Developing a neural network that uses satellite imagery to estimate carbon dioxide emissions in the Philippines

EWEMIZ C. INSIGNE, KIRSTEN DIANNE G. DELMO, NICO ANDREI R. SERRATO, and GERALD U. SALAZAR

Philippine Science High School Western Visayas Campus - Department of Science and Technology (DOST-PSHSWVC), Brgy. Bito-on, Jaro, Iloilo City 5000, Philippines

Article Info

Submitted: May 14, 2021
Approved: Jul 21, 2021
Published: Aug 30, 2021

Keywords:
artificial intelligence
carbon dioxide
satellite imagery
neural network
data science

Abstract

In recent years, satellite imagery has become a popular subject of studies involving artificial intelligence. AI and satellite imagery were used together to estimate real-life variables in various fields, such as predicting poverty levels, detecting oil spills, and analyzing coal mine areas. With the rising importance of climate change, this study aimed to develop a convolutional neural network trained with daytime and nighttime data through a transfer learning approach to predict carbon dioxide emissions in the Philippines. The developed scripts were adapted and based on code from public GitHub repositories and existing research. The trained model was concluded to have high predictive power (r^2 = 0.74), and can be used as a starting point for more fully developed software and models that can serve as an alternative method of collecting CO_2 emissions data.

Introduction. - Data science uses statistics, artificial intelligence (AI), and available data from various sources to process into valuable information [1]. AI in data science looks for patterns, predicts outcomes, and offers evidence-based information using the data that it was given [2]. More and more AI-based tools and technologies are beginning to emerge as datasets begin to be more readily available online and in other publicly-available sources. A specific type of AI, the convolutional neural network (CNN), is mainly used for the analysis of images. It takes an input and extracts features from the image to be analyzed by neurons – an interconnected system of cells that takes input and multiplies it by a specific weight that constantly changes until the CNN can correctly predict an outcome.

Recent studies were able to use satellite imagery to predict poverty levels in an area [3,4,5]. Other studies utilized satellite imagery for purposes such as detecting oil spills, analyzing coal mine areas, and detecting clouds [6,7,8]. The effectiveness of combining satellite imagery and AI and show how it has many potential uses in various fields.

With the increase of CO_2 concentration in the atmosphere, the need for “the quantification of the spatial distribution of CO_2 in the atmosphere” was felt by researchers in the last decade [9]. The main focus of currently available literature on estimating CO_2 emissions focus on mathematical models that don’t utilize AI. [9,10]. The correlation of nightlights and CO_2 emissions using satellite imagery has also been proven by studies, where highly-lit points were characterized to be highly urbanized and were seen to have high CO_2 emissions [9,11]. While using AI (specifically neural networks) to estimate CO_2 emissions is not an uncommon method [12], this study differs in that it develops a CNN that estimates CO_2 emissions in specific 1km x 1km areas within the Philippines, and utilizes a transfer learning approach. Transfer learning is a method that utilizes the learning extracted from one problem or variable (such as nightlights data) and uses it to make better predictions for related data.

The objective of this paper is to see how effective the use of AI with satellite imagery can be for the estimation of real-life variables, such as CO_2 emissions, in order to prove its versatility as a tool to be used in a wide range of applications in various fields. As a tool specifically used to estimate CO_2 emissions, it could serve as an alternative method of collecting such data. Thus, the goal of this study is to develop an AI—specifically a CNN—that uses satellite imagery to estimate CO_2 emissions in the Philippines. The specific objectives are enumerated below. They also serve as an outline for the succeeding Methods section.

(i) Design an algorithm for the CNN model;

(ii) Collect latest (at the time of development) satellite imagery and CO_2 emissions data;

(iii) Code the scripts for data preprocessing, CNN model training, and evaluating the model’s estimations using Python 3.8;

(iv) Train the CNN using the collected data; and

How to cite this article:
CSE: Insigne EC, Delmo KDG, Serrato NAR, Salazar GU. 2021. Developing a neural network that uses satellite imagery to estimate carbon emissions in the Philippines. Publiscience. 4(1): 88–92.
APA: Insigne, E.C., Delmo, K.D.G., Serrato, N.A.R., & Salazar, G.U. (2021). Developing a neural network that uses satellite imagery to estimate carbon emissions in the Philippines. *Publiscience*, *4*(1), 88–92.

For supplementary data, contact: publiscience@wvc.pshs.edu.ph.

(v) Train a ridge regression model using the CNN's extracted features for evaluation and data visualization of results

Methods. - This study consisted of five phases: *Design*, *Data Collection* (for data pre-processing), *Development*, *Training*, and *Evaluation*. The *Design* phase was for the creation of the algorithm for the CNN model; *Data Collection* phase was for the collection of images and nightlight values for pre-processing used for training the developed program; *Development* phase was for the coding and debugging of the scripts; *Training* phase was for the tuning and fitting of the model, and; *Evaluation* was for the training of the ridge regression model for evaluation and data visualization of results.

Design phase. During the *Design* phase, a flowchart and the equivalent pseudocode of the program's process were made to be used as a basis for the creation of our actual program during the *Development* phase. The flowchart presented the various parts of the process in an human-readable format that can then be converted to code. This phase employed the use of open-source codes by Jean et al. [4] and Tingzon et al. [5] that were available on the authors' GitHub repositories.

Data Collection phase. In order to properly train the AI, certain data were gathered and pre-processed. This study used the Open-Data Inventory for Anthropogenic Carbon dioxide (ODIAC) of the National Institute for Environmental Studies (NIES) Japan for the year 2015 [13]. The data points for the Philippines were filtered out, as seen in Figure 1. Daytime satellite images were then collected through a script that downloads 1 km by 1 km Google Static Maps API images according to the given data points from the CO_2 emissions data. Nighttime lights data was collected from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) data of the Earth Observation Group, Payne Institute for Public Policy [14].

Development phase. The method used in the developed scripts were adapted from the transfer learning approach used by Jean et al. [4] Daytime satellite image features were extracted by the CNN to estimate the CO_2 emission level in an area.

The scripts were developed based on the algorithm made in the design phase as well as open-source scripts and codes from similar studies publicly available on GitHub [15,16,17]. Python 3.8 was the primary programming language used for development, and the scripts were coded in JupyterLab 1.1.4 and JupyterNotebook 6.0.1.

The first three scripts were used for the pre-processing of data. The first script filtered out the CO_2 emission data points for the Philippines. The second script downloaded the daytime satellite images for each datapoint with a zoom level of 13, and the third script compiled the corresponding nighttime radiance for the coordinates of each CO_2 emission datapoint.

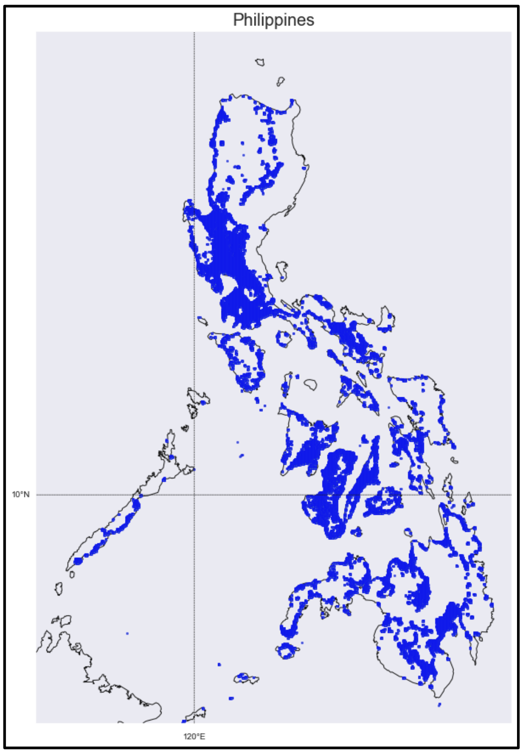

Figure 1. Philippine CO_2 emission data points. The data points for the Philippines were filtered out from the ODIAC dataset.

The first script also created a dataframe with the latitude and longitude values, the corresponding CO_2 emission. These values were then exported into a comma-separated values (.csv) file.

The second script downloads 400 px by 400 px Google Static Maps API images at a zoom level of 13 (corresponding to an estimated 1 km by 1 km area) for each data point. A total of 79942 unique images were downloaded.

The third script collected the corresponding nightlight radiance for each point in the dataset. The values were then appended to the data frame created in the first script.

The fourth script was for the training. In its initial stages, the dataset and the corresponding images were divided into two folders: 80% of the images were used for training and the remaining 20% were used for validation/testing of the model. Three nighttime light intensity classes were obtained by fitting a mixture of three Gaussian distributions to the relative frequencies of the nighttime light intensity values.

Training phase. The training phase aimed to train the CNN with images so that it could create its own model in order to analyze patterns and make accurate estimations close to the given CO_2 data.

In the first part of the training, we fine-tuned a pre-trained model, VGG F, to estimate nighttime light intensity at various locations given the corresponding daytime satellite images. This pre-trained model network is an eight-layer deep convolutional neural network (DCNN), which had been originally designed and trained for image classification on ImageNet [18].

The training method and code were heavily adapted from Mather's Predicting Poverty repository on GitHub [15] and the Pytorch CNN Training Method [19].

After training, the model was then tested by making estimations using the testing data set.

Evaluation phase. Once the CNN was trained, the extracted features were then used to train a ridge regression model. It specifically used k-fold cross-validation. Python was also used in data visualization, specifically for the creation of the graphs.

Data Analysis. For data visualization and analysis, this study used Python 3.8 and different libraries that were available and used for statistical calculations and data visualization such as NumPy, Pandas, Scikit-learn, and Matplotlib.

This analysis utilized a cross-validation technique, which is a data resampling method that assesses the generalization ability of a predictive model and prevents overfitting. The *k*-fold cross-validation technique is much less prone to selection bias compared to other cross-validation methods. In *k*-fold cross-validation, the process starts out by dividing the dataset into given k subsets and uses *k-1* subsets as the training sets while the remaining set serves as the testing set. This cross-validation method is then repeated *k* times, using different testing sets from the original *k* subsets each time. The *k* value for this study was 5, similar to that of Jean et al. [4].

Safety Procedure. As this study focused on AI, all processes were done on a computer. There were no major ethical issues dealt with. The data used is publicly available and the source codes are open-access.

Results and Discussion. - The aim of this study is to develop a convolutional neural network to analyze CO_2 emissions in the Philippines using satellite imagery, and see how effective the use of AI with satellite imagery can be for the estimation of real-life variables (such as CO2 emissions). This was done by creating the algorithm for the program, collecting data that would be used for pre-processing and training, developing the scripts used for the program, and then evaluating and visualizing the results. The evaluation phase specifically employed k-fold cross validation, with various libraries in Python such as Matplotlib used for data visualization.

The current results, shown in Figure 2, showcase strong predictive power using the model trained on a dataset with 79942 data points. With a r^2 value of 0.74, this means that the trained model fits the data well, with 74% of the total variation being accounted for.

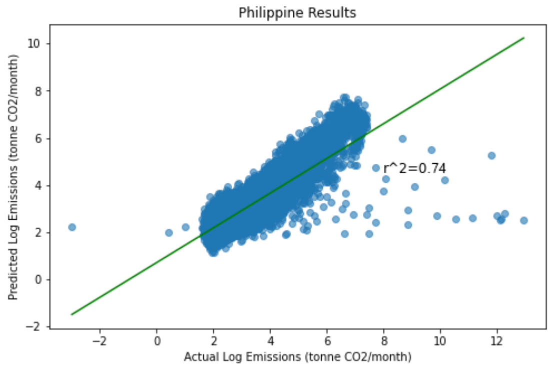

Figure 2. Philippine results. Predictions and reported r^2 value are from five-fold cross-validation. Green line shown is the line of best fit.

To assure the accuracy of this statistic, five random sampling trials were performed. Five thousand random points were selected from the data set and were processed through the final script to predict emissions. The summary of trials is presented in Table 1.

Table 1. Summary of five trials of 5000 randomly selected points.

Trial	R^2	Ridge Score (Validation)	Ridge Score (Training)
1	0.64	0.62	0.85
2	0.69	0.66	0.86
3	0.65	0.63	0.84
4	0.66	0.57	0.92
5	0.65	0.58	0.90
Ave	0.66	0.61	0.87

The average r^2 value of the five trials is 0.66, and the average ridge score for validation is 0.61. Compared to the final result's r^2 of 0.74 and ridge score for validation of 0.73, there is some difference. The small disparity can be explained by the fact that the trials only used 5000 data points, which affects the forward pass process within the predicting consumption script. Still, the results are notable as it shows that 5000 data points (6.7% of the original 79942) can already show high predictive power, explaining up to 69% of the variance.

The trials' ridge scores for training are much higher compared to their respective ridge scores for validation, which may show that there is some overfitting. In the final result, however, the difference between the training ridge score (0.78) and the validation ridge score (0.74) is much smaller, showing that a bigger dataset helps minimize overfitting.

The ODIAC dataset served as the proxy ground-truth data in this study but can also be used as a basis for a comparison of the model's performance. In a study by Chen et al. [20], the ODIAC dataset's predictions among 14 large cities were compared to emission inventory statistics. This revealed that in some cities, especially in developing countries, the dataset overestimates. As Chen states, this overestimation could be due to the "poor correlation between nightlight intensity with human activity [...] in developing countries." This is similar to a problem Jean et al. [4] cites, in which areas with very low light levels (often in developing countries) often show little variation, leading to models incapable of distinguishing differences in economic activity. Jean et al. improved upon existing studies by using a transfer learning approach in their model, which allowed them to have better estimates in countries and areas with minimal nightlight data. This study applied a similar approach; it was based on the transfer learning approach by Jean et al., which utilized both daytime images and nighttime data to estimate CO_2 emissions, thus making predictions within the Philippines - especially in areas that are darker or with low luminosity values - more accurate.

Though the results cannot be directly compared, due to the models measuring different variables, a comparison may offer some valuable insight. It was found that the developed model of this study estimates CO_2 emissions better than how the models of Jean et al. (r^2_{max} = 0.55) and Tingzon et al. (r^2 = 0.63) predict economic status.

Limitations. One of the primary concerns in this study was the lack of "true" ground truth data. Directly-measured CO_2 emissions data in the Philippines could not be found available in any of the country's database agencies, despite many efforts. Because of this, the study decided to use the ODIAC dataset as a proxy for ground truth data, which, while still a valid substitute, may contain overestimations and underestimations in urban areas and rural areas respectively as it is based on space-based nighttime light data and individual power plant emission/location profiles.

Due to restraints in time and processing power, data points with no emissions (data points with a value of 0) were removed, which left the 79942 data points which were then used for the rest of the study. It is recommended that in future studies and developments, all data points, including those with 0-emission values, should be included.

A heatmap couldn't be done in this study due to the lack of time and experience on the part of the researchers.

Conclusion. - Through the development of the scripts, adapted from and based on the publicly available GitHub repositories of Jatin Mather (2016) and Jean et al. (2016), this study was able to conclude that the developed convolutional neural network model used to estimate CO2 emissions in the Philippines has strong predictive power. It improves upon ODIAC, an existing CO_2 dataset, by looking at both daytime and nighttime images and data through transfer learning. The developed scripts and CNN model can be used as a starting point for more fully developed software and models that can serve as an alternative method of collecting CO_2 emissions data.

Recommendations. - Though the scripts and algorithms may prove themselves to be valid processes with strong predictive power, concerns may arise if the dataset itself has some issues - such as overestimations and underestimations - similar to that of ODIAC. It is recommended that, if there is access to a cleaner dataset of direct measurements from the Philippines rather than estimations, future researchers may want to utilize that instead.

This model is limited to data in 2015. It is possible that this model could be used for more recent years, but further research is required in order to confirm so.

A direct comparison between the accuracy of the ODIAC dataset and the accuracy of the developed model (compared to existing emission inventories) through graphs or a heatmap may also give additional insights.

Acknowledgment. - The researchers would like to thank Mr. Neal Jean of Stanford University and Thinking Machines Data Science Inc. for their invaluable time, knowledge, and support that was extended to this study.

References

[1] Van der Aalst WP. 2018. Data Science in Action. In: Process Mining. Berlin: Springer Berlin. p. 3–23.

[2] Dinov I. 2019. Data Science and Predictive Analytics: Biomedical and Health Applications Using R. S.l.: Springer.

[3] Xie M, Jean N, Burke M, Lobell D, Ermon S. 2016. Transfer learning from deep features for remote sensing and poverty mapping. In: Schuurmans D, Wellman MP, editors. c2016. AAAI'16. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence; 2016 Feb 12-17; Phoenix, Arizona. Palo Alto (CA): AAAI Press. p. 3929–3935.

[4] Jean N, Burke M, Xie M, Davis WM, Lobell DB, Ermon S. 2016. Combining satellite imagery and machine learning to predict poverty. Science 353: 790–794.

[5] Tingzon I, Orden A, Sy S, Sekara V, Weber I, Fatehkia M, Herranz MG, Kim D. 2019. Mapping Poverty in the Philippines Using Machine Learning, Satellite Imagery, and Crowd-sourced Geospatial Information. International Conference on Machine Learning AI for Social Good Workshop.

[6] Zhi-Ming D, Li-Xia G, Jian-Kui Z, Xue-Bin Z. 2015. Oil-Spills Detection in Net-Sar Radar Images Using Support Vector Machine. The Open Automation and Control Systems Journal 7(1): 1958–1962.

[7] Demirel N, Emil MK, Duzgun HS. 2011. Surface coal mine area monitoring using multi-temporal high-resolution satellite imagery. International Journal of Coal Geology 86:3–11.

[8] Bai T, Li D, Sun K, Chen Y, Li W. 2016. Cloud Detection for High-Resolution Satellite Imagery Using Machine Learning and Multi-Feature Fusion. Remote Sensing 8: 715.

[9] Ghosh T, Elvidge CD, Sutton PC, Baugh KE, Ziskin D, Tuttle BT. 2010. Creating a Global Grid of Distributed Fossil Fuel CO2 Emissions from Nighttime Satellite Imagery. Energies 3: 1895–1913.

[10] Makido Y, Dhakal S, Yamagata Y. 2012. Relationship between urban form and CO2 emissions: Evidence from fifty Japanese cities. Urban Climate 2: 55–67.

[11] Rayner PJ, Raupach MR, Paget M, Peylin P, Koffi E. 2010. A new global gridded data set of CO2 emissions from fossil fuel combustion: Methodology and evaluation. Journal of Geophysical Research 115.

[12] Abdullah L, Mohd Pauzi H. 2015. METHODS IN FORECASTING CARBON DIOXIDE EMISSIONS: A DECADE REVIEW. Jurnal Teknologi 75.

[13] Tomohiro O, Shamil M. 2015. ODIAC Fossil Fuel CO2 Emissions Dataset (ODIAC v1.7). Center for Global Environmental Research, National Institute for Environmental Studies, Downloaded from https://bit.ly/2SPGE5O.

[14] Elvidge, Christopher D., Kimberly Baugh, Mikhail Zhizhin, Feng Chi Hsu, and Tilottama Ghosh. "VIIRS night-time lights." International Journal of Remote Sensing 38, no. 21 (2017): 5860–5879.

[15] Mather, J. 2019. Predicting poverty replication. https://bit.ly/34IeUmm

[16] Tingzon I, Orden A, Sy S, Sekara V, Weber I, Fatehkia M, Herranz M, Kim D. 2018. Mapping Poverty in the Philippines Using Machine Learning, Satellite Imagery, and Crowd-sourced Geospatial Information. https://bit.ly/3fNvSGp.

[17] Jean N, Burke M, Xie M, Davis WM, Lobell DB, Ermon S. 2016. Combining satellite imagery and machine learning to predict poverty. https://bit.ly/3vN79HL.

[18] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. 2009. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition.

[19] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems 32 https://bit.ly/3gbmoUq.

[20] Chen J, Zhao F, Zeng N, Oda T. 2020. Comparing a global high-resolution downscaled fossil fuel CO2 emission dataset to local inventory-based estimates over 14 global cities. Carbon Balance and Management 15.

[21] Oxford University Press. 2020. Overfitting. In: Lexico. [accessed 2020 Dec 10]. https://bit.ly/3uH04Hb.

[22] Brownlee J. 2019. Overfitting and Underfitting With Machine Learning Algorithms. Machine Learning Mastery. [accessed 2020 Dec 10]. https://bit.ly/34Irm5K.

[23] Ying X. 2019. An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series 1168: 022022.