when the independent variable is non-Gaussian for the presence of threshold data points from extreme references. As with the evidence presented above, extreme upward bias over a particular range can be observed when the classical least squares and the robust least squares are used; the models overestimated CO₂ emissions over the range of 4.5– to 4.75 logarithm units of average annual salary. The extreme bias disappears when the exponential GPR model is used. Moreover, when the exponential GPR model is used, the deviations between the actual values and the predicted values are smaller for the extreme data values observed over the range of 1– to 1.5 and 5 to 5.5 logarithm units of average annual salary.

In summary, Figures 3–7 show that predicted CO₂ emissions conditional on individual components (i.e., GDP, population size, energy consumption, and industrialization) are quite close to the actual values predicted using the exponential GPR model. Even the underlying distribution of
the independent variable is non-Gaussian. Meanwhile, extreme upward bias per component in the technique can be observed when the classical least squares and the robust least squares models are used. Thus, a conclusion may be drawn upon the study that the exponential GPR model gives the most accurate predictions on CO$_2$ emissions compared with the remaining models.

CONCLUSION AND FUTURE WORKS

In this paper, the Gaussian process regression method is proposed for CO$_2$ emissions analysis in China. The traditional linear regression approach is limited by its rigid functional form and the approach often encounters an overfitting problem. The Gaussian progress regression approach relaxes the parametric assumption by applying the Bayesian nonparametric inference approach. The preciseness and exactitude of the prediction of the exponential GPR were compared and discussed with the classical least squares and the robust least-squares model. Based on the outcome of the whole study, it is proved that the Gaussian progress regression algorithms can give the most accurate predictions on CO$_2$ emissions compared with the other two traditional models and thus is applicable for CO$_2$ emissions prediction analysis to enhance forecast performance.
The prediction performances of the selected methods discussed only focus on the six predominant factors affecting carbon emissions. Future research should focus on further reviewing the completeness of the set of driving factors and the effectiveness of model predictions, compared them with other commonly used models.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

NM proposed the conceptualization, methodology and funding acquisition. WS gave the formal data analysis and wrote the original formal draft. TH performed the data collection and original arrangement. FL gave formal methodology, writing-review and editing.

FUNDING

This study was supported by Hainan College of Economics and Business (Project Reference Number: hnjmk2021301) and the Scientific Research Foundation of Yunnan University of Finance and Economics (No. 2020B03).
REFERENCES

Bhattacharya, M., Rafiq, S., Lean, H. H., and Bhattacharya, S. (2017). The Regulated Coal Sector and CO2 Emissions in Indian Growth Process: Empirical Evidence over Half a century and Policy Suggestions. Appl. Energ. 204, 667–678. doi:10.1016/j.apenergy.2017.07.061

Birdsall, N. (1992). Another Look at Population and Global Warming. In Policy Research Working Papers; No. WP5 1020. Population, Health, and Nutrition, 1020. Retrieved from http://documents.worldbank.org/curated/en/985961468761956897/another-look-at-population-and-global-warming.

Dinda, S. (2004). Environmental Kuznets Curve Hypothesis: A Survey. Ecol. Econ. 49 (4), 431–455. doi:10.1016/j.ecolecon.2004.02.011

Essandoh, O. K., Islam, M., and Kakinaka, M. (2020). Linking International Trade and Foreign Direct Investment to CO2 Emissions: Any Differences between Developed and Developing Countries? Sci. Total Environ. 712, 136437. doi:10.1016/j.scitotenv.2019.136437

Gershman, S. J., and Blei, D. M. (2012). A Tutorial on Bayesian Nonparametric Models. J. Math. Psychol. 56 (1), 1–12. doi:10.1016/j.jmp.2011.08.004

Hu, M., Li, R., You, W., Liu, Y., and Lee, C.-C. (2020). Spatiotemporal Evolution of CO2 Emissions, Energy Consumption, Financial Development, International Trade and CO2 Emissions in Indonesia. Renew. Sustain. Energ. Rev. 25, 109–121. doi:10.1016/j.rser.2013.04.009

Shahbaz, M., Mahalik, M. K., Shah, S. H., and Sato, J. R. (2016). Time-varying Analysis of CO2 Emissions, Energy Consumption, and Economic Growth Nexus: Statistical Experience in Next 11 Countries. Energy Policy 98, 33–48. doi:10.1016/j.enpol.2016.08.011

Shi, A. (2003). The Impact of Population Pressure on Global Carbon Dioxide Emissions, 1975-1996: Evidence from Pooled Cross-Country Data. Ecol. Econ. 44 (1), 29–42. doi:10.1016/s0921-8009(02)00233-9

Solow, R. M. (1956). A Contribution to the Theory of Economic Growth. Q. J. Econ. 70 (1), 65–94. doi:10.2307/1884513

Wang, R., Mirza, N., Vasileva, D. G., Abbas, Q., and Xiong, D. (2020). The Nexus of Carbon Emissions, Financial Development, Renewable Energy Consumption, and Technological Innovation: What Should Be the Priorities in Light of COP 21 Agreements?. J. Environ. Manage. 271, 111027. doi:10.1016/j.jenvman.2020.111027

Williams, C. K. (1998). Learning in Graphical Models. Dordrecht ; Boston: Kluwer Academic Publishers. doi:10.1007/978-94-011-5014-9_23Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and beyond Williams, C. K., and Rasmussen, C. E. (2006). Gaussian Processes for Machine Learning. Cambridge, Mass: MIT Press.

Williams, C. K., and Rasmussen, C. E. (1996). Gaussian Processes for Regression. In Advances in Neural Information Processing Systems. Cambridge, Mass:London: Mit Press.

Xepapadeas, A. (2005). "Chapter 23 Economic Growth and the Environment.,” Economywide and International Environmental Issues. Editors Karl-GoertanM. and J. R. Vincent (Amsterdam: North Holland), 3, 1219–1271. doi:10.1016/s0921-8009(02)00233-9

Yuan, J., Xu, Y., and Zhang, X. (2014). Income Growth, Energy Consumption, and Carbon Emissions: The Case of China. Emerging Markets Finance and Trade 50, 169–181.

Zaidi, S. A. H., Zafar, M. W., Shahbaz, M., and Hou, F. (2019). Dynamic Linkages between Globalization, Financial Development and Carbon Emissions: Evidence from Asia Pacific Economic Cooperation Countries. J. Clean. Prod. 228, 533–543. doi:10.1016/j.jclepro.2019.04.210

Zhang, J., Fan, Z., Chen, Y., Gao, J., and Liu, W. (2020). Decomposition and Decoupling Analysis of Carbon Dioxide Emissions from Economic Growth in the Context of China and the ASEAN Countries. Sci. Total Environ. 714, 136649. doi:10.1016/j.scitotenv.2020.136649

Zhang, Y.-J., Liu, Z., Zhang, H., and Tan, T.-D. (2014). The Impact of Economic Growth, Industrial Structure and Urbanization on Carbon Emission Intensity in China. Nat. Hazards 73 (2), 379–595. doi:10.1007/s11069-014-1091-x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Ma, Shum, Han and Lai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.