MODELING THE IMPACT OF ELECTRIC VEHICLE ADOPTION ON NO2 LEVELS USING MACHINE LEARNING: A PREDICTIVE ANALYSIS

ARNAV GUPTA
arnav.gupta0103@gmail.com
Delhi, India

ABSTRACT

Air pollution, notably nitrogen dioxide (NO2), poses severe health and environmental risks. The main sources of NO2 are related to transportation and stationary fuel combustion sources [U.S Environmental Protection Agency (EPA), 2018]. With a global shift to electric vehicles (EVs), their potential to mitigate pollution is becoming prominent. In India, where air quality concerns are paramount, understanding the tangible benefits of EV adoption is crucial. The research question explores whether increasing EVs adoption can visibly reduce NO2 levels. The approach combines use of various supervised machine learning models and simulations to understand the impact of socio-economic factors, adoption of electric vehicles, energy generations and emission factors on the NO2 levels and to create a model to predict the level of NO2. The created model will be used to estimate the impact different EV adoption levels can have on NO2. To train the model, data from 49 US states along with 5 Indian states where EV sales are comparable and the NO2 data was available has been used. Among diverse algorithms tested, the Neural Network turns out to be the best which predicts NO2 levels with a R2 value of 0.82. Simulations, varying EV numbers while holding other factors steady, demonstrated a clear link between higher EV count and lower NO2 levels. The results not only validate the potential of electric vehicles as catalysts for improving air quality but also highlight their crucial role within the Indian context, especially in the state of Delhi which we have used as our case study. The empirical evidence underscores the pivotal contribution that EVs could make to enhancing Delhi's air quality profile. In conclusion, the research provides substantial evidence that the widespread adoption of electric vehicles holds promise as an effective strategy to reduce NO2 levels, thereby laying the foundation for a more sustainable and healthier urban environment

INTRODUCTION

Nitrogen dioxide (NO2) primarily gets in the air from the burning of fuel. NO2 forms from emissions from cars, trucks and buses, power plants, and off-road equipment. This pollutant contributes to air pollution, respiratory issues, and cardiovascular problems in humans [U.S. Environmental Protection Agency (EPA), 2016]. Prolonged exposure to NO2 from fossil-fueled vehicles has serious health implications, especially in densely populated urban areas. Transitioning to electric vehicles (EVs) and adopting cleaner fuels can help reduce NO2 emissions, improving air quality and public health while addressing the environmental impact of transportation.

In response to these concerns, many countries are actively transitioning their transportation systems away from fossil fuels. EVs powered by renewable energy sources offer a promising alternative, producing minimal tailpipe emissions and reducing GHG emissions from transportation. EV sales have been steadily increasing over the years. In 2020, despite the challenges posed by the COVID-19 pandemic, global yearly EV sales reached around 3 million units [IEA (2021)].

India has recognized the importance of transitioning to EVs as part of its strategy to address air pollution, reduce greenhouse gas emissions and promote sustainable transportation. The country has implemented a range of policies and initiatives to encourage the adoption of EVs and to create a supportive environment for their growth. It has set an ambitious goal of achieving 30% EV penetration for new vehicle sales by 2030 [E-AMRIT 2019]. The Delhi Government aims to have 1 out of every 4 vehicles sold in Delhi by 2024, to be an EV [Delhi Electric Vehicles (EV) policy (2020)]. While there is substantial theoretical understanding of how EVs could contribute to cleaner air, translating these benefits into real-world impacts requires comprehensive observational evidence. Empirical assessments of environmental benefits made by an extensive transition to EVs is missing.

This study aims to examine the impact of EV numbers on pollution levels within a particular state through the use of machine learning models. The research aims to determine the benefits of EVs by using predictive modeling based on factors like EV adoption, socio economic census data, energy generation, and emission factors, focusing on chosen US states and Indian regions. We chose the USA as the basis of our model due to its favorable EV policies, which align with India's evolving efforts (as documented in the Niti Aayog paper outlining new policies). Moreover, the extensive and diverse US dataset includes state-level information akin to India's context. The investigation employs AI techniques to establish connections between EV usage and NO2, forecasting NO2 levels based on EV-related variables.

BACKGROUND

Several research studies have explored the connections between EV adoption, air quality, and socioeconomic conditions. While Singh, Anuradha & Yadav, Jyoti & Shrestha, Sarahana & Varde, Aparna. (2023) [1] provides valuable insights into the associations between adoption patterns, socio-economic status and air quality, it does not take into account other important factors like energy generation methods which also affect air pollution. The research confirms that Air Quality Index (AQI) scores and Alternative Fuel Vehicles (AFV) counts are positively correlated. However, the research is limited to counties in New Jersey while in our approach we have taken into account data from 49 states in the US and 5 states in India.

Another study by Holland, S. P.; Mansur, E. T.; Muller, N. Z. and Yates, A. J., 2019, [2] explores the consequences of EV adoption for air quality and subsequent distributional effects. These studies [1][2] contribute valuable insights into the interactions between vehicle adoption and air quality, but they do not comprehensively consider factors such as energy production and emission.

Furthermore, Miconi F, Dimitri GM (2023) [3] explore machine learning techniques to analyze and estimate factors which impact the distribution of EVs within Italy. While concentrating on a specific geographic context, this study's data-driven approach aligns with our methodology.

In the context of the United States and the 5 states in India, our study extends the analysis by examining the relationship between EV adoption and NO2 emissions at the state level. Our dataset includes the number of petrol, diesel, CNG, and EVs, alongside essential socioeconomic factors such as GDP, electricity production, power sector emissions, and the land area of the state. By employing machine learning techniques, including linear regression, Ridge regression, Decision Tree Regressor, Random Forest Regressor, XGB Regressor, and Multi-Layer Perceptron, we aim to model and predict the potential reduction in NO2 emissions associated with increased EV adoption.

DATASET

Construction of dataset was done in 3 steps:

1. **Identification of Variables:**
   Initially, we looked at the factors that could impact air quality concerning EVs. We narrowed it down to socioeconomic factors like GDP, energy production and emissions.
2. **Data Collection:**
   Once we had determined the specific features, we proceeded to collect data related to them. We obtained data for 49 U.S. states and 5 states in India. More details on this are provided in the following section.
3. **Data Compilation:**
   Finally, after obtaining data for all the variables, we merged them into a single dataset. During this process, we also conducted various data transformations and corrections, because the data were in different formats.

AIR QUALITY DATA

For US states, the “Annual mean NO2 data” and SO2 was picked for every state from the EPA site [4]. Average was calculated for all counties in a state for years 2016 to 2021. For India, we shortlisted 9 states based on data availability of EV sales on Vahan site [6] for the past few years. The vehicle category we looked at was 2, 3 and 4 wheelers. 5 states namely Delhi, Haryana, Uttar Pradesh, Rajasthan and Karnataka were then selected for which NO2 data could be scrapped from CPCB site. Annual mean of NO2 , SO2 data was webscrapped from the data provided by Central Pollution Control Board (CPCB) at the central control room site for air quality management [5] for years 2019 to 2022.

NUMBER OF VEHICLES

Data was collected for the number of registered vehicles in a state. The vehicle categories chosen were petrol, diesel, CNG and Electric. For India, the data was scraped from Vahan site [6]. For the US, the registration data was scraped from Alternatives Fuel Data Centre [7].

ELECTRICITY PRODUCTION AND POWER SECTOR EMISSION

Data was collected from the IEA site for all US and India states [8]. This included production and emission data from following sources - bioenergy, coal, gas, hydro, nuclear, solar, other renewable, other fossil sources. With the increase in the number of electric cars, there will be more demand for electricity. Hence the generation data is an important factor to consider.

OTHER DATA

Since the parameters are for various states which would vary in sizes and socio-economic status, data was also collected for population size, area of the state and GDP of the state. Later GDP was converted to GDP per capita and the population parameter was dropped.

In the final data set there were in total 314 data records. These data records were for 5 Indian states from 2019 to 2022 and for 49 US states from 2016 to 2021. Table 1 lists all the features and their type in the final data set.

Category	Feature	Type	Unit
Air Quality	NO2	float	ppb
Socio-economic	GDP per capita	int	USD
Registered Vehicles	Petrol	int	Number
	Diesel	int	Number
	CNG	int	Number
	Electric	int	Number
Demographics	Male	int	Number
	Female	int	Number
	Area of state	float	km2
Electricity generation Data from various sources	Bioenergy	float	TWh
	Coal	float	TWh
	Gas	float	TWh
	Hydro	float	TWh
	Nuclear	float	TWh
	Other Fossil	float	TWh
	Other Renewables	float	TWh
	Solar	float	TWh
	Wind	float	TWh
Power Emission Data from various sources	Bioenergy	float	mtCO2
	Coal	float	mtCO2
	Gas	float	mtCO2
	Hydro	float	mtCO2
	Nuclear	float	mtCO2
	Other Fossil	float	mtCO2
	Other Renewables	float	mtCO2
	Solar	float	mtCO2
	Wind	float	mtCO2

Table 1: Input parameters

DATA PRE-PROCESSING

When looking at the correlation matrix between above parameters, it was found that electricity production and Power emission parameters had correlation values as high as 1 between them. Hence, we kept only one of the factors. Data was scaled using Min-Max scaler. The final correlation coefficients table has been added in Appendix.

METHODOLOGY/MODELS

We tried the following well-established machine learning algorithms for supervised learning in our research

- Ridge Regressor (RR)
- Decision Tree Regressor (DT)
- Linear Regression (LR)
- Random Forest Regressor (RF)
- XGB Regressor (XGB)
- Neural Network (NN)

RIDGE REGRESSION

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated. In general, the method provides improved efficiency in parameter estimation problems in exchange for a tolerable amount of bias (see bias–variance tradeoff) . [16]

DECISION TREE REGRESSION

Decision Tree algorithm uses a tree-like model of decisions to either predict the target value (regression) or predict the target class (classification). Decision trees where the target variable or the terminal node can take continuous values (typically real numbers) are called regression trees [17]. We will be using the regression tree in our study.

Creating a regression tree relies on binary recursive partitioning, an iterative technique that divides data into segments. Precisely, commencing from the root node, the algorithm dissects data via all feasible binary divisions, opting for the optimal partition. The partition's effectiveness is gauged by variance reduction, hence the finest division minimizes variance. This segmentation approach extends to every tree node, persisting until each achieves a predefined minimum observation count, transitioning into a leaf node.[17]

LINEAR REGRESSION

In this study, we focused on Multivariate Linear Regression. Multiple linear regression is a generalization of simple linear regression to the case of more than one independent variable, and a special case of general linear models, restricted to one dependent variable [22]. The basic model for multiple linear regression is :

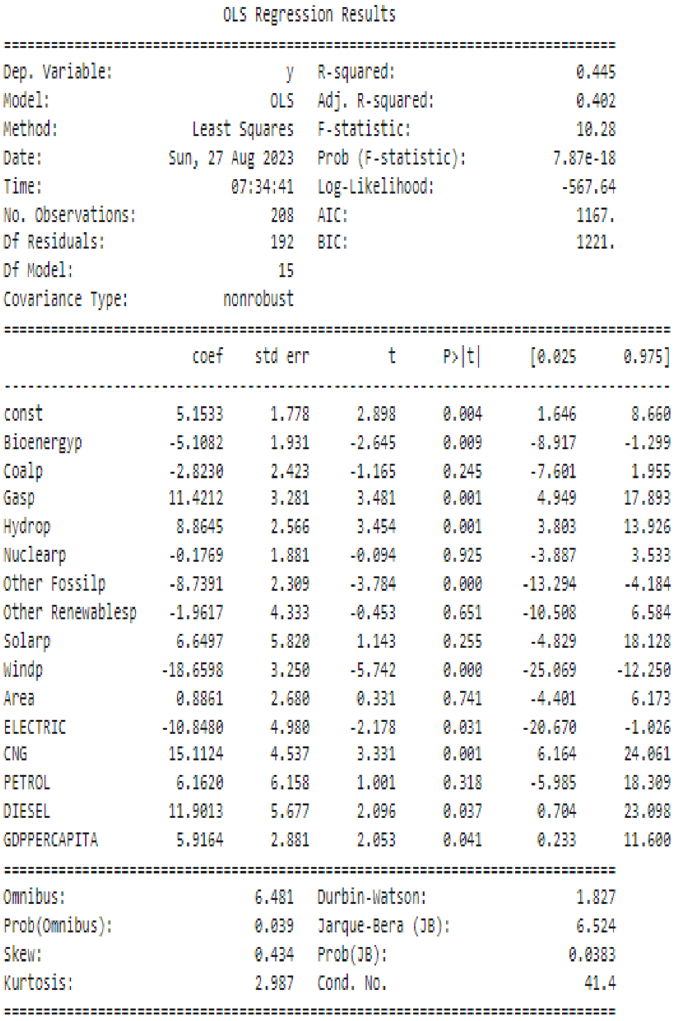
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

In the formula above we consider n observations of one dependent variable and p independent variables. Thus, Yi is the i^{th} observation of the dependent variable, X_{ij} is i^{th} observation of the j^{th} independent variable, j = 1, 2, ..., p. The values βj represent parameters to be estimated, and ε_i is the i^{th} independent identically distributed normal error. [22]

The null hypothesis is the claim that no relationship exists between two sets of data or variables being analyzed. The null hypothesis is that any experimentally observed difference is due to chance alone, and an underlying causative relationship does not exist, hence the term "null". In addition to the null hypothesis, an alternative hypothesis is also developed, which claims that a relationship does exist between two variables. [9]

p-value is the probability that an observed difference is due to random chance when the null hypothesis is true[26]. A common threshold of the P-value is 0.05. A P-value of 0.05 means that 5% of the time, we will falsely reject the null hypothesis. If the P-value is lower than 0.05, we can reject the null hypothesis with 95% confidence and conclude that there exists a relationship between the variables. [10]

To test if the coefficients from the LR function has a significant impact on the dependent variable, we used the NULL value hypothesis. Ordinary least-squares (OLS) models assume that the analysis is fitting a model of a relationship between one or more explanatory variables and a continuous or at least interval outcome variable that minimizes the sum of square errors, where an error is the difference between the actual and the predicted value of the outcome variable [22]. We looked at the p values, coefficients [Figure 1] returned by running the OLS model and the NULL value hypothesis confirmed our initial belief that the Number of Electric cars does affect the dependent variable “NO2”. p-value for ELECTRIC is .031 which < .05 confirming our belief that the “Number of Electric cars” has a significant relation with dependent variable NO2. The coefficient being negative for Electric cars also indicates a negative relationship between the two. This indicates that as the number of EVs increase, we can expect the NO2 levels to decrease.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.445
Model:                            OLS   Adj. R-squared:                  0.402
Method:                 Least Squares   F-statistic:                     10.28
Date:                Sun, 27 Aug 2023   Prob (F-statistic):           7.87e-18
Time:                        07:34:41   Log-Likelihood:                -567.64
No. Observations:                 208   AIC:                             1167.
Df Residuals:                     192   BIC:                             1221.
Df Model:                          15
Covariance Type:            nonrobust
====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const                 5.1533      1.778      2.898      0.004       1.646       8.660
Bioenergyp           -5.1082      1.931     -2.645      0.009      -8.917      -1.299
Coalp                -2.8230      2.423     -1.165      0.245      -7.601       1.955
Gasp                 11.4212      3.281      3.481      0.001       4.949      17.893
Hydrop                8.8645      2.566      3.454      0.001       3.803      13.926
Nuclearp             -0.1769      1.881     -0.094      0.925      -3.887       3.533
Other Fossilp        -8.7391      2.309     -3.784      0.000     -13.294      -4.184
Other Renewablesp    -1.9617      4.333     -0.453      0.651     -10.508       6.584
Solarp                6.6497      5.820      1.143      0.255      -4.829      18.128
Windp               -18.6598      3.250     -5.742      0.000     -25.069     -12.250
Area                  0.8861      2.680      0.331      0.741      -4.401       6.173
ELECTRIC            -10.8480      4.980     -2.178      0.031     -20.670      -1.026
CNG                  15.1124      4.537      3.331      0.001       6.164      24.061
PETROL                6.1620      6.158      1.001      0.318      -5.985      18.309
DIESEL               11.9013      5.677      2.096      0.037       0.704      23.098
GDPPERCAPITA          5.9164      2.881      2.053      0.041       0.233      11.600
==============================================================================
Omnibus:                        6.481   Durbin-Watson:                   1.827
Prob(Omnibus):                  0.039   Jarque-Bera (JB):                6.524
Skew:                           0.434   Prob(JB):                       0.0383
Kurtosis:                       2.987   Cond. No.                         41.4
==============================================================================
```

Figure 1: OLS Regression Results

RANDOM FOREST REGRESSION

Random forest is a type of supervised learning algorithm that uses ensemble methods (bagging) to solve both regression and classification problems. The algorithm operates by constructing a multitude of decision trees at training time and outputting the mean/mode of prediction of the individual trees.[18]

XGB REGRESSION

The Extreme Gradient Boosting (XGBoost) Regressor is a powerful machine learning algorithm. It's an ensemble method that combines multiple weak regression models to create a robust and accurate predictor. XGBoost iteratively refines predictions, focusing on data points with higher residuals. This boosts predictive accuracy and reduces overfitting. It handles missing values and can handle both linear and nonlinear relationships in data. XGBoost's effectiveness lies in its ability to capture complex patterns, making it a popular choice for regression tasks, especially in cases where high predictive performance is desired.[19]

NEURAL NETWORK - MULTILAYER PERCEPTRON

The Multi-Layer Perceptron (MLP) is a neural network architecture used in machine learning. It consists of interconnected layers of nodes, including an input layer, one or more hidden layers, and an output layer. Each node applies a nonlinear activation function to its input, allowing the network to model complex relationships in data. MLPs can handle various data types and are effective in tasks like classification and regression. Training involves adjusting weights through backpropagation, optimizing the network's ability to learn patterns and make predictions. MLPs are versatile and widely applied in deep learning for their capacity to handle intricate data relationships. [20]

In this study, we used a neural network with 5 hidden layers (16,14,12,10,8), learning rate of .01 and Adam optimizer. The model was developed using Keras Regressor library and parameters were tuned using GridSearchCV.

MODEL METRICS

We used the following 4 metrics to evaluate the model's performance: coefficient of determination (R2), mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE).

The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.[21]

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}\,|$$

where

$\hat{y} - predicted\ value\ of\ y$

$\overline{y} - mean\ value\ of\ y$

Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals.[21]

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\,(y_i - \hat{y})^2$$

Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals.[21]

$$RMSE = \sqrt{MSE}$$

The coefficient of determination or R-squared represents the proportion of the variance in the dependent variable which is explained by the linear regression model. It is a scale-free score i.e. irrespective of the values being small or large, the value of R square will be less than one.[21]

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \overline{y}\,)^2}$$

For the test train split, we used a 80:20 split with random seed and Shuffle selected as TRUE. K-fold cross-validation was used for evaluating the predictive models with k set as 5. The dataset was divided into k=5 subsets or folds. The model was trained and evaluated k times, using a different fold as the validation set each time. Performance metrics from each fold were averaged to estimate the model's generalization performance. Neural network is a non-deterministic model which gives different results with every run. So with every k-split, the model was run 5 times and average was taken over all the runs to take care of the non-deterministic behavior.

RESULTS AND DISCUSSIONS

In this section, we analyze the outcomes of our predictive models and explore the impact of EVs on NO2 levels. Our investigation involved multiple regression techniques, including Linear Regression (LR), Ridge Regression (RR), Decision Tree (DT), Random Forest (RF), Extreme Gradient Booster (XGB), and Neural Network (NN).

Table below provides details of various metrics measured with the predictive models.

Metrics	RR	DT	LR	RF	XGB	NN
RMSE	3.87	3.239	3.87	3.73	2.210	1.802
MSE	15.03	10.88	15.1	13.97	4.912	3.360
MAE	3.008	1.922	3.03	2.935	1.60	1.333
R2	0.415	0.584	0.41	0.467	0.811	0.823

Table 2: Model metrics

We set a criterion of excluding models with R2 values below 50% from further consideration. Out of the two remaining models XGBR and Neural Network, we chose to employ the Neural Network model for subsequent simulations due to its high R2 value and lower RMSE error compared to the XGBR.

Visualizing the actual versus estimated scores on the test data for each model through graphs provided a visual view of the errors. However, the inclusion of data from the year 2020 in our prediction model led to outliers in these plots. The nationwide lockdown in various countries during 2020, due to the outbreak of COVID-19, reduced traffic and industrial activities, leading to a significant reduction of NO2 as reported by Cooper, M.J., Martin, R.V., Hammer, M.S. et al.[25]. This is not representative of normality. Hence the data of that year was removed. The graphs for actual vs estimates scores are in Figure 2 and 3:

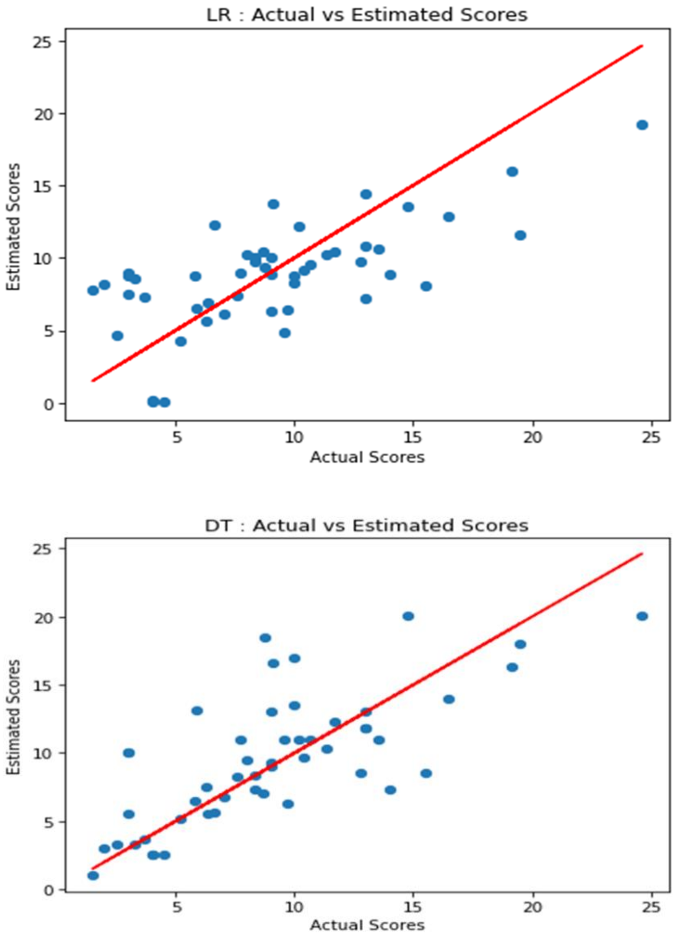


Figure 2: Estimated vs Actual NO2 values for various models (LR, DT)

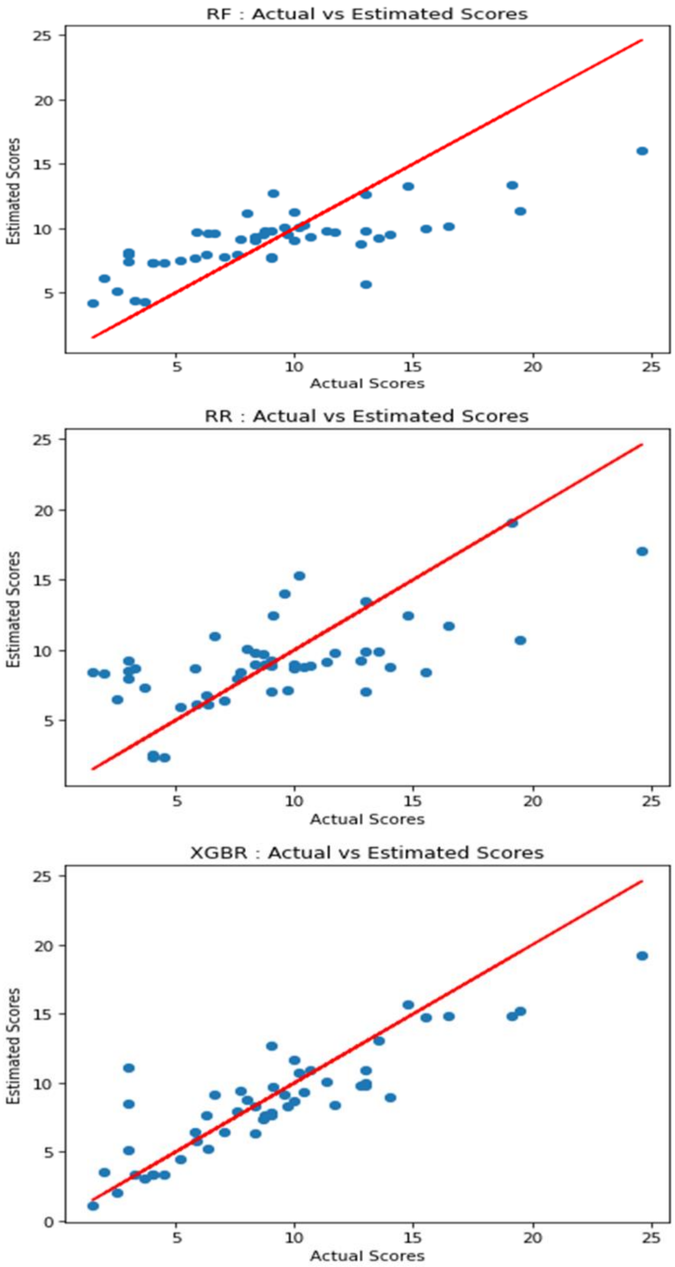


Figure 3: Estimated vs Actual NO2 values for various models (RF, RR, XGBR)

IMPACT OF EVS ON NO2

Our subsequent analysis aimed to assess the influence of increasing EV adoption on NO2 concentrations. This investigation involved simulations where EV quantities were scaled from 0 to 1.5 million, while holding other variables constant. We chose 1.5 million as the upper limit as the maximum EV numbers we have in our training dataset was 0.5 million. We have scaled numbers to 3 times that value as estimates of NO2 level with EV numbers much higher than in the training data set would make NO2 predictions much less reliable. The resulting trends demonstrated a consistent decrease in NO2 levels with the rise of EV numbers.

Interestingly, some states like Texas, Pennsylvania, Rajasthan exhibited an initial spike in NO2 levels followed by a subsequent decline upon increasing the number of EVs. This phenomenon might be attributed to the interplay between EV adoption and electricity production. As the demand for electricity increases alongside EVs, reliance on fossil fuels for energy production may temporarily elevate NO2 levels. However, this effect is outweighed by the long-term benefits of EVs in mitigating NO2 pollution.

The graphs below are the simulation results with x axis denoting the EV's (0 to 1.5 million) and y axis denoting the predicted NO2 level. For each year, a different color is used for showing the predicted value of NO2 in a particular year. Mean of the predicted NO2 value over years is also calculated and plotted as the red line.

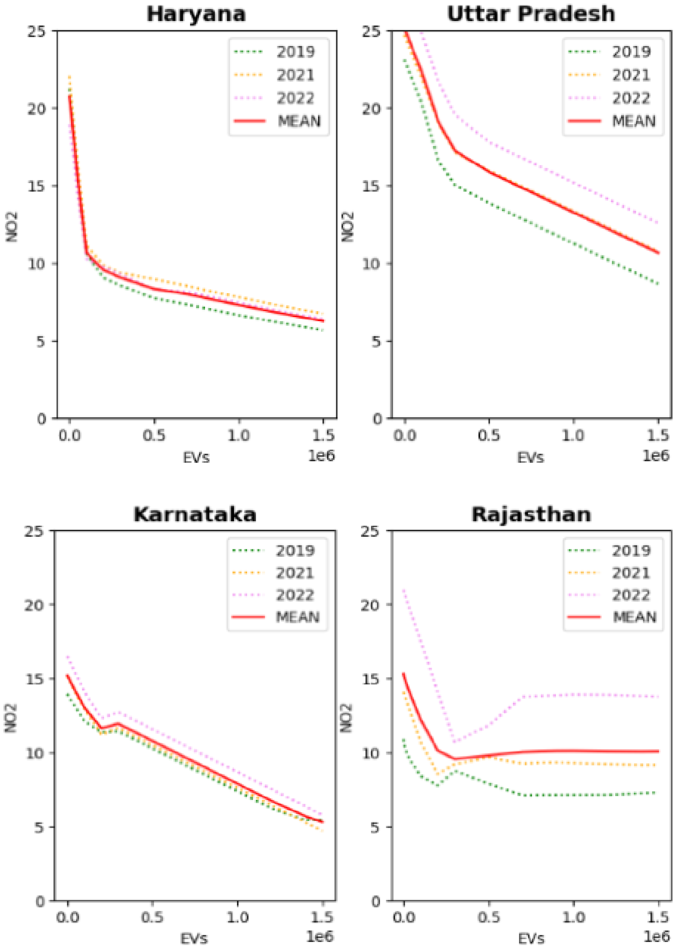


Figure 4: Simulations Scaled EV's vs NO2 (Indian States)

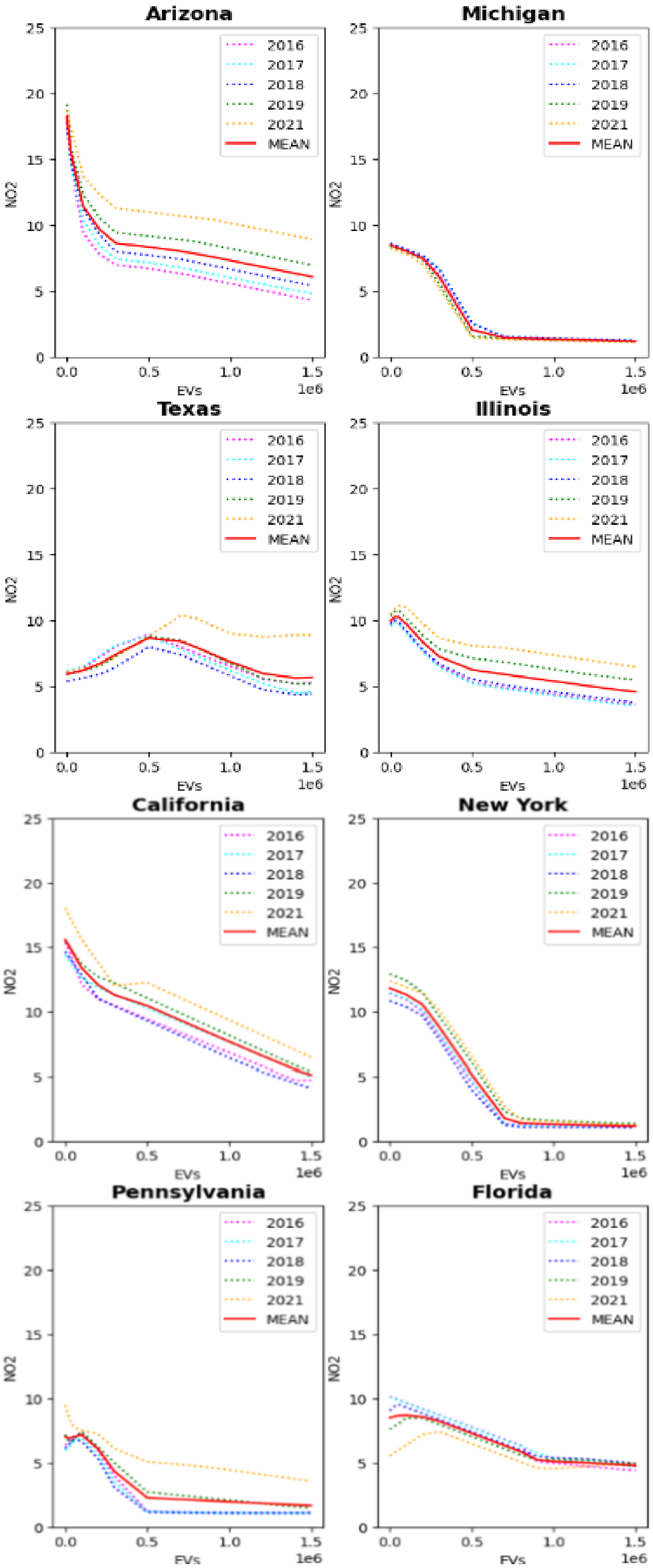


Figure 5: Simulations Scaled EV's vs NO2 (US States)

DELHI AS A CASE STUDY

Focusing on Delhi, India, we conducted similar simulations by increasing the number of EVs while keeping the other factors constant. Our model predicts a visible decrease in NO2 level as the number of EVs would be scaled to 1.5 million. The dots on the graph indicate the actual values for NO2 for those years which are very close to the predicted values shown on the chart.

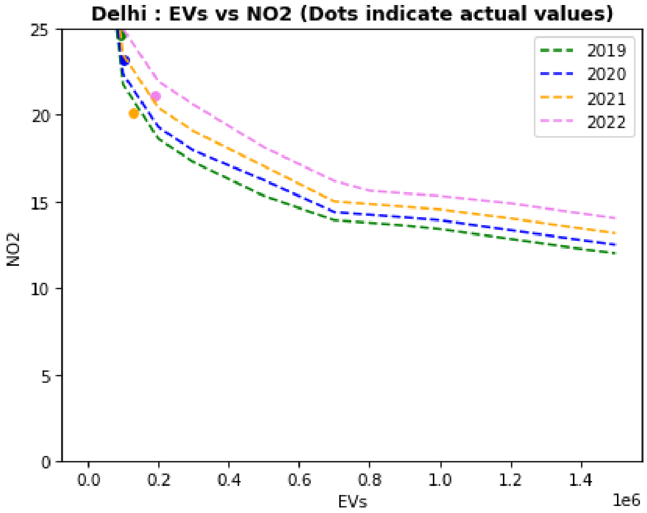


Figure 6: Delhi simulation run with scaled EV's

It is also observed that NO2 levels have increased over the years. This might be attributed to the fact that along with EVs, the number of Diesel and petrol vehicles are also increasing. So we decided to run more simulations to observe how replacement of diesel vehicles with EVs affect the NO2 levels. We ran these simulations for the year 2022. In this scenario, we reduced the number of diesel vehicles by 10% and redistributed this portion as EVs while keeping the total vehicle count constant. The results of this simulation indicated a promising trend—NO2 levels decreased as the share of EVs increased, in line with our broader findings.

These trends align well with Delhi's policy initiatives aimed at combating pollution. The regulation to phase out diesel vehicles older than a decade and the proposal to restrict diesel passenger vehicles in Indian cities by 2027, has been advocated by the Ministry of Petroleum and Natural Gas [24]. Our model predicts that if the policy has the intended effect of reducing the number of diesel vehicles and increasing the number of EVs, NO2 levels shall decrease.

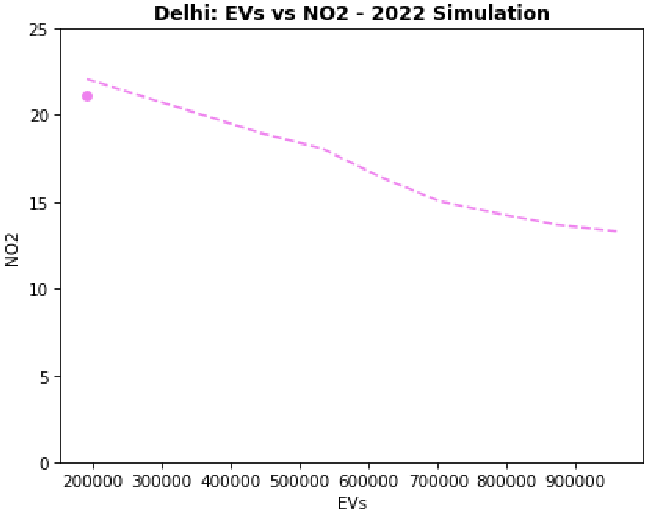


Figure 7: Delhi simulation with varying diesel and EV's

CONCLUSION

In conclusion, our study highlights the potential of EVs in reducing NO2 levels. Through predictive modeling and simulations, we underscore the importance of transitioning toward cleaner transportation alternatives to mitigate air pollution. We experimented with a multitude of regression techniques and found that the neural network performs best for our given data set with a R2 of 0.82. The observed trends, particularly evident in Delhi, offer actionable insights for policy makers and urban planners seeking effective strategies to improve air quality and public health.

By systematically analyzing the outcomes of our models and simulations, our study contributes to the understanding of the complex relationship between transportation choices and air quality, shedding light on the path towards sustainable urban development. However, shifting to EV's alone won't solve the air quality issues alone. A switch to public transportation and overall reduction in the use of personal/low-occupancy vehicles is necessary too. Another important factor to consider is to shift to use of renewable sources for production of electricity for the EV's.

Building upon this study's findings and considering the potential advancements highlighted by recent research developments, several promising avenues for future investigation emerge. The following areas present opportunities to enhance the understanding of the relationship between EV adoption and nitrogen dioxide

(NO2) levels, with a particular focus on India's unique context.

1. Advanced Modeling Techniques

While the current study leveraged established machine learning algorithms, further exploration into more advanced modeling techniques could yield richer insights. Incorporating time series analysis, for instance, could provide a dynamic understanding of the evolving relationship between EV adoption and NO2 levels over different time intervals. Time series models such as ARIMA and Prophet, demonstrated in [23], could offer a comprehensive view of temporal trends and patterns.

2. Expansion of Dataset and Features

Enhancing the dataset's scope by including a broader range of countries, especially developing nations that share similarities with India's environmental challenges, can yield more robust insights. Introducing features that characterize the energy grid's composition, the availability and planned capacity of EV charging infrastructure, and socio-economic factors specific to each region can enhance the predictive power of the models. This approach aligns with the methodology proposed in [1], which incorporates socioeconomic status.

3. Improved Data Granularity and Resolution

Accessing more granular EV data, such as precise counts of EVs by type and location, could augment the predictive accuracy of models. Additionally, adopting higher temporal resolution—transitioning from yearly to monthly data—can provide a more detailed representation of EV adoption dynamics and their impact on NO2 levels. This aligns with the approach used in [3] where a machine learning approach is employed to analyze and predict EV scenarios.

4. Inclusion of Additional Environmental Factors

Broadening the scope of the analysis to incorporate a wider range of environmental factors—such as sulfur dioxide (SO2) concentrations—can enhance the holistic representation of air quality. This more comprehensive approach resonates with studies like [2], which explores the distributional effects of air pollution from EV adoption. Including multiple pollutants and their interactions can lead to a more nuanced understanding of EVs' impact on overall environmental quality.

In summary, the next steps for research involve a multifaceted approach, encompassing advanced modeling techniques, expanded datasets, finer data granularity, and the inclusion of broader environmental factors. By pursuing these avenues, researchers can attain a more nuanced and comprehensive understanding of the potential of EVs to influence NO2 levels and contribute to India's air quality improvement.

ACKNOWLEDGEMENTS

My sincere thanks to Inspirit AI for providing me with the opportunity to do this research and Justas Janonis for teaching, supporting, and helping me along the way.

REFERENCES

[1] Singh, Anuradha & Yadav, Jyoti & Shrestha, Sarahana & Varde, Aparna. (2023). Linking Alternative Fuel Vehicles Adoption with Socioeconomic Status and Air Quality Index. [https://doi.org/10.48550/arXiv.2303.08286]

[2] Holland, S. P.; Mansur, E. T.; Muller, N. Z. and Yates, A. J., 2019, 'Distributional Effects of Air Pollution from Electric Vehicle Adoption,' Journal of the Association of Environmental and Resource Economists, 6, pp. S65–S94. https://doi.org/10.1086/701188.

[3] Miconi F, Dimitri GM (2023) A machine learning approach to analyse and predict the electric cars scenario: The Italian case. PLoS ONE 18(1): e0279040. https://doi.org/10.1371/journal.pone.0279040

[4] https://www.epa.gov/outdoor-air-quality-data/air-quality-statistics-report

[5] https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/data

[6]https://vahan.parivahan.gov.in/vahan4dashboard/vahan/view/reportview.xhtml

[7] https://afdc.energy.gov/vehicle-registration?.

[8] https://ember-climate.org/data/data-catalogue/?topic=electricity

[9] https://en.wikipedia.org/wiki/Statistical_hypothesis_testing

[10]https://www.w3schools.com/datascience/ds_linear_regression_pvalue.asp

[11] U.S. Environmental Protection Agency (EPA) (2018). "Review of the Primary National Ambient Air Quality Standards for Oxides of Nitrogen," 40 CFR Part 50, Federal Register. 83, No. 75/ Wednesday, April 18, 2018. Available at: https://www.govinfo.gov/content/pkg/FR-2018-04-18/pdf/2018-07741.pdf

[12] U.S. Environmental Protection Agency (EPA) (2016). Integrated Science Assessment for Oxides of Nitrogen— Health Criteria (2016 Final Report). Research Triangle Park, NC: U.S. EPA, National Center for Environmental Assessment.

EPA/600/R–15/068. Available at: https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=310879

[13] IEA (2021), Global EV Outlook 2021, IEA, Paris https://www.iea.org/reports/global-ev-outlook-2021, License: CC BY 4.0

[14] Delhi Electric Vehicles (EV) policy (2020) (https://ev.delhi.gov.in/files/Accelerating-Electric-Mobility-in-Delhi8497bf.pdf

[15] E-AMRIT (2019) (https://e-amrit.niti.gov.in/assets/admin/dist/img/new-fronend-img/report-pdf/rmi-niti-ev-report.pdf)

[16] Wikipedia. 2023. “Ridge Regression” Last Modified 29 August 2023, at 16:11 (UTC). https://en.wikipedia.org/wiki/Ridge_regression

[17] Wikipedia. 2023. “Decision tree learning” Last Modified 4 September 2023, at 18:39 . https://en.wikipedia.org/wiki/Decision_tree_learning

[18]Wikipedia. 2023. “Random Forest” Last Modified 9 September 2023, at 09:27 (UTC).https://en.wikipedia.org/wiki/Random_forest

[19] https://www.geeksforgeeks.org/xgboost/

[20] https://www.sciencedirect.com/topics/computer-science/multilayer-perceptron

[21] https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

[22] https://link.springer.com/referenceworkentry/10.1007/978-94-007-0753-5_2008#:~:text=Ordinary%20least%2Dsquares%20(OLS)%20models%20assume%20that%20the%20analyst,the%20predicted%20value%20of%20the

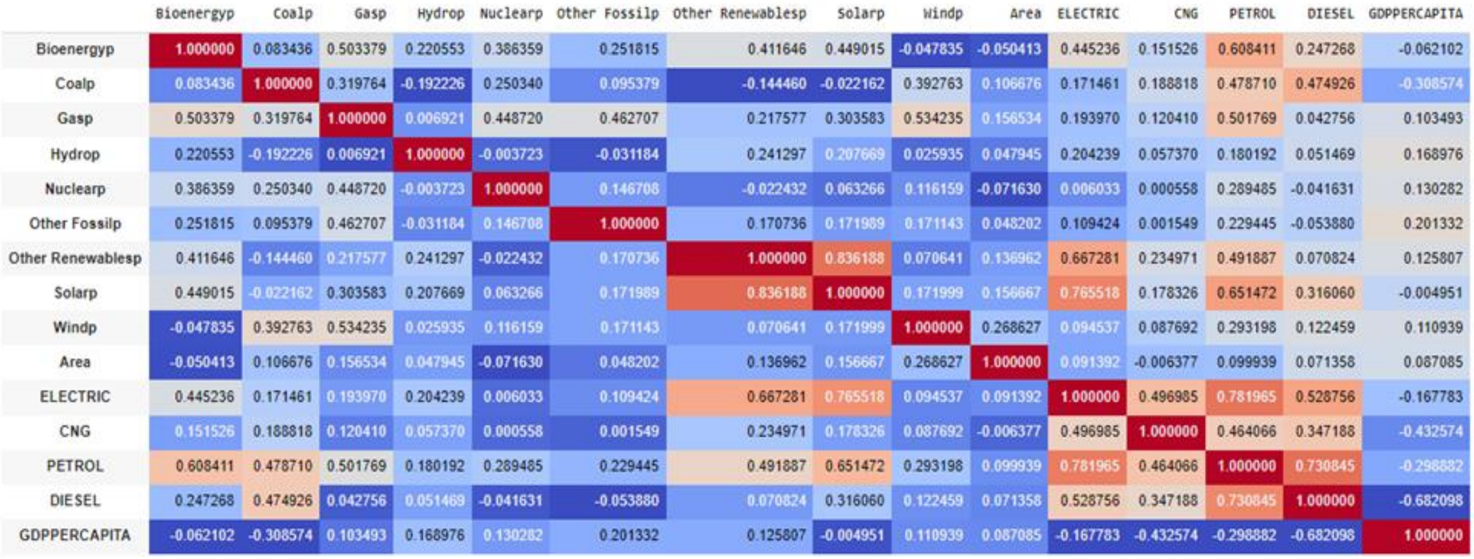
[23] Hasnain Ahmad, Sheng Yehua, Hashmi Muhammad Zaffar, Bhatti Uzair Aslam, Hussain Aamir, Hameed Mazhar, Marjan Shah, Bazai Sibghat Ullah, Hossain Mohammad Amzad, Sahabuddin Md, Wagan Raja Asif, Zha Yong. (2022). Time Series Analysis and Forecasting of Air Pollutants Based on Prophet Forecasting Model in Jiangsu Province, China [https://www.frontiersin.org/articles/10.3389/fenvs.2022.945628] [DOI=10.3389/fenvs.2022.945628]

[24]https://mopng.gov.in/files/uploads/ETAC_2023_FINAL_PRINT.pdf

[25] Cooper, M.J., Martin, R.V., Hammer, M.S. et al. Global fine-scale changes in ambient NO2 during COVID-19 lockdowns. Nature 601, 380–387 (2022). https://doi.org/10.1038/s41586-021-04229-0

[26] Flechner L, Tseng TY. Understanding results: P-values, confidence intervals, and number need to treat. Indian J Urol. 2011 Oct;27(4):532-5. doi: 10.4103/0970-1591.91447. PMID: 22279324; PMCID: PMC3263226.

APPENDIX



	Bioenergyp	Coalp	Gasp	Hydrop	Nuclearp	Other Fossilp	Other Renewablesp	Solarp	Windp	Area	ELECTRIC	CNG	PETROL	DIESEL	GDPPERCAPITA
Bioenergyp	1.000000	0.083436	0.503379	0.220553	0.386359	0.251815	0.411646	0.449015	-0.047835	-0.050413	0.445236	0.151526	0.608411	0.247268	-0.062102
Coalp	0.083436	1.000000	0.319764	-0.192226	0.250340	0.095379	-0.144460	-0.022162	0.392763	0.106676	0.171461	0.188818	0.478710	0.474926	-0.308574
Gasp	0.503379	0.319764	1.000000	0.006921	0.448720	0.462707	0.217577	0.303583	0.534235	0.156534	0.193970	0.120410	0.501769	0.042756	0.103493
Hydrop	0.220553	-0.192226	0.006921	1.000000	-0.003723	-0.031184	0.241297	0.207669	0.025935	0.047945	0.204239	0.057370	0.180192	0.051469	0.168976
Nuclearp	0.386359	0.250340	0.448720	-0.003723	1.000000	0.146708	-0.022432	0.063266	0.116159	-0.071630	0.006033	0.000558	0.289485	-0.041631	0.130282
Other Fossilp	0.251815	0.095379	0.462707	-0.031184	0.146708	1.000000	0.170736	0.171989	0.171143	0.048202	0.109424	0.001549	0.229445	-0.053880	0.201332
Other Renewablesp	0.411646	-0.144460	0.217577	0.241297	-0.022432	0.170736	1.000000	0.836188	0.070641	0.136962	0.667281	0.234971	0.491887	0.070824	0.125807
Solarp	0.449015	-0.022162	0.303583	0.207669	0.063266	0.171989	0.836188	1.000000	0.171999	0.156667	0.765518	0.178326	0.651472	0.316060	-0.004951
Windp	-0.047835	0.392763	0.534235	0.025935	0.116159	0.171143	0.070641	0.171999	1.000000	0.268627	0.094537	0.087692	0.293198	0.122459	0.110939
Area	-0.050413	0.106676	0.156534	0.047945	-0.071630	0.048202	0.136962	0.156667	0.268627	1.000000	0.091392	-0.006377	0.099939	0.071358	0.087085
ELECTRIC	0.445236	0.171461	0.193970	0.204239	0.006033	0.109424	0.667281	0.765518	0.094537	0.091392	1.000000	0.496985	0.781965	0.528756	-0.167783
CNG	0.151526	0.188818	0.120410	0.057370	0.000558	0.001549	0.234971	0.178326	0.087692	-0.006377	0.496985	1.000000	0.464066	0.347188	-0.432574
PETROL	0.608411	0.478710	0.501769	0.180192	0.289485	0.229445	0.491887	0.651472	0.293198	0.099939	0.781965	0.464066	1.000000	0.730845	-0.298882
DIESEL	0.247268	0.474926	0.042756	0.051469	-0.041631	-0.053880	0.070824	0.316060	0.122459	0.071358	0.528756	0.347188	0.730845	1.000000	-0.682098
GDPPERCAPITA	-0.062102	-0.308574	0.103493	0.168976	0.130282	0.201332	0.125807	-0.004951	0.110939	0.087085	-0.167783	-0.432574	-0.298882	-0.682098	1.000000

Figure 8: Correlation coefficient