Automobile Emissions Prediction Through Artificial Neural Networks

Randy G. Chapa

College of Engineering and Computer Science
Electrical Engineering
University of Texas Rio Grande Valley - Edinburg, Texas USA
randy.chapa02@utrgv.edu

Abstract— **Climate change has been a topic of discussion all around the world for several years. Greenhouse gases (GHG) contribute to the warming of the Earth and to prevent further effects, countries such as the United States and Canada have implemented plans to achieve net-zero emissions in the future. It is no secret that the primary contributor to GHG in the atmosphere come from human activities, mostly through the burning of fossil fuels. Most citizens though have no impact on how these fossil fuels are burned, but they do have control over consumption through transportation. Through transportation, most vehicles have tailpipes that emit carbon dioxide (CO_2) which makes up the highest percentage of GHG. However, if a driver looking for a vehicle to purchase can know the projected CO_2 emissions over time, this can influence their decision on what to buy. This gives an ordinary citizen the power to contribute to the call for less GHG emissions. Vehicle manufacturers are doing their part to produce electric vehicles, which considerably have less emissions, but through the used market standard fuel reliant vehicles will still dominate the roads. Hence, being able to anticipate CO_2 emissions can help the environment and reach net-zero emissions in the next 20 or 30 years. This paper will showcase artificial neural networks as the most accurate technique to predict CO_2 emissions, while detailing advantages of this application.**

Keywords— **Neural networks, Greenhouse gases, linear regression**

I. Introduction

Climate change has been a hot topic around the world, with many countries establishing plans to limit the warming of the Earth. According to [1], transportation generates about 33% of GHG emissions, the most of any sector. This sector, arguably is the most accessible to the general public, allowing regular citizens to impact emissions throughout their lives. As a result, countries such as the United States and Canada have implemented plans to reduce GHG emissions. Additionally, the state of California has been actively seeking to ban the sales of new gas cars by the year 2030 [2]. Autonomous vehicles are continuing development and advancement too, and California is pushing for all future builds be electric. However, it is highly likely that even by the year 2030, the roads will still be dominated by gas vehicles, with used sales continuing to dominate the market. However, if an average driver wanted to make a difference to the reduction of emissions, researching a specific vehicle model's projected emissions can lead to a smart purchase if they must purchase a gas reliant vehicle. Access to CO_2 emissions by vehicle builds can be found online, and in this paper, we will reference the 2020 Canada fuel consumption ratings [3]. Simply by knowing the build of a vehicle, the fuel type, fuel consumption rating, and the size of the engine, one can make an accurate prediction of CO_2 tailpipe emissions per km driven. It is worth noting that studies on car manufacturers impact on reducing have been done such as in [4], recommending more strict guidelines to meet climate change regulations in their target countries. This isn't a method that everyday drivers can impact however, which is the goal of this study. If forecasting emissions is the objective, the question now lies with the method of prediction to be implemented for the best possible accuracy. Generally, the most dominant method on automobile data, regression techniques are tested as seen in [5], [6], [7], [8]. The objective of this paper is to introduce neural networks techniques for emissions prediction, in hopes to provide higher accuracy than regression methods.

The structure of the rest of the paper is as follows. Section II will be a description of GHG and its contributors in Canada. Section III will detail the methodologies of regression and neural networks. Here previous applications will be compared to the proposed neural network architecture, mainly through formulaic representations. Section IV will showcase the main results of the paper on the Canadian fuel consumption ratings 2020 data.

Details regarding this data and pre-processing will be present there as well. Finally, Section V contains the conclusion of this paper, with final takeaways regarding the emissions forecasting.

II. Greenhouse Gases in Canada

This paper will focus on the impact of transportation sector in Canada. Like other countries including the United States and China, Canada has taken steps to reduce emissions through the next 20 years. More specifically Canada aims to reduce overall GHG emissions by 30% by the year 2030 (in relation to 2005 emissions) and reach net-zero emission by 2050. [9] This plan comes from the Pan-Canadian Framework on clean growth and climate change, which began in 2016. Among the economic sectors that contribute to GHG emissions, transportation has been consistently the largest contributor. GHG emissions by economic sector dating back to 1991 can be seen in Figure 1.

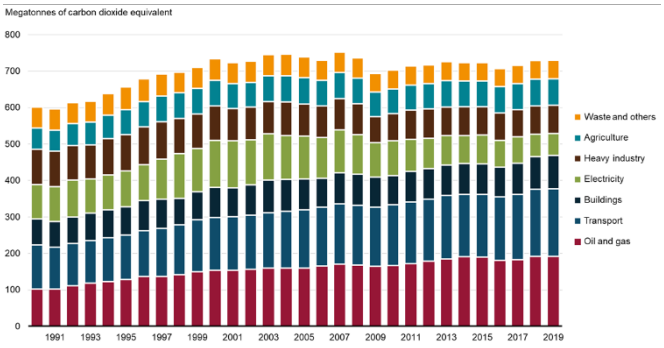

Fig. 1 GHG emissions by economic sector, Canada, 1990 to 2019.

As seen, in 2019, the transport sector was only behind oil and gas in CO_2 emissions. Transport accounted for 25% of total national emissions. These metrics should come as no surprise, as the number of cars and trucks on roads have increased substantially over the years. This holds true not only in Canada but all over the world. Back in 2007, 23% of global GHG emissions come from transportation sectors, with 73% coming from roads [10]. This is significant, as regular drivers can heavily impact this sector without much effort. If one wanted to impact emissions from air travel, they would have to pursue a career related to energy or engineering and hope to develop new methods of fuel. Being aware of what road vehicles release, this opens the door to planning when purchasing a vehicle, for any use. Getting back to Canada specifically, Figure 2 shows the breakdown of the transportation sector, among freight and passenger vehicles.

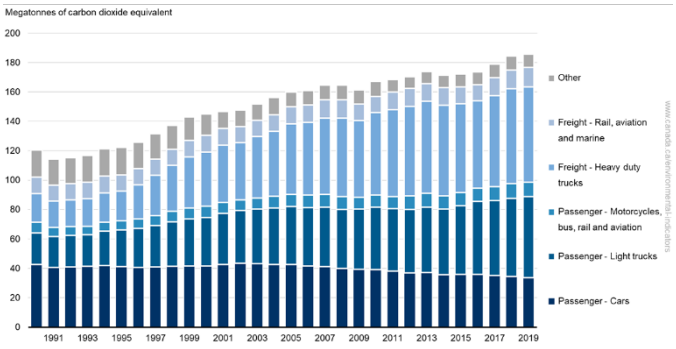

Fig. 2 Transport sector GHG emissions, Canada, 1990 to 2019.

Light trucks consist of sports utility vehicles, vans, and trucks, which doubled their emissions from 1990 to 2019 while cars emissions declined by 21%. Additionally, freight trucks have tripled their emissions. Once again it is reassured that most drivers can reduce emissions simply by purchasing more fuel-efficient vehicles, helping Canada progress towards its net-zero goal.

A. *Electric Vehicles*

It is important not to forget about the presence of electric vehicles (EVs), and how it will help with the battle to reduce CO_2 emissions. EVs undoubtedly have a significant difference in lifetime emissions compared to fuel dependent vehicles. However, we are not at a point where the electricity generated is entirely clean. Research on fuel-cycle emissions of EVs in China and U.S. is done in [11]. Results indicate that EVs impact on emissions reduction depend on cleanliness of electricity mix. In regions that don't rely mostly on coal-based electricity (such as California), EVs can reduce GHG emissions significantly compared to conventional vehicles. However, in China and Midwestern states in the U.S., coal is heavily relied on for electricity mix. EVs in these regions don't reduce GHG emissions as much and increase air pollutants. Projections in the study indicate that EVs charged with about 80% clean electricity are enough for 60-85% reductions in GHG. This goal has not been hit, and time will tell when

cleaner energy sources are discovered. Therefore, to bridge the gap to cleaner EVs, drivers must have an idea of emissions from road vehicles. Drivers looking to purchase vehicles through the new or used market need a resource to predict CO_2 emissions on the vehicle they are interested in. Vehicle investments are generally long-term and can help get us to the cleaner EV market 15-20 years down the road.

B. Other Efforts

Transportation is not the only sector contributing to GHG emissions, meaning there are other sectors we can investigate for reduction practices. The term net-zero regarding GHG simply means that the number of gases being released should be equal to the number of gases being taken out of the atmosphere. CO_2 removal mostly comes from photosynthesis, highlighting the importance of keeping forests alive. The issue lies with the use of forest bioenergy, which is discussed in [12]. Authors of the study aimed to identify bioenergy paths to contribute to the Pan-Canadian Framework's decarbonization targets. What these methods have in common, is how little the general public can contribute to the cause. Like mentioned previously, a young driver can know the emission rates of a vehicle and aim to get the most environmentally friendly choice. What's left is the process of prediction the emissions of a vehicle, and what parameters do drivers need to know in order to prediction emissions. The next section will review methods of prediction that are popular in dataset analysis.

III. Prediction Methodologies

Emissions analysis in the transportation sector have been investigated in many studies. These studies vary in the types of variables explored that relate to emissions which can include distance travelled, size of the vehicles, age of the vehicle, manufacturer, and model. Depending on the complexity of the data in question, a proper method can be chosen to fit the model. Such methods have been explored in literature, such as time series study, regression analysis, decomposition analysis, bottom-up method, and system optimization tools. In [13] the methods are compared to one another, with detailed tables listing pros and cons. Most prediction methods are valid, as each study varies in the data being looked at to determine emissions. One example is [14] that explores the connection between gross domestic product (GDP), salary, urbanization, and energy demand. Using regression techniques, it is determined that in growing cities, alternate energy sources must be studied to reduce CO_2 emissions. Similarly, in [15], economic growth, energy usage and emissions connections are explored. Saboori et al. went with a time series approach in this case. Time series techniques rely heavily on past behaviour to make predictions. While it is a valid technique there are flaws. Some variables that impact the emissions may not be an accurate indicator as time passes on. For example, emissions data going back to 1995 considers vehicles with much different builds and fuel consumption ratings. Also, if the data in question is large, computational costs become large. Most prediction models drive a high computational load however, especially if datasets are large. In the case of this paper though, the data in question is large with over 7,300 rows, but the dependent variables are not too complex to formulate. Therefore, the 2 methods to be compared are regression and neural networks (NN) techniques. There aren't many literature studies that support the use of NN, but this paper hopes to showcase valid results using a NN model. Interestingly, there are many studies that compare regression techniques with NN such as in [16,17 shrimp] to predict electricity consumption, as well as disease occurrence, respectively.

A. Linear Regression

Linear regression models depict a relationship between a dependent variable (y) and one or several independent variables (x) as:

$$\hat{y} = b_1 x + b_0$$

The method of linear regression this paper will focus on is the ordinary least squares (OLS) method. For this method, we are adjusting the values of b_1 and b_0 for the total sum of squares of the difference between actual and calculated measures of y is as low as possible. The formula for OLS is:

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_1 x_1 - b_0)^2$$
$$= \sum_{i=1}^{n} (\hat{\varepsilon}_i)^2$$

Where $\hat{y}_i$ is the predicted value for i[th] observation, y_i is actual value, $\hat{\varepsilon}_i$ is the error, and n is the total number of observations. To measure the performance of the model, R^2 scores are calculated which is just the percentage of the dependent variable variation that is created by the OLS model. The higher the percentage, the better the model is to predict the data.

B. Neural Networks

Neural networks are a type of model that are designed to emulate human brain neurons. Each neuron has connections that transmit signals to other neurons, which are triggered through actions involving the brain. In practice, each neuron carries a specific weight, in attempts to result in a single output. Outputs depend on the connections of the neurons, and the connections depend on the weighted neurons from the previous layer. A model depicting this is shown in Fig. 3. There are different algorithms to consider when designing a NN, but this paper will focus on the backpropagation (BP). BP is highly effective for pattern recognition and predictions.

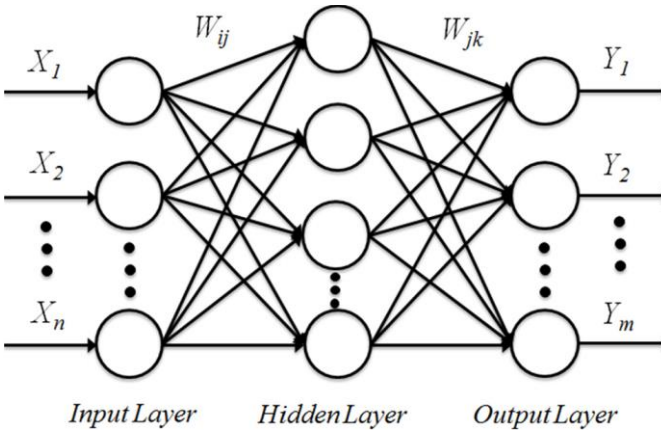

Fig. 3 Backpropagation neural network system

BP of a NN always have at least three layers: input, hidden and output. As noted in [18] the layer outputs are as follows:

Output value of hidden layer:

$$o_j = f \sum_{i=1}^{n} \left(w_{i_j} x_i - d_j \right) j = 1,2, \dots, l$$

Output value of output layer:

$$Y_k = f \sum_{j=1}^{l} \left(o_j w_{j_k} - d_k \right) k = 1,2, \dots, m$$

The number of hidden layers and number of neurons that make up a hidden layer is up to the user. As the data becomes more complex, the model will also become more complex as more neurons are needed. Unsurprisingly these models can become computationally expensive, needing thousands of datapoints to test and train on to learn behaviours. In the next section, we will utilize Python libraries to build our model for us, making the process extremely simple.

IV. MAIN RESULTS

Before getting into the main findings of the paper, a brief section will detail the dataset used for forecasting the CO_2 emissions.

A. 2020 Canadian Fuel Consumption Ratings

As mentioned previously, Canada is the country being investigated therefore it was appropriate to find data originating there. From [4], the data provides fuel consumption ratings and estimated CO_2 emissions for new light-duty vehicles for the year 2020. In this dataset, we can find information on the make, model, vehicle class, engine size, number of cylinders, transmission type, fuel type, fuel consumption and finally estimated CO_2 emissions. Some columns are irrelevant to this study such as make and model but can easily be incorporated into a similar study on the same data if one is interested. Table I lists the relevant columns in the data, along with their units. For fuel consumption columns, there are entries considering only city ratings, only highway ratings and combined ratings.

TABLE I
RELEVANT COLUMNS WITH UNITS

Engine Size	Cylinders	Fuel Consumption	CO2 Emissions
L	Unitless	L/100km	g/km

For unused columns such as fuel type and transmissions, one can convert the categorical data into numbers with ease to consider in another test. However, in this study they will not be considered. With over 7300 unique entries of data, there is plenty to build and test prediction models. Both methods being compared require testing and training datapoints. Training data are the points used to train the model and understand the behaviours. Testing is then used to test our model, to see if it can make valid predictions. Generally, it is best to have testing data account for 20-30%, and the rest into training. For this paper, 20% is being set for testing data and remaining 80% for training. Now we can get into the simulations and results.

B. Regression Results

As mentioned in the previous section, linear regression focuses on the relationship between independent and dependent variables. In this case there are five independent variables; engine size, cylinders, fuel consumption city, fuel consumption highway and fuel consumption combined. Before considering all independent variables though, let's look at how some of the variables relate to emissions individually. First, considering only fuel consumption combined, one can infer that if the vehicle has a high fuel consumption over 100km, then the emissions per km will also be high. This is confirmed from Fig. 4.

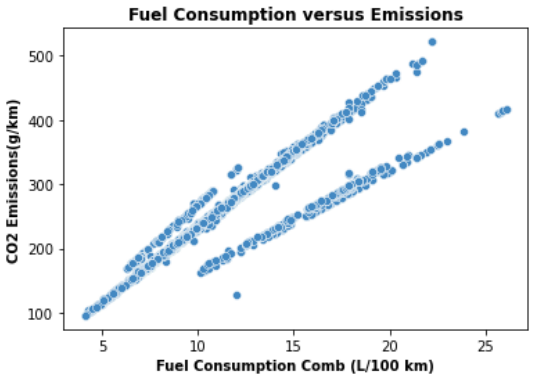

Fig. 4 Fuel Consumption vs CO2 Emissions

Similarly, if we consider the number of cylinders and how it impacts emissions. Generally, the more cylinders in a vehicle indicate better performance and more power. Figure 5 considers all unique cylinder entries and displays the impact on CO_2 emissions.

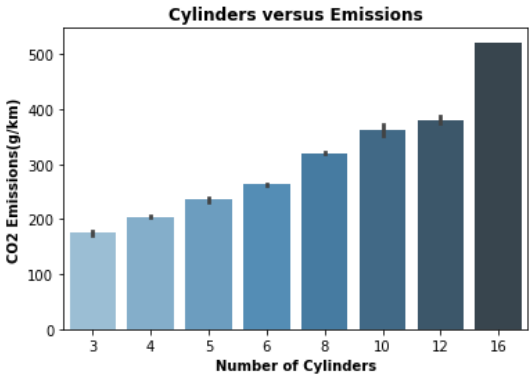

Fig. 5 Cylinders vs CO2 Emissions

Given the relationships of cylinders and fuel consumption, we can expect high accuracy scores when running the multiple variable regression model. To get results, the data gets ran through a Python script, and an OLS model is tested. Finally, a summary of the results is found in Figure 6.

OLS Regression Results

Dep. Variable:	CO2 Emissions(g/km)	R-squared:	0.874
Model:	OLS	Adj. R-squared:	0.874
Method:	Least Squares	F-statistic:	1.276e+04
Date:	Mon, 22 Nov 2021	Prob (F-statistic):	0.00
Time:	21:25:18	Log-Likelihood:	-32890.
No. Observations:	7385	AIC:	6.579e+04
Df Residuals:	7380	BIC:	6.582e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	61.3416	1.054	58.183	0.000	59.275	63.408
Engine Size(L)	12.8378	0.325	39.444	0.000	12.200	13.476
Fuel Consumption City (L/100 km)	2.9750	3.137	0.948	0.343	-3.175	9.125
Fuel Consumption Hwy (L/100 km)	0.8319	2.589	0.321	0.748	-4.243	5.907
Fuel Consumption Comb (L/100 km)	9.4575	5.691	1.662	0.097	-1.699	20.614

Omnibus:	2382.474	Durbin-Watson:	1.679
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9414.930
Skew:	-1.562	Prob(JB):	0.00
Kurtosis:	7.565	Cond. No.	575.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fig. 6 OLS Regression Results

The most important score is the R^2 score of around 87%. This is a valid score, and our model is suited to make predictions on our data. However, for regression testing it is best to have more than one measure for rating performance. So, for this study we will consider residual error to validate our model. Residual error is simply the difference between predicted and actual values. In our Python script, we run the model on the training and testing data and

compare the predicted Y values with the actual values. Results are shown in Figure 7.

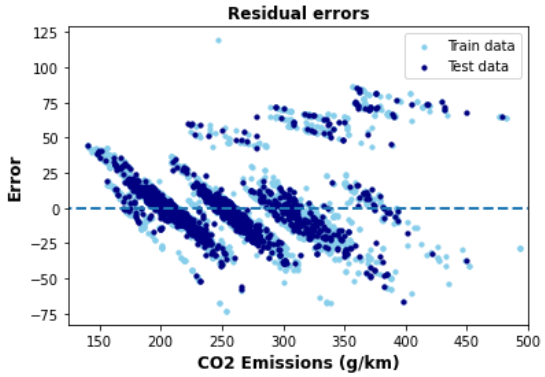

Fig. 7 Residual Errors

From the plot in Fig. 7, the dotted line origin represents absolute zero error. The closer the points are to the dotted line, indicates small error. As the points move away from the origin, the larger the error is. For the most part, errors are small, with some extreme outliers in the 70s and 80s. It is expected these vehicles are designed for performance only, not considering fuel economy or emissions. With a decent R^2 score and solid overall residual errors, it is valid to say this regression model performs well enough to forecast CO_2 emissions. However, the goal of this paper is to introduce a simple neural network model and check if the improved R^2 score will justify this method.

C. Neural Network Results

Considering the same independent variables from the regression model, a NN model is built to predict the dependent variable, CO_2 emissions. Once again, a Python script is setup to build a regression model, with 20% testing data. Adam is the algorithm of choice, which is one of the most common for NN models. For BP models, an activation function is utilized, and, in this case, the results did not differ much, ultimately going with a hyperbolic tangent function (tanh). Default learning rate is invscaling, which is left the same for this model. More information on these choices can be found in [20, 21]. To get high performance, the model must run up to several hundred iterations until the change in R^2 score is almost zero. For this script, the max possible iterations are set to five hundred. Once the model is setup through Python's sklearn library, simulations can run. Main results are found in Fig. 8, containing the predicted Y test data and the actual Y test data.

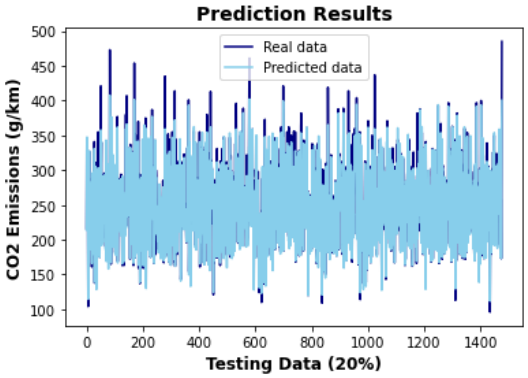

Fig. 8 Neural Networks Prediction Results

As seen in the plot, the NN model does a great job attempting to match the real data, with the navy-blue figures indicating the difference between the predicted and actual values. Additionally, a R^2 score of about 93% is achieved in most test runs. This leaves a difference of about 5% compared to the previous regression model. The only downside is the time it takes to complete a run. For NN models, it is generally more computationally expensive, which can take time if scripts are ran on lower end hardware. More problems arrive if using TensorFlow libraries in Python, but this isn't an issue in this study. It also helps that the data isn't too complex, not many float values so it takes under three minutes to run this script, based on 50 runs. Figure 9 visualizes the runtime of the script, and this can be much larger for complex data such as image analysis or if running on low end hardware.

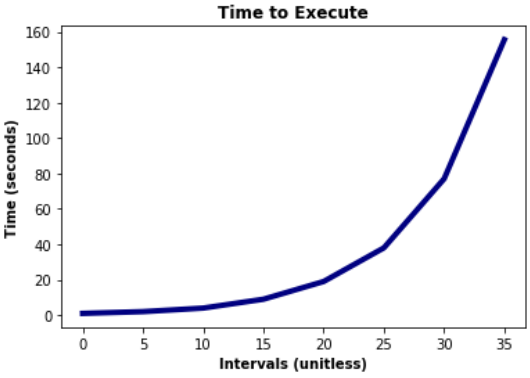

Fig. 9 Neural Networks Time Results

The x axis are just intervals to plot the elapsed time, and y axis is the elapsed time according to the created intervals. On average, elapsed time is under 180 seconds or under 3 minutes.

V. CONCLUSIONS

From the results, it is proven that the neural networks model outperforms the multiple linear regression model. However, the difference in score isn't significant enough to convincingly promote neural networks as the "best" method. In this study, it is more accurate, and the computational expense isn't drastic, therefore it is a viable option for GHG emissions forecasting. It should be utilized more often for forecasting, but it all depends on the dataset in question, more so involving the independent variables if they become more complex. Running tests for multiple prediction techniques is imperative to achieve the best possible performance, and this paper hopes to contribute to already existing research.

ACKNOWLEDGMENT

I would like to thank Dr. Nazari for motivating students to pursue meaningful topics for research. (This is a little informal given the assignment, but can be removed if further worked on.)

REFERENCES

[1] Thomas, C. E. (Sandy). "Greenhouse Gases by Sector." Stopping Climate Change: The Case for Hydrogen and Coal. Cham: Springer International Publishing, 2017. 9–11. Web.

[2] Executive Order. No. N-79-20, 2020, p. 5.

[3] Fuel consumption ratings - 2020 fuel consumption ratings (2021-09-29) [Internet]. Open Government Portal. [cited 2021Nov25]. Available from: https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64/resource/56a89c09-d609-41cd-8838-9dd9905d3cfc

[4] Voltes-Dorta A, Perdiguero J, Jiménez JL. Are car manufacturers on the way to reduce CO2 emissions?: A DEA approach. Energy Economics. 2013;38:77–86.

[5] Libao Y, Tingting Y, Jielian Z, Guicai L, Yanfen L, Xiaoqian M. Prediction of CO 2 emissions based on multiple linear regression analysis. Energy Procedia. 2017;105:4222–8.

[6] Xu B, Lin B. Does the high–tech industry consistently reduce CO 2 emissions? results from nonparametric additive regression model. Environmental Impact Assessment Review. 2017;63:44–58.

[7] Sajid MJ. Machine learned artificial neural networks vs linear regression: A case of Chinese carbon emissions. IOP Conference Series: Earth and Environmental Science. 2020;495(1):012044.

[8] Cha J, Park J, Lee H, Chon MS. A study of prediction based on regression analysis for real-world CO2 emissions with light-duty Diesel Vehicles. International Journal of Automotive Technology. 2021;22(3):569–77.

[9] Giuntoli J, Searle S, Pavlenko N, Agostini A. A systems perspective analysis of an increased use of forest bioenergy in Canada: Potential carbon impacts and policy recommendations. Journal of Cleaner Production. 2021;321:128889.

[10] JAMA - Japan Automobile Manufacturers Association, inc [Internet]. [cited 2021Nov25]. Available from: http://www.jama-english.jp/publications/2008_CO2_RoadTransport.pdf

[11] Huo H, Cai H, Zhang Q, Liu F, He K. Life-cycle assessment of greenhouse gas and air emissions of electric vehicles: A comparison between China and the U.S. Atmospheric Environment. 2015;108:107–16.

[12] Giuntoli J, Searle S, Pavlenko N, Agostini A. A systems perspective analysis of an increased use of forest bioenergy in Canada: Potential carbon impacts and policy recommendations. Journal of Cleaner Production. 2021;321:128889.

[13] Alhindawi R, Abu Nahleh Y, Kumar A, Shiwakoti N. Projection of greenhouse gas emissions for the road transport sector based on multivariate regression and the double exponential smoothing model. Sustainability. 2020;12(21):9152.

[14] Sadorsky P. The effect of urbanization on CO2 emissions in emerging economies. Energy Economics. 2014;41:147–53.

[15] Saboori B, Sapri M, bin Baba M. Economic growth, energy consumption and CO2 emissions in OECD (Organization for Economic Co-operation and Development)'s transport sector: A fully modified bi-directional relationship approach. Energy. 2014;66:150–61.

[16] Pombeiro H, Santos R, Carreira P, Silva C, Sousa JMC. Comparative assessment of low-complexity models to predict electricity consumption in an institutional building: Linear regression vs. Fuzzy Modeling vs. Neural Networks. Energy and Buildings. 2017;146:141–51.

[17] Leung PS, Tran LT. Predicting shrimp disease occurrence: Artificial neural networks vs. logistic regression. Aquaculture. 2000;187(1-2):35–49.

[18] Kim MK, Kim Y-S, Srebric J. Predictions of electricity consumption in a campus building using occupant rates and weather elements with sensitivity analysis: Artificial neural network vs. linear regression. Sustainable Cities and Society. 2020;62:102385.

[19] Shao Y, Dietrich FM, Nettelblad C, Zhang C. Training algorithm matters for the performance of Neural Network Potential: A case study of adam and the Kalman filter optimizers. The Journal of Chemical Physics. 2021;

[20] Anastassiou GA. Multivariate hyperbolic tangent neural network quantitative approximation. Intelligent Systems Reference Library. 2011;:89–107.