STEM Fellowship Big Data Challenge 2024

Analyzing The Correlation Between Biodiversity Indexes, Carbon Emissions, and Government Investment Into Public Transport to Identify Feasible Pathways to Improving Biodiversity in the Future

Ashank Chaturvedi, Atharva Rao, Joel Zou, Owen Kuang, and Yunhe Zhang

Webber Academy

January 14, 2024

Word Count: 2995

1 Abstract

In the world of increasing CO2 emissions, where transportation alone contributes to 28% of this environmental burden, its impact on climate has been a widely discussed topic. The main impact is global warming, which can change biodiversity drastically by disrupting ecosystems and changing living conditions for organisms. With this in mind, our study aims to investigate and establish the correlation between CO2 and biodiversity to evaluate periods of increased CO2 emission to biodiversity change in particular ecosystems in North America. Our second objective of this study is to seek out potential financial solutions governments could implement to mitigate CO2 emissions' negative impacts on biodiversity. To identify relationships among the data we used, we used a fine-tuned chatbot we made on Open AI's token-based platform and Python to sort multiple datasets, including biodiversity data from the Living Planet Index Organization and annual CO2 emissions data from Our World Data. To take this further, we continued to use Python to create graphs to visualize relationships and correlations between the data used. We hypothesize that there is an inverse relationship between biodiversity and carbon emissions. However, the effect of carbon emissions on biodiversity can be significantly decreased by putting into practice strong environmental solutions by receiving financial backing from governments. We identified a strong negative correlation between CO2 emissions and biodiversity and a moderate positive correlation between investment in public transport and CO2 levels.

Keywords
Climate Change, Biodiversity, Biology, Data Science, Government Policy

2 Introduction

Ever since the advent of the Industrial Revolution, the use of fossil fuels such as coal and oil has exponentially increased, and with that, carbon dioxide emissions into the atmosphere. There now exists issues with rising CO2 levels and the warming of the Earth, the phenomenon now widely known as climate change [1] [2]. Rising carbon emissions have left many negative impacts on Earth. Many of these negative impacts revolve around biodiversity. Firstly, climate change is expected to cause the global temperature to increase by more than 1.5°C within the next 20 years. This drastic change will force many organisms to either rapidly adapt or go extinct, which is extremely damaging to biodiversity worldwide. Climate change also leads to more invasive species harming ecosystems, as many species that typically survive in hotter climates can now survive in more northern regions. Invasive species wreak havoc upon ecosystems, by disrupting the food chain and natural flow of energy and matter [3]. Governments have only begun implementing solutions recently, as climate activism has grown stronger. Some solutions include carbon taxes, cap and trade, international agreements, clean energy standards, and investments. Carbon taxes, which function by having companies pay for their emissions, have worked well in Sweden, where they have reduced emissions by 27% while doubling their GDP [4]. While these outcomes are a step in

the right direction, CO2 levels are still drastically rising, so even more action is required to combat this issue. In this study, we will first confirm the correlation between carbon emissions and biodiversity. We will accomplish this by analyzing large data sets to identify trends where a time period of higher carbon emissions corresponded to a greater degree of biodiversity loss. Secondly, we will focus on transportation, as over 28% of carbon emissions are produced in these processes [5]. Lastly, we would like to use robust data to seek realistic financial resolutions which governments could implement to create real change in terms of carbon emissions. The main question we want to answer in this study is to what extent is government investment into public transportation a feasible path for improving biodiversity? We hypothesize that there is a negative correlation between carbon emissions and biodiversity. However, through the implementation of compelling environmental solutions verified by this study and financial support from governments, the impact of carbon emissions on biodiversity can potentially be reduced.

3 Materials & Methods

3.1 Data Selection

Prior to selecting our data, our group fine-tuned a chatbot for the purpose of being a research assistant. The ChatGPT API was utilized, and the chatbot was built through Python using the OpenAI library and Gradio to host it for our group to access it through any device. This chatbot was essentially a contextualized version of GPT-3.5 that provided us with notions on how to develop the project further and seek potential correlations. Our group also utilized the ChatGPT token-based platform to create an AI assistant that was capable of being trained on files uploaded to it. Rather than going through websites and testing data files to see if they fit the project, we uploaded documentation for the files instead to see if it was usable. We picked our data based on three main criteria: data diversity, scope of data, and dataset synergy with other datasets. We can use this chatbot to review the documentation of these data sets and give us recommendations based on these criteria. By using this chatbot that we fine-tuned, we can go through many more data sets before determining which ones we want. Ultimately, we found three data sets that fit all three criteria. The first file is from the Living Planet Index Organization, which measures biodiversity from 1970 to 2018. The second file is from Our World In Data, which is about annual CO2 emissions globally measured in kilotons [6][7]. Our last file is from the Bureau of Transportation Statistics (BTS), which uncovers American government spending on various public investments such as transportation [8].

3.2 Data Cleaning

To clean our data, we used Python as our language of choice, along with multiple libraries such as Pandas, Numpy, CSV, and Matplot. Additionally, our machine learning included EvalML, PygWalker, AutoClean, and Streamlit, which streamlined the process. Visual Studio was our IDE of choice throughout the entire project. To clean our data, we first had to merge the biodiversity and carbon emission data into one file. To do this, the Pandas library was used to ensure all chart variables were standardized in terms of naming and timeline before merging. We had to trim the timeline to 1970-2018 so that a variable on one of the files would always have a corresponding variable in the other file. Additionally, we believe that 1970 was a good cut-off as it was just a decade before when global warming became a concern to world governments. During the merging of the files, every year's biodiversity index and corresponding carbon emission were paired up. To better visualize the data, the y-intercept was scaled by 100. It is important to note that scaling does not affect r-values or any other methods of analyzing data this paper will be using. Because the scope of our research was in North America, we also took out locations that are not in North America in the merged CSV file. Additionally, we took the data from the Bureau of Transportation Statistics regarding government investment in public transportation and also merged it into one file with carbon emission data. For similar reasons, the new file started in 2007 and ended in 2018 to ensure every data point had a corresponding data point. This is because BTS only started collecting data in 2007. All of this was done through a cleaner file, which we made in Python using Pandas and CSV libraries.

3.3 Data Analysis and Algorithms

Because our group was primarily interested in using Linear Regression as an analysis, we first had to ensure that the residual plot for a given data correlation was randomly distributed.

$$e = y - \hat{y}$$

e : Residual
y : Observed Value
$\hat{y}$: Predicted Value

The formula for residuals to establish linear correlation[9]

To find the residual, we would have to get the (y-axis real value - (slope of regression line) * (corresponding variable on x-axis) + y-intercept of the regression line). We used Numpy to find the slope of the regression line twice, once with the independent variable "thousands of inflation-adjusted USD" plotted against the dependent variable "Kilotons of carbon emissions emitted annually". The second time, the independent variable was "Kilotons of Carbon Emissions emitted annually vs the dependent variable, biodiversity measured with the Living Planet Index scaled by a factor of 100".

After confirming that the residuals were randomly scattered, we utilized linear regression to find trends. We did this two-step method twice for each variable comparison listed previously. After plotting down all the results with the Matplot library, we had four graphs. The type of linear regression used during this project was least squared regression, denoted in the formula below. The Numpy library comes with functions to perform such calculations. After performing linear regression, we extract the slope of the regression line, R and R-squared, and standard deviation to analyze.

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

The formula for the slope of the regression line can be seen above, where b is the slope and n is the number of data points. This function is built into Numpy. [10]

We also wanted to create a model in order to gain insights from our data. We researched new approaches in machine learning, and decided on using automated machine learning so that our hyperparameter optimization, data cleaning, and model selection could all be automated. EvalML by Alteryx was our library of choice, as we could feed its AutoMLSearch function with time series data. After we prepared our data with the AutoClean library to drop certain null values and scale data, we fed the data to AutoMLSearch. Our target column was the Living Planet Index, our date column was a date-time column version of the original Year column in the data, and our feature columns were the remaining columns. Through this method, we could find insights into our data, such as 1) Exponential Smoothing with Drop Columns Transformer 2) Imputer 3) Time Series Featurizer 4) DateTime Featurizer 5) One Hot Encoder The best method for our data was exponential smoothing. This is a method for univariate data that also includes a time trend and/or a seasonal component, meaning that there are potential trends or repeating cycles in our data. Univariate means that our data consists of observations on only a single characteristic, which is our Living Planet Index column.

4 Results

4.1 Comparison of Carbon Emissions vs Biodiversity

Firstly, we created a residual plot in Figure 1 to verify if a linear regression fits the North American Annual Carbon Emissions and Biodiversity Index. We saw a random dispersion indicating that linear regression would be acceptable to use. Although there is a slight concern about the gap in the lower half of the domain, 5.5-7.5, we believe that the gap is there purely out of chance, as every other part of the residual plot appears to be randomly scattered. This random scattering of the residual plot indicated that utilizing linear regression was valid for comparing these two datasets. Figure 2 shows the linear regression graph we used for these two data sets and we had an R-value of -0.87, indicating a strong negative correlation between North America's annual carbon emissions and the biodiversity index in North America. We would attribute this to the effects of greenhouse gases on local wildlife. We saw a slope of around -1.11e-8, indicating that on average, for every hundred million kilotons of carbon emissions, the predicted decrease in biodiversity was by about 0.01 on a scale of 1 in the Living Planet Index unit. Our R-squared value was approximately 0.76, meaning that 0.76 of the variation in North American biodiversity can be attributed to North American Annual Carbon Emissions.

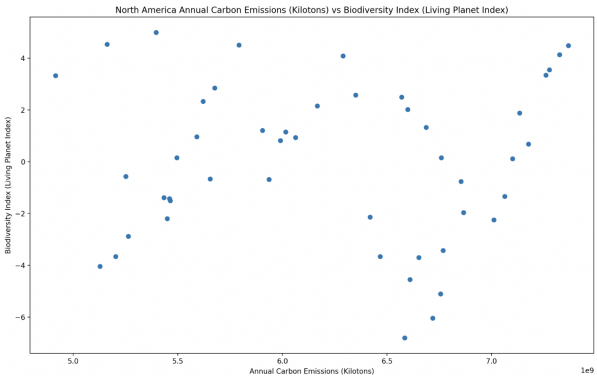

Figure 1: This graph relates annual carbon emissions across North America on the x-axis to the biodiversity index from the Living Planet Index (y-axis) from 1970 to 2018.

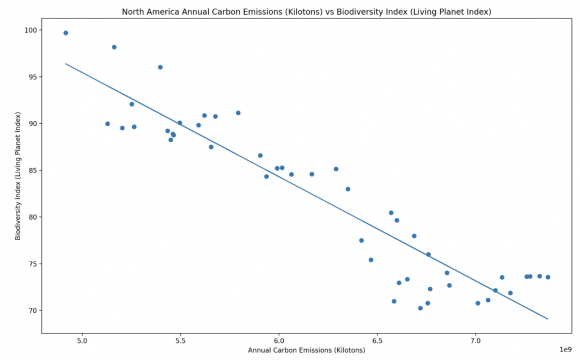

Figure 2: This scatter plot relates the residual of the annual carbon emissions across North America on the x-axis to the biodiversity index from the Living Planet Index on the y-axis from 1970 to 2018.

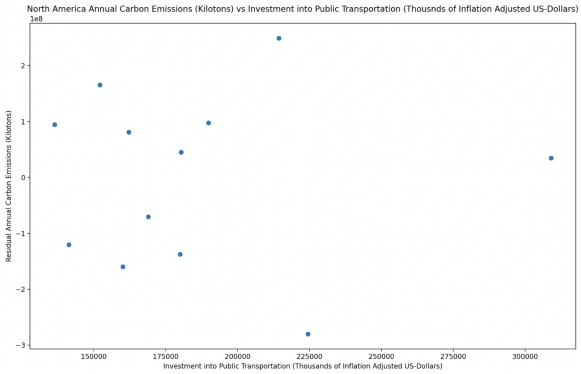

Figure 3: This scatter plot showcases the residual investment into public transportation in the United States on the x-axis to annual carbon emissions across North America on the y-axis from 2007 to 2018.

4.2 Comparison of Investment Into Public Transportation vs Carbon Emissions

We once again created a residual plot in Figure 3 in order to better understand what form of regression would be best to use for comparing investment into public transportation and carbon emissions. The residual plot of North American Investment into Public Transportation vs. Carbon Emission is randomly scattered. This lead us to believe that linear regression was appropriate. There were some potential high-leverage points that could affect the data heavily, including the point close to 300,000 USD invested into public transport. We also noted that this was also a potential outlier that could have impacted our data. Due to the fact that data points in Figure 3 are randomly scattered, this meant that linear regression would be a valid way to measure correlation in our datasets. In Figure 4, we used linear regression in order to discover our R, R-squared, and slope for all of the prospective relationships that we wanted to analyze. We had an R-value of 0.67, indicating a moderately strong positive correlation between North American investment in public transport and carbon emission. This is most likely because fewer people are taking cars as a mode of transport, which results in less carbon emissions. Our R-squared value was approximately 0.46, meaning that 0.46 of the variation in carbon emissions is attributable to the North American investment in public transport. The slope of the regression line was around 4732, meaning that for every thousand dollars invested into Public transport, the predicted carbon emissions would, on average, increase by 4732 kilotons. We will further analyze this counter-intuitive statistic later in the paper. Additionally, we calculated that the standard deviation of the linear regression line was 3377369520 kilotons. This alerted us to potential high variation

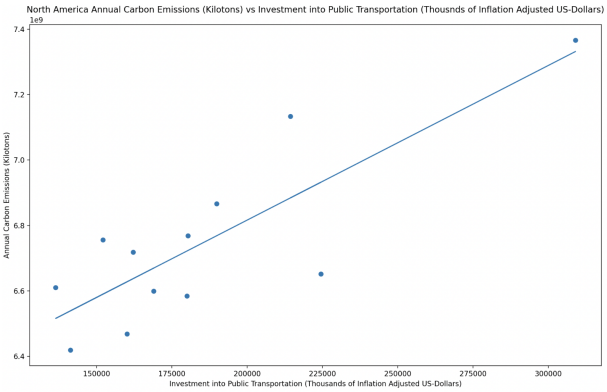

Figure 4: This graph illustrates the relationship between investment into public transportation in the United States on the x-axis and annual carbon emissions across North America on the y-axis from 2007 to 2018.

in the dataset; however, we then verified our work by checking the scale of the y-axis of the graph. Since the y-axis is measured at a scale of 1e9, we confirmed that variation in the data was moderately low.

4.3 Machine Learning Results

Another interesting insight we gained from the automated machine learning choosing exponential smoothing was that recent data points were more important for explaining data during the training. This is because exponential smoothing assigns a high weighting for recent data points, and decreases the weighting exponentially for data that is further in the past. Further transformers EvalML applied to our data were a Drop Columns Transformer, to drop columns with no data in them, and an imputer that filled in remaining missing values with the mean (where dropping them or filling them in

with zeros would be unreasonable). Other further transformers were a time series feature that delayed input features and the target variable and a DateTime feature, which extracted years from our DateTime column. Finally, there was a one-hot encoder that encoded categorical features into a one-hot numeric array.

5 Discussion

5.1 Discussion of Investments and Correlation to Biodiversity

From our results, we can see that there is a strong negative correlation between carbon emissions and biodiversity and a moderate positive correlation between investment into public transportation and carbon emissions. We accomplished the first aim of our study, as the clear correlation between biodiversity and carbon emissions was established, as seen with the R-value of -0.87. We believe that the loss in biodiversity can be attributed to greenhouse gases, global warming, and the excessive use of fossil fuels. The moderately strong positive correlation between investment and carbon emission is seen with the counter-intuitive result of the R-value of 0.67. This indicates one of two potential scenarios: either investment into public transport is ineffective or that more investment into public transportation is required to see real results. Currently, we cannot see which is the case as there is not enough public data available to determine which of the outcomes is true. Due to the positive correlation between investment in public transportation and carbon emissions and the negative correlation between carbon emissions and biodiversity, a negative correlation between biodiversity and investment in public transportation through the confounding variable of carbon emissions a negative correlation between biodiversity and investment in public transport would be expected. Seeing that our results only support two potential solutions, we would like to recommend governments to investigate a period of large investments into public transport to determine its effectiveness in improving biodiversity. Upon completing this investigation, a proper analysis of our hypothesis could be tested, meaning real solutions to improving biodiversity could potentially still be created through the incorporation of governmental investment plans into public transport. Seeing that we are left with two potential solutions, we have accomplished the second aim of our study, which was to seek out robust and realistic solutions to the biodiversity crisis.

5.2 Applications

The results of our findings conclude that a potential government test period for investment into public transport is required to determine potential solutions. The rationale for this testing period is further backed by the fact that other papers in the field have found a negative correlation between carbon emissions and public transport development, contrasting our results. For example, the paper The Impact of Public Transportation on Carbon Emissions—From the Perspective of Energy Consumption by Qin-Lei Jing et al discovered that the development of public transportation has "a significant carbon emission reduction effect" [11]. Our results about the correlation between carbon dioxide emissions and biodiversity align with findings from other researchers in the field. More precisely, a negative correlation was also found by K.R. Shivanna in Climate Change and its Impact on Biodiversity and Human Welfare in ProcIndian National Science Academy Journal, 2022 [12]. Real-world application of this research is in supporting policy decisions with scientific research for the purposes of preserving the environment. Shivanna researched how agricultural practices across Asia and North America accounted for CO2 emissions, and how scientific research that supported the need for more sustainable methods of agriculture would help convince governments, regulatory bodies, and other organizations about the need for preserving biodiversity. Our findings could be used for government policy regarding federal investments in public transportation. For example, the methods used in our study can work with data from any region, allowing for region-specific plans for solutions that governments can undertake. This will allow governments to implement plans that will be the most effective for their nation or province/state and take the best financial routes available to them.

5.3 Sources of Errors

The main source of error for our study was the limited pool of data available to the public. For the data we used for carbon emissions, we used data that combined both the United States and Canada's carbon emissions, as that was what the source provided us with. For the data for investments into public transport, we were only able to utilize data regarding the United States, as Canada has not released any data related to public spending on transportation. However, we still believe that our study still holds true overall as Canadian carbon emissions account for less than 15% of the carbon emissions between Canada and the US [13]. As the difference between the two data sets is not extremely drastic, our study will still hold largely true for Canada and the United States. An area of improvement

that we noticed for our data would be the inclusion of investment into specific sectors of public transport. If this were to be included in our data sets, the usability of our study would improve greatly, as specific technologies could be pinpointed as being the most worthwhile for improving biodiversity. This would also allow for more region-specific solutions that could be implemented for an even greater degree of success.

6 Conclusion

We found a strong negative correlation between carbon emissions and biodiversity, and a moderate positive correlation between carbon emissions and investments into public transportation. Based on these results, either government spending on public transport is ineffective, or more investment is required to establish a correlation between biodiversity and government spending on public transport. But, based on the data, we can not confirm which one it is. Hence, we first believe that there is a need for a larger pool of data regarding these topics to the public. We recommend governments to have a test period for investment into public transport, to assess its true result on biodiversity and release the data to the public. If this were to be achieved, our study could be utilized by governments to create region-specific solutions to improving biodiversity and prioritizing investments in the most economically efficient technologies. We hope that our study can be used in tandem with others to help create a multifaceted and effective approach to the biodiversity crisis, to create a better and greener future for everyone.

7 Acknowledgements

Our group would like to sincerely thank Mr. Niven for being a mentor who provided us with statistical insights. We would also like to thank STEM Fellowship for hosting this challenge, which provides students with a space to show their research findings. We would also like to acknowledge the papers we used to initially research the topic of utilizing AI to preserve Biodiversity [14] [15] [16] [17] [18] [19].

References

[1] Rebecca Lindsey. Climate change: Atmospheric carbon dioxide.

[2] Nasa. How do we know climate change is real.

[3] Lisa Hendry Holly Chetan-Welsh. How are climate change and biodiversity loss linked?

[4] How do governments combat climate change?

[5] United States Environmental Protection Agency. Sources of greenhouse gas emissions.

[6] Hannah Ritchie and Max Roser. Co2 emissions.

[7] Living Planet Index. Living planet index.

[8] Bureau of Transportation Statistics. Transportation expenditure by mode, federal (chained 2017 dollars).

[9] ABS. Residual in regression analysis.

[10] Dobrimir Dikov. Least-squares method to estimate the cost function.

[11] Wei-Qing Yu Qin-Lei Jing, Han-Zhen Liu and Xu He. The impact of public transportation on carbon emissions—from the perspective of energy consumption.

[12] K. R. Shivanna. Climate change and its impact on biodiversity and human welfare.

[13] Worldometer. Co2 emissions by country.

[14] A. Holzinger, K. Keiblinger, P. Holub, K. Zatloukal, and H. Müller. Ai for life: Trends in artificial intelligence for biotechnology. *N. Biotechnol*, pages 16–24, 2023.

[15] D. Silvestro, S. Goria, T. Sterner, and A. Antonelli. *Improving biodiversity protection through artificial intelligence*, volume 5, page 415–424. Nature Sustainability volume, 2022.

[16] A. Sen, B. Sterner, N. Franz, C. Powel, and N. Upham. Combining machine learning reasoning for biodiversity data intelligence. *AAAI*, 35(17):14911–14919, 2021.

[17] Rout George Kerry et al. An overview of remote monitoring methods in biodiversity conservation. *Environ. Sci. Pollut*, 29(53):80179–80221, 2022.

[18] E. K. Nti, S. J. Cobbina, E. E. Attafuah, E. Opoku, and M. A. Gyan. Environmental sustainability technologies in biodiversity, energy, transportation and water management using artificial intelligence: A systematic review. *Sustainable Future*, 4, 2022.

[19] Kellenberger B. Beery S. et al Tuia, D. Perspectives in machine learning for wildlife conservation. *Nature Communications*, 13(792), 2022.