Machine Learned Artificial Neural Networks Vs Linear Regression: A Case of Chinese Carbon Emissions

M J Sajid\textsuperscript{a,b}

School of Management, Xuzhou University of Technology, Xuzhou, Jiangsu, China
\textsuperscript{a}jawad.jaws@outlook.com; \textsuperscript{b}11910@xzit.edu.cn

Abstract. China is the topmost source of world’s carbon emissions. Keeping this in view, a lot of work has focused on evaluating the relation between the Chinese carbon emissions and its drivers. However, these works mostly employ different types and extensions of the regression model to estimate the relations. The popular machine learning approaches like the artificial neural networks (ANN) are mostly overlooked in this regard. Furthermore, the studies based on the ANN and its different extensions often boast its superiority over the regression analysis. This claim has also not yet analysed for the relationship between a region’s carbon emissions and their drivers. This study fills these critical research gaps. The results showed that the linear regression model with lesser ‘mean squared error’ outperformed the ANN model with linear activation code. This study can be a good starting reference for advanced future work on this much neglected research gap.

1. Introduction
It’s the duty of all people to take care of environment [1-3], but the increase in global energy emissions in 2017 indicate that we are failing in this duty [4]. It’s a well-established fact now that the energy related emissions are a major cause of the global warming. The warming up of the world’s average temperature beyond the IPCC average target of 1.5 degree Celsius can devastate the environment and destroy the balance of the ecosystem [5]. China has the largest global carbon footprint and simultaneously emits the highest amount of global carbon emissions [6]. Thus, China’s approach to carbon mitigation is vital to global climate protection efforts.

Over the years, regression and some other regression-based models have been commonly used to frame the relationship between carbon emissions and their key drivers. Specifically, for China, a number of studies have been undertaken to understand the relationship between important factors and China’s direct carbon emissions. [7] Modelled the relationship between carbon emissions and oil and gross domestic product, using panel data of 98 nations from 1971 to 2007; used the dynamic panel threshold model (DPTM) based on the regression model. [8] Employing ‘VAR (Vector auto-regression) and STRIPAT (Stochastic impacts by regression on population)’ models studied the impact of some key factors on carbon emissions of Chinese ‘commercial sector’. [9] used the ‘LMDI’ and the ‘STRIPAT’ models to analyse different scenarios for carbon emissions reduction of the ‘Chongqing city’ of China. [10] Also used the ‘STRIPAT’ model to study the impact of ‘Energy intensity, population aging, population size, and consumption/capita’ on the carbon emissions of thirty Chinese states. [6] Graphically presented the non-linear relation between the Chinese direct carbon emissions and the carbon footprint with some key socio-economic factors, again using various regression models.
Artificial neural network (ANN) is a type of machine learning which is modelled on the principles of a human brain [11]. Simply, an ANN model is based on the working of the biological brain neurons, functions as an ‘artificial brain’. Where the neurons perform non-linear processes based upon the inputs from a source to forecast the final yield [12]. The ANN has been proven as a powerful model in various fields of study. Artificial neural networks (ANNs) are now well developed and popular in literature. This technique has a significant ability to simulate and control different processes in which different parameters interact at the same time [13]. The ANN can predict factors much more powerfully compared to different forms of the regression analysis [14]. The ANN has the capacity to model the non-linear relationship between the independent input and the dependent output quantities [15]. The ANN model has been extensively used by scientific community from different fields for accurately predicting outcomes of miscellaneous models see for example [12], [13], [15-17]. Most of these studies argue the superiority of the ANN approach over the miscellaneous conventional regression approaches.

However, there has been not much work on the relationship between the carbon emissions and their ‘socio-economic’ factors through artificial intelligent (AI) machine learning approaches like the ‘artificial neural networks’. Specifically, for China which is the largest single source of world’s energy based carbon emissions. Furthermore, it’s also very important to answer the not attended important question that: does the ANN model is also superior to regression in predicting the relation between carbon emissions and its drivers? Or does the already available huge volume of literature using linear and advanced regression models is sufficient in modelling the linear or non-linear relationship between the China’s carbon emissions and various driving factors? This work tries to fill these important gaps in the literature.

2. Materials and methods

2.1. Materials
The socio-economic factors selected in this study are based upon the literature review by the author. On the basis of the literature some important factors like ‘GDP, urban agglomerations, urbanization, energy consumption, fossil fuel use, and population are selected for this study. The temporal data related to these factors was retrieved from the database of the World Bank [18]. The data related to the energy emissions of China was taken from Eora database1. This database provides China’s yearly energy data with a larger time span from 1970 to 2015. But due to the missing data for some variable relating to the earlier years, the time span taken for this study is from 1990 to 2015. Furthermore, the socio-economic factors of ‘Energy, carbon intensity and fossil fuel use’ respectively have missing values for the year of 2015, the 2015 values of these factors are forecasted using the function described below in the methodology section.

2.2. Methods
Pattern identifications, predicting and taxonomy especially for non-linear relations can be approximated much strongly by the artificial neural networks. The model self-adopts and is very ‘robust’, providing the answer to condition specified question [17]. Of the many models available, multi-layer feed-forward model owing to its formidable modelling capability, is the most common method for prediction under the ANN [12]. The three-layer feed forward model (used in this study) consists of three layers: (1) single input layer, (2) few hidden layers, (3) and a single output layer [17]. The train and the test set are proportioned in this study as 0.70 and 0.30, which is the usual ratio in case of artificial neural networks. So, it implies that we have three hidden layers for our model. In the first step the author forecasted missing data for the year 2015 for the factors of ‘carbon intensity, energy and fossil fuel use ‘based upon the following linear forecasting formula:

\[ y = k + \gamma x \]  

(1)
\[ k = y - \gamma x \]  
\[ \gamma = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})} \]

Where \( y \) denotes the dependent variable, \( x \) the independent variable, \( k \) denotes constant (also known as \( y \)-intercept and biasness), \( \gamma \) represents the regression coefficient. Equation number one is also known as the formula of the linear regression. In the next step the author performed the “feature scaling” of the data. If we choose a very great ‘learning rate’ the errors become greater in every period, due to the over shooting of the cost minimum, on the other hand a very trivial ‘learning rate’ will require the algorithm to converge a very large number of periods. The feature scaling gives the data of a study the property of a ‘standard normal distribution’. The formula used for scaling of the data set is described below:

\[ x' = \frac{x - \bar{x}}{d} \]

Where \( x' \) represents the scaled value of the dependent socio-economic variable. \( x \) is its original value, \( \bar{x} \) and \( d \) represent its mean and standard deviation. Because the purpose of this paper is to compare ANN with linear regression we create a simple linear activation function for the ANN. An ANN linear activation function is almost same as the linear regression function; it changes the weighted inputs of the ANN model linearly to the outputs. It is presented as below.

\[ y = \sum_{i=1}^{m} g_i x_i + s \]

Where \( y \) represents the single output obtained through the linear function, \( g_i \) are the weights assigned to the different input quantities \( i \). \( x_i \) presents the input quantities and \( s \) is the biasness. Finally, to compare the performance of the two models, we calculate the mean square errors of the two models.

\[ E_{MS} = \frac{1}{m} \sum_{i=1}^{m} (x_b - x_i)^2 \]

Where \( E_{MS} \) presents the mean square error of a given model, \( X_b \) presents original (observed) input values, \( X_i \) represents estimated values.

3. Results

3.1. Linear regression model

The residual error of the regression model was 0.01381 on 7 degrees of freedom, in simple words the residual shows the difference amongst the expected and the actual values. \( R^2 \) shows the amount of variance in the observed values presented through the model, so a value closer to one is always preferred. The \( R^2 \) value under the linear regression model was 0.9999, which indicates that the model is almost a perfect fit. Adjusted fit present the models’ goodness of fit, again a value closer to one should always be preferred, the value in our case was 0.9998, which also shows a quite high value almost near to a 100 per cent fit. F-statistic indicates the significant of a model, the F-statistic value for this model was 1.001e+04 on 10 and 7 DF, which shows the model is quite significant in predicting the values. P value presents the occurrence of the results by chance, thus a low P value should be preferred, and the P value of the regression model was 1.436e-13 which is extremely low. All these indicators indicate the linear regression is an extremely reliable model to present the relationship between the Chinese energy emissions (referred as PBA (production based emissions) and the driving factors.

Table 1, contains the details of the factors wise results from the application of the linear regression model. The first column presents the coefficients, including the constant also known as the \( y \)-intercept and all other dependent factors of the model. The \( y \)-intercept or constant in other words presents the
value related to the dependent variable when the value of all other input factors is zero. The second column represents the estimates, which are commonly known as regression coefficients, which show that if all other variables are held constant a unit increase in the value of that variable would result in how much increase in the value of the dependent variable. For example the estimate for GDP is positive 0.004795, which means the GDP will have a positive impact of 0.004795 on the dependent variable, per unit of increase in GDP value. Similarly, a unit increase in the use of renewable energy will impact the carbon emissions by -0.244129.

| Coefficients  | Estimate  | Std. Error | t value | Pr(>|t|) |
|---------------|-----------|------------|---------|---------|
| (Intercept)   | -0.00691  | 0.008749   | -0.79   | 0.45564 |
| GDP           | 0.004795  | 0.161745   | 0.03    | 0.97718 |
| Agglomeration | -0.30287  | 0.226131   | -1.339  | 0.22231 |
| Urbanization  | 0.24362   | 0.275167   | 0.885   | 0.40535 |
| Savings       | 0.294505  | 0.221227   | 1.331   | 0.22483 |
| Renewable     | -0.24413  | 0.051398   | -4.75   | 0.00208 ** |
| Intensity     | -0.01233  | 0.025677   | -0.48   | 0.64567 |
| Energy        | 0.33345   | 0.139199   | 2.395   | 0.04778 * |
| Fossil        | 0.070818  | 0.089357   | 0.793   | 0.45407 |
| Population    | 0.089859  | 0.076859   | 1.169   | 0.28061 |
| FDI           | -0.05318  | 0.018887   | -2.816  | 0.02593 * |

** presents the confidence of values, here *** shows 100 per cent confidence, ** presents 99.99 percent confidence, * shows 95 percent and a dot behind the p value shows 90 percent confidence, the p values without any of these indicates their confidence is not sure.

3.2. ANN linear model

The error of the ANN model was 0.0198, reached threshold was 0.0082 and a total of 60 steps were involved. Figure 1 exhibits the three hidden layer ANN model. The input nodes (dependent variables) are presented at the starting of the black lines, the single arrowed line at the finish presents the output layer for the Chinese carbon emissions (referred as PBA). The weights are presented as the values on the black lines. The bias known as y-intercept in linear regression model added at each step is presented by the blue line. Figure 1, presents the values of the weights and biases (value of intercept) added at the three hidden layers. For example, From GDP to hidden layer 1 the weight was positive 2.43, it became 0.35 at the hidden layer 2, and finally at the hidden layer 3 it was 0.01. Similarly, for renewable energy it was 0.79 at first hidden layer, 0.22 at second, and 0.05 at the third hidden layer. The value of bias (y-intercept) at first layer 0.63, at second 1.34 and at third hidden layer 0.47 respectively.
3.3. ANN vs linear regression

Now that the results from the linear regression and the artificial neural network with linear regression activation function have been presented in detail, the question that needed to be answered is that which one of them have performed better than the other in predicting the values. For this purpose the author has considered the mean squared errors. The model with the minimum value of the mean squared errors
will automatically consider to outperform the other. The closeness of the regression line to the observed (actual) values can be estimated via the ‘mean squared error’. The distances known as errors are estimated between the line of regression and the observed values by squaring the values, squaring removes the negative signs. Figure 2, presents the predicted and the observed (real) values under the two models. The ‘mean squared error’ for the ANN model was almost 0.0906 and for the linear regression model it was 0.0020. Based on these results it’s quite obvious that the linear regression model clearly outperformed the ANN model.

![Figure 2](image_url)

**Figure 2.** The predicted and the observed (real) values under the two models.

### 4. Conclusion

The purpose of this study was to present the relationship between the Chinese carbon emissions and the key factors, by using both the conventional regression and the machine learning approach of the artificial neural networks. Most of the works on the topic of carbon emissions and their drivers, have failed to consider the growing popularity of the ANN based prediction models for determining the relationship between the dependent and the independent variable. Also, the literature on the artificial neural network models boasts about the superiority of the ANN over the various extensions of the regression model. This study was also conducted to answer whether or not this also holds true for all kinds of situations, especially for modelling the relationship between the carbon emissions and their key drivers. The results were in strong support of the contrary, i.e. the linear regression model outperformed the ANN model with linear activation function. However, this cannot be considered as a general rule for the prediction of the relationship between different other factors and the carbon emissions of a region. At the same time only linear model for both the regression and the ANN modelling was considered in this study. Future work based on different factors or more advanced non-linear models for both the regression and ANN approaches can shed more light on this topic in detail. Still, this study can be a good starting reference for advanced future work on this much neglected research gap.

### References

[1] Q. Cao, W. Kang, M. J. Sajid, and M. Cao, “Measuring China’s carbon emissions based on final consumption,” *Energy Procedia*, vol. 152, pp. 853–862, 2018.

[2] Q. Cao, W. Kang, S. Xu, M. J. Sajid, and M. Cao, “Estimation and decomposition analysis of
carbon emissions from the entire production cycle for Chinese household consumption,” *J. Environ. Manage.*, vol. 247, pp. 525–537, Oct. 2019.

[3] M. Cao, W. Kang, C. Qingren, and M. J. Sajid, “Estimating Chinese rural and urban residents’ carbon consumption and its drivers: considering capital formation as a productive input,” *Environ. Dev. Sustain.*, pp. 1–22, 2019.

[4] M. J. Sajid, Q. Cao, and W. Kang, “Transport sector carbon linkages of EU’s top seven emitters,” *Transp. Policy*, vol. 80, pp. 24–38, May 2019.

[5] M. J. Sajid, X. Li, and Q. Cao, “Demand and supply-side carbon linkages of Turkish economy using hypothetical extraction method,” *J. Clean. Prod.*, vol. 228, pp. 264–275, Aug. 2019.

[6] M. J. Sajid, “Modelling Best Fit- curve Between China’s Production and Consumption-based Temporal Carbon Emissions and Selective Socio-economic Driving Factors,” in 2019 3rd International Conference on Power and Energy Engineering (ICPEE 2019), 2019.

[7] K. M. Wang, “Modelling the nonlinear relationship between CO2 emissions from oil and economic growth,” *Econ. Model.*, vol. 29, no. 5, pp. 1537–1547, 2012.

[8] A. Wang and B. Lin, “Assessing CO2 emissions in China’s commercial sector: Determinants and reduction strategies,” *J. Clean. Prod.*, vol. 164, pp. 1542–1552, 2017.

[9] X. Tan, L. Dong, D. Chen, B. Gu, and Y. Zeng, “China’s regional CO2 emissions reduction potential: A study of Chongqing city,” *Appl. Energy*, vol. 162, no. 2015, pp. 1345–1354, 2016.

[10] Y. Wang, Y. Kang, J. Wang, and L. Xu, “Panel estimation for the impacts of population-related factors on CO2 emissions: A regional analysis in China,” *Ecol. Indic.*, vol. 78, pp. 322–330, 2017.

[11] S. Chandrasekhar, “Artificial Intelligence (AI), Machine Learning, Deep Learning, Neural Networks – What Do These Terms Mean?,” *IredDrop*, 2019. [Online]. Available: https://1reddrop.com/2019/01/13/artificial-intelligence-ai-machine-learning-deep-learning-neural-networks-what-do-these-terms-mean/%0D. [Accessed: 30-Dec-2019].

[12] E. Tosun, K. Aydin, and M. Bilgili, “Comparison of linear regression and artificial neural network model of a diesel engine fueled with biodiesel-alcohol mixtures,” *Alexandria Eng. J.*, vol. 55, no. 4, pp. 3081–3089, 2016.

[13] D. Liu, Y. Yuan, and S. Liao, “Artificial neural network vs. nonlinear regression for gold content estimation in pyrometallurgy,” *Expert Syst. Appl.*, vol. 36, pp. 10397–10400, 2009.

[14] Missinglink, “Neural Networks for Regression (Part 1)—Overkill or Opportunity?,” *Neural Network Concepts*, 2018. [Online]. Available: https://missinglink.ai/guides/neural-network-concepts/neural-networks-regression-part-1-overkill-opportunity/. [Accessed: 29-Dec-2019].

[15] A. Castrounis, “Artificial Intelligence, Deep Learning, and Neural Networks, Explained,” *InnoArchiTech*, 2019. [Online]. Available: https://www.innoarchitech.com/blog/artificial-intelligence-deep-learning-neural-networks-explained. [Accessed: 30-Dec-2019].

[16] W. Wu, “Beyond Travel & Tourism competitiveness ranking using DEA , GST , ANN and Borda count,” *Expert Syst. Appl.*, vol. 38, pp. 12974–12982, 2011.

[17] F. Nakhai, M. R. Mosavi, A. Sam, and Y. Vaghei, “Recovery and grade accurate prediction of pilot plant flotation column concentrate: Neural network and statistical techniques,” *Int. J. Miner. Process.*, vol. 110–111, pp. 140–154, 2012.

[18] World Bank, “World Bank Open Data,” *Database*, 2019. [Online]. Available: https://data.worldbank.org/. [Accessed: 30-Dec-2019].