ISABELLA DE MELO SOUSA

PREDICTION OF SOCIOECONOMIC INDICATORS IN VALE DO RIBEIRA USING DEEP LEARNING AND SATELLITE IMAGERY

São Paulo
2022

ISABELLA DE MELO SOUSA

PREDICTION OF SOCIOECONOMIC INDICATORS IN VALE DO RIBEIRA USING DEEP LEARNING AND SATELLITE IMAGERY

Work presented to the Escola Politécnica da Universidade de São Paulo in order to obtain the Bachelor's Degree in Computer Engineering.

São Paulo
2022

ISABELLA DE MELO SOUSA

PREDICTION OF SOCIOECONOMIC INDICATORS IN VALE DO RIBEIRA USING DEEP LEARNING AND SATELLITE IMAGERY

Work presented to the Escola Politécnica da Universidade de São Paulo in order to obtain the Bachelor's Degree in Computer Engineering.

Advisor:

Prof. Dr. Pedro Luiz Pizzigatti Corrêa

Co-Advisor:

PhD Marina Jeaneth Machicao Justo

São Paulo
2022

Prof. Dr. Pedro Luiz Pizzigatti Corrêa

ACKNOWLEDGMENTS

I would like to thank my family for their unconditional support throughout my graduation. As well as my dear friends for their encouragement, in special Gabriel Barbutti de Lima Baker, Megumi Tsuru and Vinicius Akira Imaizumi, for their support and insightful discussions about the project.

I also would like to express my sincere gratitude to my supervisors, Prof. Dr. Pedro Luiz Pizzigatti Corrêa and PhD Marina Jeaneth Machicao Justo, for guiding me in this difficult and rewarding journey that was developing this project. I thank them for all the help, attention and above all, their patience.

Finally, I am also grateful to the PARSEC group and all its researchers for their support in conducting this project.

RESUMO

Medidas-chave dos indicadores socioeconômicos são essenciais para a tomada de decisões políticas informadas, mas devido aos altos custos e dificuldades operacionais dos esforços tradicionais de coleta de dados, a obtenção de dados socioeconômicos confiáveis continua sendo um desafio, principalmente nos países em desenvolvimento. Este trabalho apresenta uma metodologia de aprendizado profundo para estimar indicadores socioeconômicos utilizando imagens de satélite. O modelo de rede neural desenvolvido foi treinado na região brasileira de São Paulo e Paraná com o objetivo de analisar o indicador socioeconômico de renda na região do Vale do Ribeira. O modelo resultou em um R-quadrado de 0,4016 e teve um desempenho significativamente melhor do que o modelo treinado apenas com as bandas RGB.

Palavras-Chave – aprendizagem profunda, imagens de satélite, indicadores socioeconômicos.

ABSTRACT

Key measures of socioeconomic indicators are essential for making informed policy decisions, but due to the high costs and operational difficulties of traditional data collection efforts, obtaining reliable socioeconomic data remains a challenge, particularly in developing countries. This work presents a deep learning methodology to estimate socioeconomic indicators using satellite imagery. The neural network model developed was trained at the Brazilian region of São Paulo and Paraná with the goal of analyzing the socioeconomic indicator of income in the Vale do Ribeira region. The model yielded a R-squared of 0.4016 and performed significantly better than the model trained only on RGB bands.

Keywords – deep learning, satellite imagery, socioeconomic indicators

LIST OF FIGURES

LIST OF TABLES

CONTENTS

PART I

INTRODUCTION

1 INTRODUCTION

1.1 Motivation

Data on key measures of socioeconomic development can be a powerful tool to assist government policies decisions. However, reliable data on such measures at the local level is still lacking in many parts of the world, especially in developing countries. This mainly occurs due to the difficulties of scaling up traditional data collection efforts, such as government surveys, which can be expensive, time-consuming, and overlook hard-to-reach areas.

In the Brazilian scenario, the main census institute is known as Instituto Brasileiro de Geografia e Estatística (IBGE) and the data collection is done usually at 10-year intervals. This periodicity produces a data gap that prejudices the interpretability of the collection results, since they may not reflect the Brazilian reality in real-time. In this context, the use of machine learning to analyze satellite imagery can help provide valuable information for the socioeconomic monitoring in Brazil.

The motivation for employing the association of these two technologies comes from the recent research of Yeh et al. (2020) [1], that applies deep neural network techniques to estimate socioeconomic criteria in African countries using satellite images of their territories. This type of methodology is very promising due to the low consumption of resources and the great availability of satellite imagery, allowing the coverage of larger and more remote areas and making the process automated and more efficient, without the necessity of on-site collection.

1.2 Objectives

This work aims to develop a deep learning method that uses satellite imagery to estimate the socioeconomic indicator of income in the Brazilian municipalities of São Paulo and Paraná, focusing on the study of the Vale do Ribeira region.

Furthermore, this research is being developed as part of the PARSEC project [2] and it aims to share the results with the group, in which one of its goals is to promote adequate monitoring of the environmental situation in the various biomes in Brazil and to analyze the socioeconomic influence of protected areas on the cities that surround them.

PART II

THEORETICAL FOUNDATION AND STATE-OF-THE-ART

2 THEORETICAL FOUNDATION AND STATE-OF-THE-ART

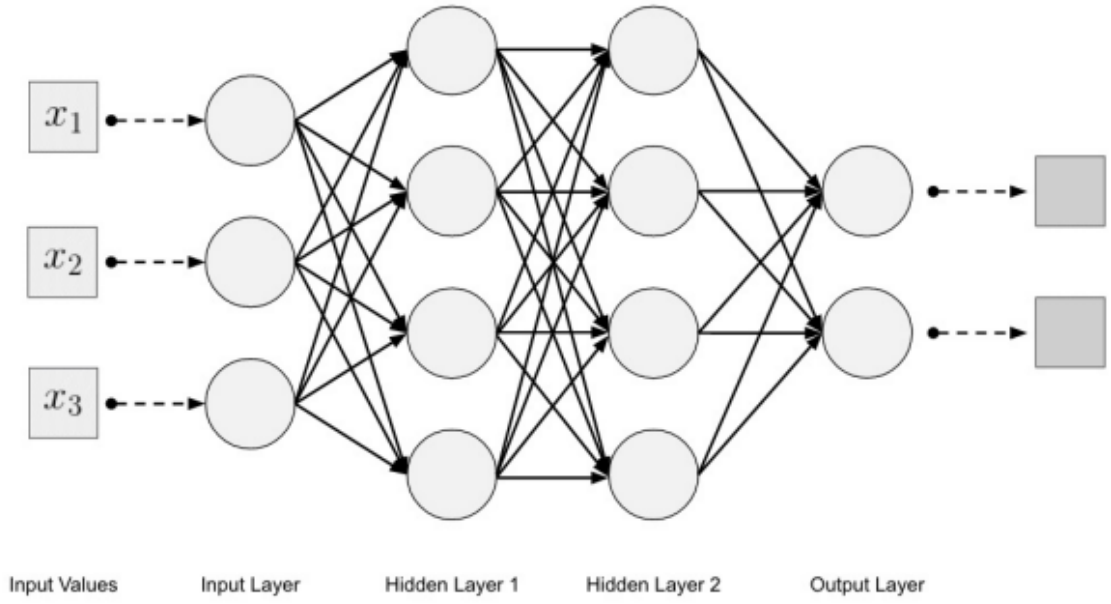
2.1 Deep Learning

Artificial Neural Networks are a computational model that shares some properties with the animal brain in which many simple units are working in parallel with no centralized control unit. The weights between the units are the primary means of long-term information storage in neural networks. Updating the weights is the primary way the neural network learns new information. (PATTERSON; GIBSON, 2017) [3].

The behavior of neural networks is shaped by its network architecture, which can be essentially defined by the following: number of neurons, number of layers and types of connections between layers. [3]

The most well-known and simplest-to-understand neural network is the feedforward multilayer neural network. It has an input layer, one or many hidden layers, and a single output layer. Each layer can have a different number of neurons and each layer is fully connected to the adjacent layer. The connections between the neurons in the layers form an acyclic graph, as illustrated in Figure 1. [3]

Figure 1: Multilayer neural network topology.



Source: Patterson and Gibson, (2017) [3].

Deep Learning refers to the use of Artificial Neural Networks with three or more layers [4]. Note that additional hidden layers can help optimize and refine the accuracy of the model, but on the other hand they increase its complexity and computational cost.

2.1.1 Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is an a Deep Learning technique in which the goal is to learn higher-order features in the data via convolutions.

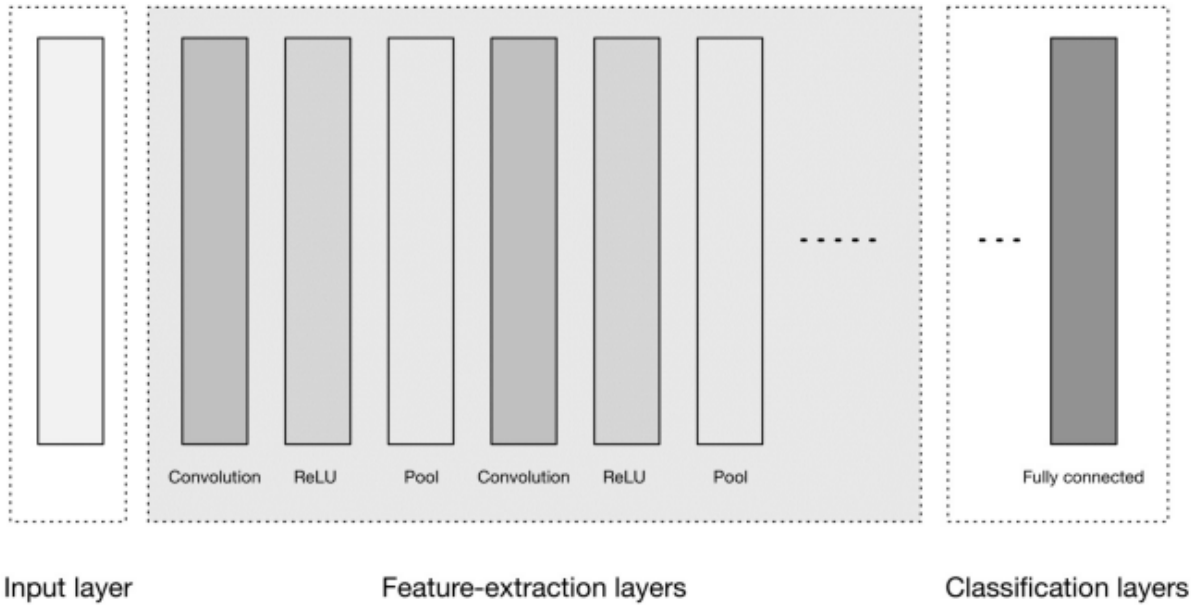
According to Patterson and Gibson (2017) [3], although there are several variations in the architecture of a convolutional neural network, in general it is based on the pattern of layers shown in Figure 2, consisting essentially of three major groups:

1. Input layer: accepts three-dimensional input generally in the form spatially of the size (width × height) of the image and has a depth representing the color channels (generally three for RGB color channels);

2. Feature-extraction (learning) layers: find a number of features in the images and progressively construct higher-order features;

3. Classification layers: consists in one or more fully connected layers to take the higher-order features and produce class probabilities or scores.

Figure 2: High-level general CNN architecture.



Source: Patterson and Gibson, (2017) [3].

2.2 Regression

2.2.1 Linear Regression

Linear regression is a linear approach for modelling the relationship between a variable of interest and one or more explanatory variables (also known as dependent and independent variables). If there is just one explanatory variable, the method is known as simple linear regression; if there are more, it is known as multiple linear regression.

The multiple linear regression model has the form:

$$y_i = \hat{y}_i + \epsilon_i \tag{2.1}$$

And

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_0 x_{ij}, i = 1, 2, ..., n \tag{2.2}$$

Where:

- y_i represents the ith observed response value;
- $\hat{y}_i$ represents the ith response value that is predicted;
- $\beta_0, \beta_1, ..., \beta_p$ represents the coefficients to be estimated (also known as weights);

- x_{ij} represents the independent variable;
- n represents the size of the sample;
- p represents the number of independent variables.

To estimate the regression coefficients, a common approach is the linear least squares function, that consists in minimizing the residual sum of squares (RSS), that can be defined as:

$$RSS = \sum_{i=1}^{n} \epsilon_i. \tag{2.3}$$

Where $\epsilon_i = y_i - \hat{y}_i$ represents the ith residual, which is the difference between ith observed response value and the ith response value that is predicted by the linear model.

In this methodology, the model is not penalized for its choice of weights. This means that for some features the model may place weights that are too large, leading to overfitting.

2.2.2 Ridge Regression

The Ridge Regression solves a regression model where the loss function is the linear least squares function and regularization is given by the l2-norm. It avoids overfitting because it shrinks the regression coefficients by imposing a penalty on their size and it is also very useful to deal with multicolinearity between the independent variables.

The ridge coefficients minimize a penalized residual sum of squares, given by:

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.4}$$

Here $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage. [5].

2.3 Evaluation Metrics

In order to evaluate the performance of a statistical learning method on a given data set, it is necessary to measure how well its predictions actually match the observed data. This section presents some metrics used in this project.

2.3.1 Mean Squared Error (MSE)

The Mean Squared Error (MSE) its one of the most common measures when fitting a regression model. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases.

The MSE value is calculated by the expression:

$$MSE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.5}$$

Where:

- y_i represents the ith observed response value;
- $\hat{y}_i$ represents the ith response value that is predicted;
- n represents the size of the sample.

2.3.2 Coefficient of Determination (R^2)

The coefficient of determination is used to identify the strength of the model, it captures how well the predictions match the observations, or how much of the variation in the observed data is explained by the predictions. Usually the $\mathbf{R^2}$ varies between 0 and 1, being expressed as a percentage (the closer to 1, the better).

According to James et al. (2013) [6], the R^2 can be calculated by the following equation:

$$R^2(y, \hat{y}) = 1 - \frac{RSS}{TSS} \tag{2.6}$$

With

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \tag{2.7}$$

And

$$\bar{y}_i = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.8}$$

Where:

- y_i represents the ith observed response value;
- $\hat{y}_i$ represents the ith response value that is predicted;
- RSS represents residual sum of squares defined by Equation 2.3;
- TSS represents the total sum of squares;
- $\bar{y}_i$ represents the arithmetic mean of all values of the response variable in the sample;
- n represents the size of the sample.

2.3.3 Person Correlation Coefficient (r)

The Pearson correlation coefficient is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.9}$$

Where:

- x_i, y_i are the individual sample points indexed with i;
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sample mean) and analogously for $\bar{y}$;
- n represents the size of the sample.

2.3.4 Spearman's Rank Correlation (rank)

The Spearman rank correlation coefficient is a nonparametric measure of the monotonicity of the relationship between two datasets. [7]

$$rank = \frac{\sum_{i=1}^{n}(R(x_i) - R(\bar{x}))(R(y_i) - R(\bar{y}))}{\sqrt{\sum_{i=1}^{n}(R(x_i) - R(\bar{x}))^2}\sqrt{\sum_{i=1}^{n}(R(y_i) - R(\bar{y}))^2}} \tag{2.10}$$

Where:

- R(x) and R(y) are the ranks of the x and y variables;
- R(x) and R(y) are the mean ranks;
- n represents the size of the sample.

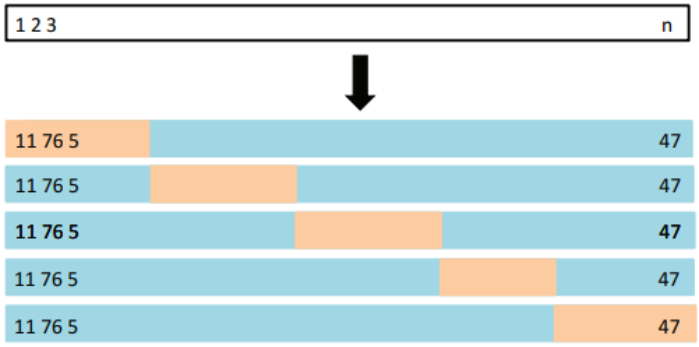
2.4 Cross-Validation

Cross-validation is a technique used to evaluate the capacity of generalization of a prediction model. In this project, two Cross-Validation techniques are adopted: k-Fold Cross-Validation and Leave-one-group-out Cross-validation.

2.4.1 k-fold Cross-Validation

This approach involves dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining k - 1 folds. The mean squared error, MSE, is then computed on the observations in the held-out fold. This procedure is repeated k times; each time, a different group of observations is treated as a validation set. At the end, you can estimate the model error by averaging the errors of all the validations performed. (JAMES et al., 2013 [6]).

Figure 3: A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.



Source: JAMES et al., 2013 [6].

Note that in some approaches, besides the segregation of the data set into training and validation subsets, it is proposed the separation of a holdout subset (test). In such method, the test subset provides a final estimate of the machine learning model's performance after it has been trained and validated.

Figure 4: Separation of the sample into training, validation and test subsets.

Source: HASTIE, TIBSHIRANI e FRIEDMAN (2009) [5].

2.4.2 Leave-one-group-out Cross-Validation

This approach provides train/test indices to split data such that each training set is comprised of all samples except ones belonging to one specific group.

Compared to the 5-Fold Cross Validation, it validates the machine learning model more times resulting in more precise metrics. However, it is more computationally expensive and time consuming.

2.5 Literature Review

2.5.1 Yeh, C. et al. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa (2020)

In this article, public satellite imagery was used to train a combined deep learning model to estimate socioeconomic conditions over space and time in several regions of sub-Saharan Africa.

For this purpose, two ResNet-18 [8] models were trained separately on multispectral daytime imagery from 30m/pixel Landsat and <1 km/pixel nighttime lights imagery. Then, the final layers of the separate models were concatenated in a final fully connected layer and a ridge regression was used to fine-tune the model.

To gather the dataset, the authors assembled data on asset wealth using the Demographic and Health Surveys (DHS) [9] conducted between the years 2009 and 2016, as well as an independent smaller set of household level panel data, the Living Standards Measurement Surveys (LSMS) [10]. With this, the authors were able to establish the villages of the African countries as 'clusters'. Then, Landsat surface reflectance and nightlights images centered on each cluster location were obtained with 30m/pixel resolution, spanning 6.72km on each side.

The model was able to explain on average 70% ($R^2 = 0.7$) of the variation in ground-

based wealth measurements, with the results never below 50% ($R^2 = 0.5$). For the predictive performance over time, the authors used different approaches, with the most predictive one yielding an $R^2 = 0.35$.

Note that for the adaptation of the model to the Brazilian scenario, the studies were focused on replicating only the performance over space, since the dataset used is not of multiple years.

2.5.2 Triñanes et al. Application of a deep learning algorithm for predicting socioeconomic data through satellite images in the Vale do Ribeira (2020)

This article adapted the article of Jean et al. (2016) [11] to predict socioeconomic data for the Brazilian scenario, using the Vale do Ribeira region as a case study. In this approach, a VGG11 [12] model is fine-tuned to estimate nighttime light intensity at various locations given the corresponding daytime satellite images. After extracting the features of the daytime satellite imagery by discarding the nighttime light classification layer, those features are used to train ridge regression models that can estimate cluster-level expenditures or assets.

The socioeconomic data was calculated using the IBGE census of 2010 and the methodology of Abreu et al. (2011) [13]. In this model, 10km x 10km clusters were generated around the center point of each sector. For each cluster at least 10 satellite images were obtained.

Even though the methodology that is used in Triñanes et al. (2020) [14] and the one that is developed in this report are different, part of the socioeconomic data studied is the same (income dimension of the Human Development Index, as explained in section 6.1.3), therefore, the performance of the model developed by Triñanes et al. can be used as a benchmark for comparison.

2.5.3 Megumi Tsuru, Samuel Vieira Ducca, Vitor Dias Souza. Online Platform for inference of indicator data (2021).

This research uses machine learning methods for the analysis of satellite images with the goal of predicting environmental, social, and economic indicators. It adapted the article of Jean et al. (2016) [11] to develop and train a neural network in four estates of the brazilian northwest, focusing on the Caatinga biome.

In this work, the highest value obtained for the coefficient of determination of the socioeconomic variable GDP per capita (i.e income) was $R^2 = 0.349$.

2.5.4 SOUSA, I. A deep learning approach to predict socioeconomic indicators in Vale do Ribeira from satellite imagery (2022)

This research, written by myself, adapted the article of Yeh et al. (2020) [1] to predict socioeconomic data for the Brazilian scenario, using the Vale do Ribeira region as a case study.

In this first approach, the data collection methodology consisted in obtaining the images centered on the census sectors of the Vale do Ribeira's municipalities, where a census sector is defined as a group of approximately 300 households within a defined area. Therefore, similar to Triñanes et al. (2020) [14], since IBGE publish its data at the municipality level, it was necessary to adapt the income indicator to the right granularity of the model.

The model yielded a low performance of $R^2 = 0.289$, attributed to the considerably small size of the dataset (only 880 images or 4.4% of the dataset used in the original study of Yeh et al.) and to the possibility of overlapping between the images.

This present research has the goal of continuing the work previous developed by attempting to improve the results obtained. In order to do so, a careful analysis of the Vale do Ribeira region is conducted, resulting in a new methodology for the data acquisition, as better explained in section 6.1.2. The overall deep learning methodology in both studies is the same, differing in the details of the execution, such as training with different number of epochs and learning rates, and of course in the results analysis. In addition, the scripts used to develop the model were adapted to work with the most current version of the technologies used in the environment of execution, as outlined in section 3.2.

PART III

TECHNOLOGIES USED

3 TECHNOLOGIES USED

3.1 Satellite Imagery Acquisition

The Google Earth Engine API was used for the collection of the satellite imagery. Google Earth Engine (GEE) [15] is a cloud-based geospatial analysis platform that enables users to visualize and analyze satellite images of our planet. Its public data archive includes more than forty years of historical imagery and scientific datasets, being free for academic and research uses. Two types of dataset were obtained:

- Landsat Surface Reflectance, whose dataset was made available on the platform by the U.S. Geological Survey (USGS);
- Nighttime lights, whose dataset was made available on the platform by Defense Meteorological Satellite Program - Operational Linescan System (DMSP-OLS).

3.2 Data Processing and Visualization

3.2.1 Computational Platform

The computational platform used to develop the project was the Google Colaboratory (Colab), a Jupyter notebook based runtime environment which allows you to run code entirely on the cloud.

Running a stable python version, it facilitates the process of managing different libraries. Furthermore, it provides the allocation of graphics processing units (GPUs) to its notebooks. In this manner, it presents itself as a suitable environment for running machine learning models, avoiding an extensive environment setup in the local machine.

Nevertheless, the free version of Colab has its limitations, only allowing a 12GB of RAM to be allocated for execution. For this project, the Colab Pro version was used since the memory RAM necessary for running the training of the model was approximately

23GB.

3.2.2 Programming Language and Main Libraries

The project was developed in python, a programming language widely used by the academic and professional environment for the collection, processing, and analysis of data. Being a high-level language and providing open-source libraries, it is an effective tool for building the data pipelines necessary for developing machine learning models.

Among the main python libraries used in the project for data processing and visualization, the following can be highlighted:

- Pandas: a library for data manipulation, processing and analysis. In particular, it provides structures and operations for manipulating tabular data;
- Geopandas: an extension of the datatypes used by pandas to allow spatial operations on geometric types;
- Tensorflow: an open source library for machine learning. It is optimized for prototyping and implementing deep learning models, being able to perform calculations using tensors (n-dimensional vectors) with strong acceleration through GPUs;
- Scikit-learn: an open source machine learning library that includes several classification, regression, and clustering algorithms;
- Numpy: an open source library designed to perform operations on multidimensional arrays and matrices. It supplies a large collection of high-level mathematical functions to operate on these data structures;
- Matplotlib: a comprehensive library for fast processing and high-quality graphics generation, used in data visualization and graphical plotting.

Note that the scripts developed by Yeh et al. (2020) [1] were done using tensorflow 1.15 and python 3.7. In this work, the scripts were adapted to the most recent version of tensorflow (2.9.2) and python (3.8) used in Colab.

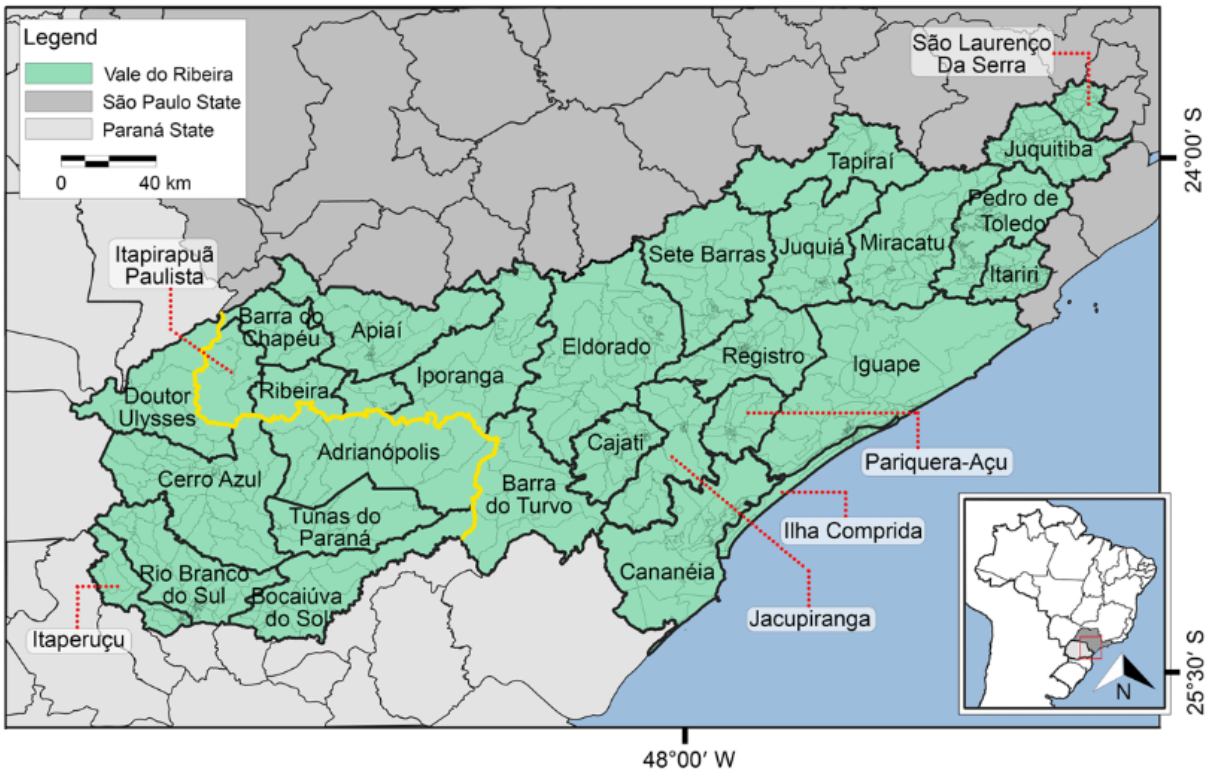
PART IV

CASE STUDY: VALE DO RIBEIRA

4 CASE STUDY: VALE DO RIBEIRA

Vale do Ribeira (VR) was chosen as a case study due to its data availability and its ecological and economic importance. It is located in the southern region of Brazil, encompassing the states of Paraná and São Paulo. With 28,306 km², it has 30 municipalities and it is the largest continuous area of preserved Atlantic Forest in Brazil, being declared as a Natural Heritage of Humanity by UNESCO (United Nations Educational, Scientific and Cultural Organization) in 1999. Finally, even though it has a high economic development, this area is considered economically poor and has the lowest human development index in the state of São Paulo [16], being very interesting to study.

Figure 5: Vale do Ribeira and its municipalities.



Source: Machicao et al. [17]

However, due to the data image collection methodology proposed in this project, the

number of images obtained considering only this region was very small, as explained in section 6.1.2. Therefore, the other cities of São Paulo (SP) and Paraná (PR) were also included in the training of the model, totalizing 1044 municipalities.

PART V

METHODOLOGY

5 OVERALL WORKFLOW

The workflow used in this project to replicate the work of Yeh et al. [1] was divided in four steps:

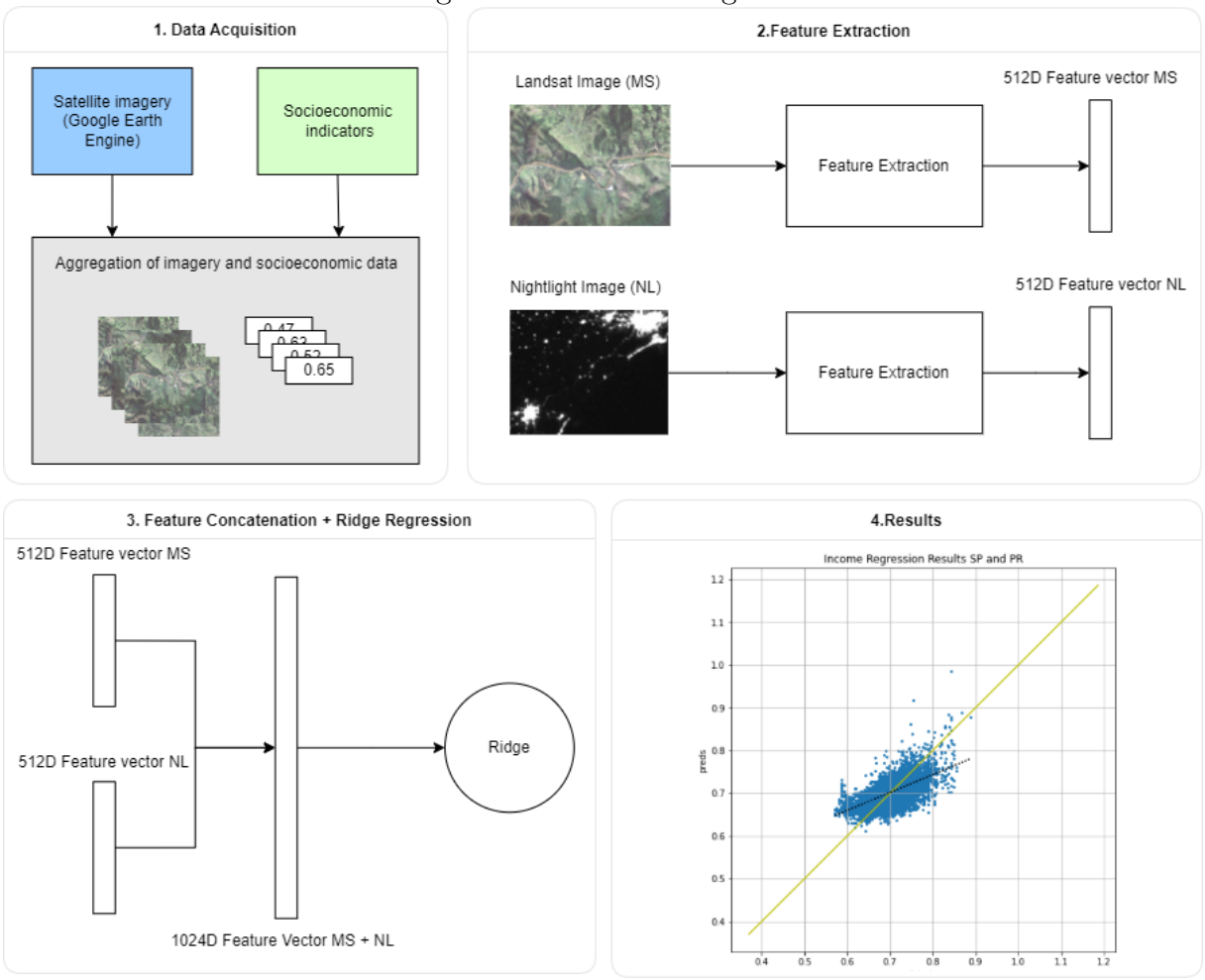
1. Data Acquisition and Aggregation: overlay of a grid over the region of interest, with blocks of approximately 45km² in area. For each block, an average income indicator was calculated as the weighted average of the income of the municipalities, provided by the IBGE census of 2010, and their respective areas contained in the block. Multispectral daytime imagery (MS) and nighttime lights imagery (NL) were obtained through the Google Earth Engine API.

2. Feature extraction: pretrained ResNet-18 networks models were modified to adapt multi-band satellite imagery and used to extract the feature vectors of the images. The loss function used was mean squared error (MSE). The training followed a 5-fold cross-validation.

3. Feature Concatenation and Ridge Regression: concatenation of the feature vectors and refinement of the model using a ridge regression. The loss function used was MSE. The training followed a leave-one-group-out cross-validation.

4. Results analysis, in which performance metrics such as the coefficient of determination (R^2) were calculated and graphs were drawn to better analyze the results.

Figure 6: Workflow diagram.



Source: Author's Compilation.

6 DATA ACQUISITION AND AGGREGATION

6.1 Data Acquisition

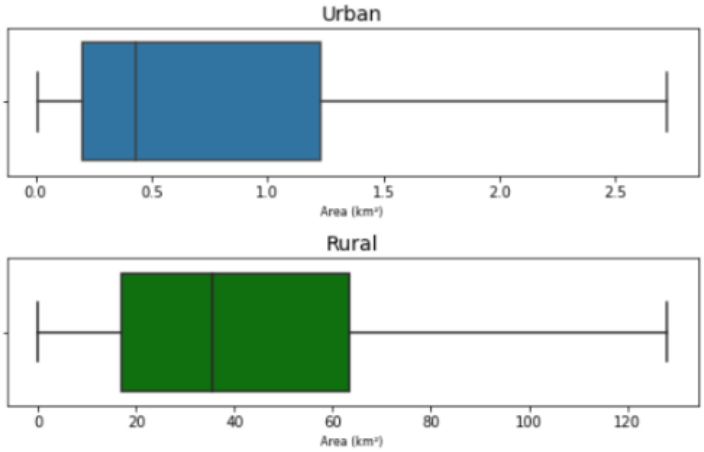
As mentioned in section 2.5.4, in the preliminary study done in the paper SOUSA, I et al. (2022) [18], the data collection methodology consisted of obtaining the images centered on the census sectors of the municipalities of Vale do Ribeira. This approach resulted in a low performance of $R^2 = 0.289$, attributed to the considerably small size of the dataset and the possibility of overlap between the images.

6.1.1 Analysis of the Vale do Ribeira Region

This section aims to describe the data analysis done in the Vale do Ribeira region in order to better analyze whether there is superposition between the images.

The calculations were done using the geometric information of the territories and its rural/urban classification, provided by the IBGE in its official website. With that, the area of each census sector was calculated and the distribution of the census tract areas was obtained.

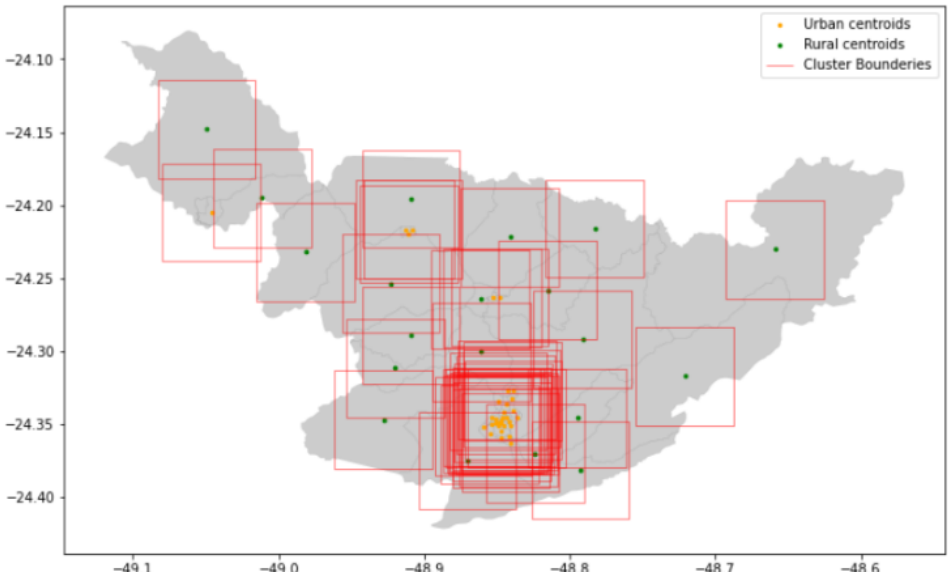
Figure 7: Boxplot of census areas by urban or rural types.



Source: Author's Compilation.

It is possible to visualize that there is a major discrepancy between the size of the urban areas when compared to the size of the rural areas. Due to the proximity of the urban tracts and the area of 45 km² covered by one image in the model's methodology (images resolution of 30m/pixel and CNN with input 224x224 pixels - as explained in section 6.1.4), the clusters selected in the model will probably be composed of multiple urban tracts. This implies that the same sector can be present in several images, harming the interpretability of the results. This overlap between images can be exemplified in Figure 8 and can also happen with less intensity for rural areas.

Figure 8: Cluster boundaries of the municipality of Apiai, SP.



Source: Author's Compilation.

6.1.2 Grid with tiles of 45km²

To overcome the problem of overlapping between images, a new methodology for the imagery collection is proposed. The idea is to overlay of a grid over the region of interest, with blocks of approximately 45km² in area. To adapt the socioeconomic indicator of income (see section 6.1.3) to the new methodology, for each block an average income indicator was calculated as the weighted average of the income of the municipalities and their respective areas contained in the block.

With this, the total number of images for the Vale do Ribeira region was approximately of 500 images, which is very low. Therefore, to increase the number of the images, the other municipalities of SP and PR were also included in the training of the model, totalizing 9748 images.

Note that Albers Equal-Area Conic was chosen as the reference map projection for the grid. The Albers Equal-Area Conic projection parameters follow the cartographic standards defined by IBGE, with SIRGAS-2000, the official planimetric datum of Brazil, as the reference geodetic system. It is recommended for equal-area mapping, being best suited for land masses extending in an east-to-west orientation at mid-latitudes [19]. In this way, the blocks were automatically generated with the same corresponding sizes, always defined by Albers Equal-Area Conic plane coordinates.

6.1.3 Socioeconomic Indicators

The socioeconomic proxies utilized in the model were obtained using data gathered from IBGE from the 2010 census, the latest census available.

The IBGE and Atlas Brazil [20] publish a municipal Human Development Index (HDI) every decade based on the census data and other surveys. The HDI is a statistical index composed of income, longevity, and education indicators. It was developed and compiled by the United Nations to measure the various levels of social and economic development of countries. In this work, we focus only on the income indicator of this index at the municipal level, which can be found here.

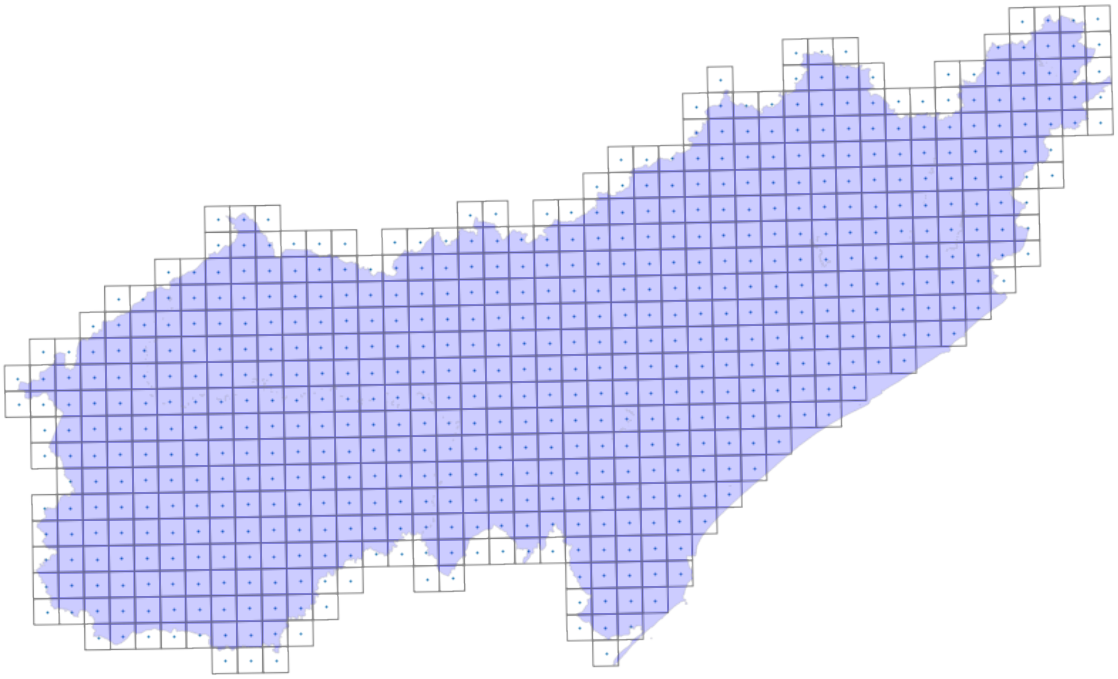
Note that since the number of images gathered only depends of the spacial extend of the area of interest, there is no longer the need of developing the model at the census sectors level. Instead, the municipality level is used, eliminating the necessity to adapt the indicators to a different granularity than the one provided by IBGE.

6.1.4 Satellite Imagery

The satellite imagery chosen was multispectral daytime imagery from 30m/pixel Landsat and <1 km/pixel nighttime lights imagery. The motivation for choosing nighttime lights imagery was that earlier studies demonstrated that they can serve as an indicator of economic activities at night [21] and, when this indicator is compared across regions and over time, it can be used to measure economic performance. Regarding the choice of daytime imagery, earlier works demonstrated that high-resolution (<1 m/pixel) imagery from private-sector providers can be used to measure spatial variation in local economic outcomes in several developing and middle-income countries. Nevertheless, high-resolution imagery remains very expensive nowadays, so the idea is to focus on more coarser public imagery that can be found for free.

The methodology adopted by Yeh et al. (2020) [1] focuses on obtaining the Landsat surface reflectance and nightlights images centered on each cluster location. The same is done in this project, in which the clusters are the cells of the grid that was overlayed in the regions of SP and PR.

Figure 9: Grid zoomed to the Vale do Ribeira region.



Source: Author's Compilation.

Following the methodology of Yeh et al. (2020) [1], it was obtained a 3-year median composite for the Landsat surface reflectance. This composite was created by taking the median of each cloud free pixel available during the period of 3 years, which in this case is 2009 to 2011, since the Brazilian census survey was done in 2010. As the author's describe in their original paper, the motivation for using three-year composites was two-fold. First, multi-year median compositing has seen success in similar applications as a method to gather clear satellite imagery. Second, the outcome we are trying to predict tends to evolve slowly over time, and we similarly wanted the inputs to not be distorted by seasonal or short-run variation.

The Landsat surface reflectance imagery was captured by the Landsat 5 and Landsat 7 satellites, with seven bands which we refer to as the multispectral (MS) bands: RED, GREEN, BLUE, NIR (Near Infrared), SWIR1 (Shortwave Infrared 1), SWIR2 (Shortwave Infrared 2), and TEMP1 (Thermal), all with a spatial resolution of 30 m/pixel. Those are all the surface reflectance bands available for both the satellites in question. See Table 1

for a description of the bands mentioned.

Name	Units	Wavelength	Description
B1		0.45-0.52 µm	Band 1 (blue) surface reflectance
B2		0.52-0.60 µm	Band 2 (green) surface reflectance
B3		0.63-0.69 µm	Band 3 (red) surface reflectance
B4		0.77-0.90 µm	Band 4 (near infrared) surface reflectance
B5		1.55-1.75 µm	Band 5 (shortwave infrared 1) surface reflectance
B6	Kelvin	10.40-12.50 µm	Band 6 surface temperature.
B7		2.08-2.35 µm	Band 7 (shortwave infrared 2) surface reflectance

Table 1: Description of Landsat 5 and 7 surface reflectance bands.

Note that in the paper of Yeh et al., Landsat 8 was also used since images taken from these satellites have been available since 2013 and in the original study the surveys were from several years (2009 to 2017).

For comparability, it was also created 3-year median composites for the nightlights imagery (NL). Only DMSP imagery was used, since the first year of availability of the VIIRS dataset at GEE is 2014. The main difference here is that the authors used DMSP and VIIRS because no single satellite captured nightlights for all that period. Note that the nightlight imagery is much more coarse than the daytime imagery we acquired, with a resolution of 927.67 meters. The images are resized using nearest-neighbor resampling to cover the same spatial area as the Landsat images, this is done by default by GEE during reprojection.

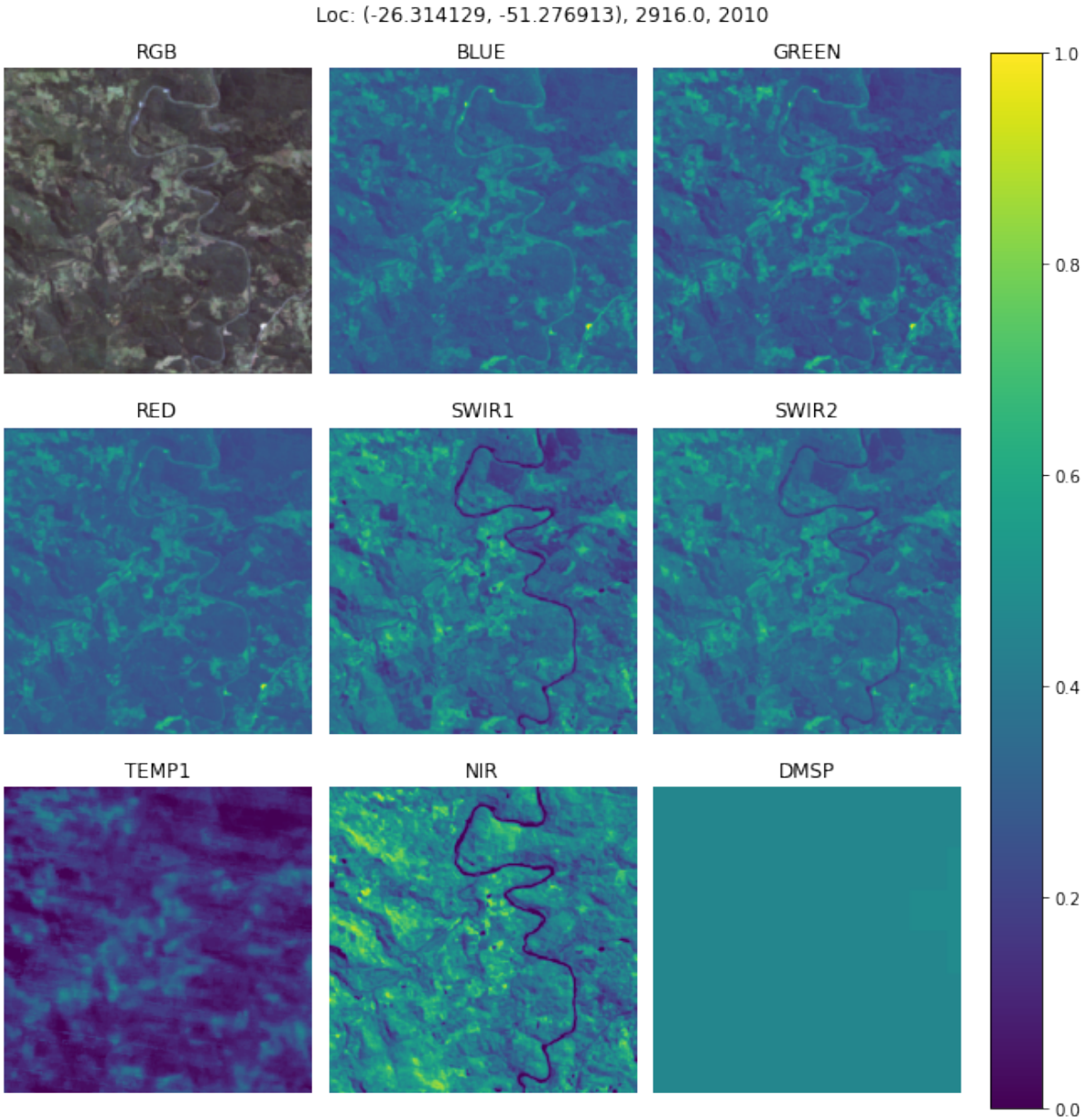
Finally, both MS and NL images were processed in and exported from GEE in 255 × 255 tiles, with a 30m/pixel resolution, then at the training stage center-cropped to 224 × 224 (the input size of the convolutional neural network architecture), spanning 6.72 km on each side (30 m Landsat pixel size × 224px). Note that the reason why 255 x 255 tiles are used is to have the flexibility of using 'random crops' as a form of data augmentation.

6.2 Data Aggregation

The files were exported as TFRecords, a simple format for storing a sequence of binary records, including all the bands (MS + NL) but as well other relevant data such as the values of the socioeconomic data labels. Inspecting those records, I was able to obtain

the Figure 10, in which it is possible to see the bands plotted for the cell number 2916, centered at the coordinates (26.314129° S, 51.276913° W). Note that each band highlights different aspects of the image, being useful to identify different features. The red, blue and green (RGB) all together create a true color band combination. The goal of displaying that combination in the figure was to create a visual map of the area that is intuitive to the human eye.

Figure 10: Bands plotted for block number 2916, centered at the coordinates (26.314129° S, 51.276913° W). The color map refers to the normalized pixel values.



Source: Author's Compilation.

7 DEEP LEARNING METHODOLOGY

The goal of this project is to train a convolutional neural network (CNN) to predict socioeconomic indicators using a combined model that incorporates both multispectral daytime imagery and nighttime lights in a deep learning model trained end-to-end. For this purpose, pretrained ResNet-18 models were modified and trained separately on the Landsat bands and nightlights bands. Then, their feature vectors were fed to the fully connected layer of the Resnet-18 and a ridge regression was done on top for fine-tuning.

7.1 Feature Extraction

7.1.1 ResNet-18 Modifications

Residual Network (ResNet) is a convolution neural network that was introduced in 2015 by Kaiming He et Al. [8], being the winner of the ILSVRC Imagenet 2015 competition and becoming quite popular for image recognition and classification tasks because of its ability to solve the vanishing gradient problem.

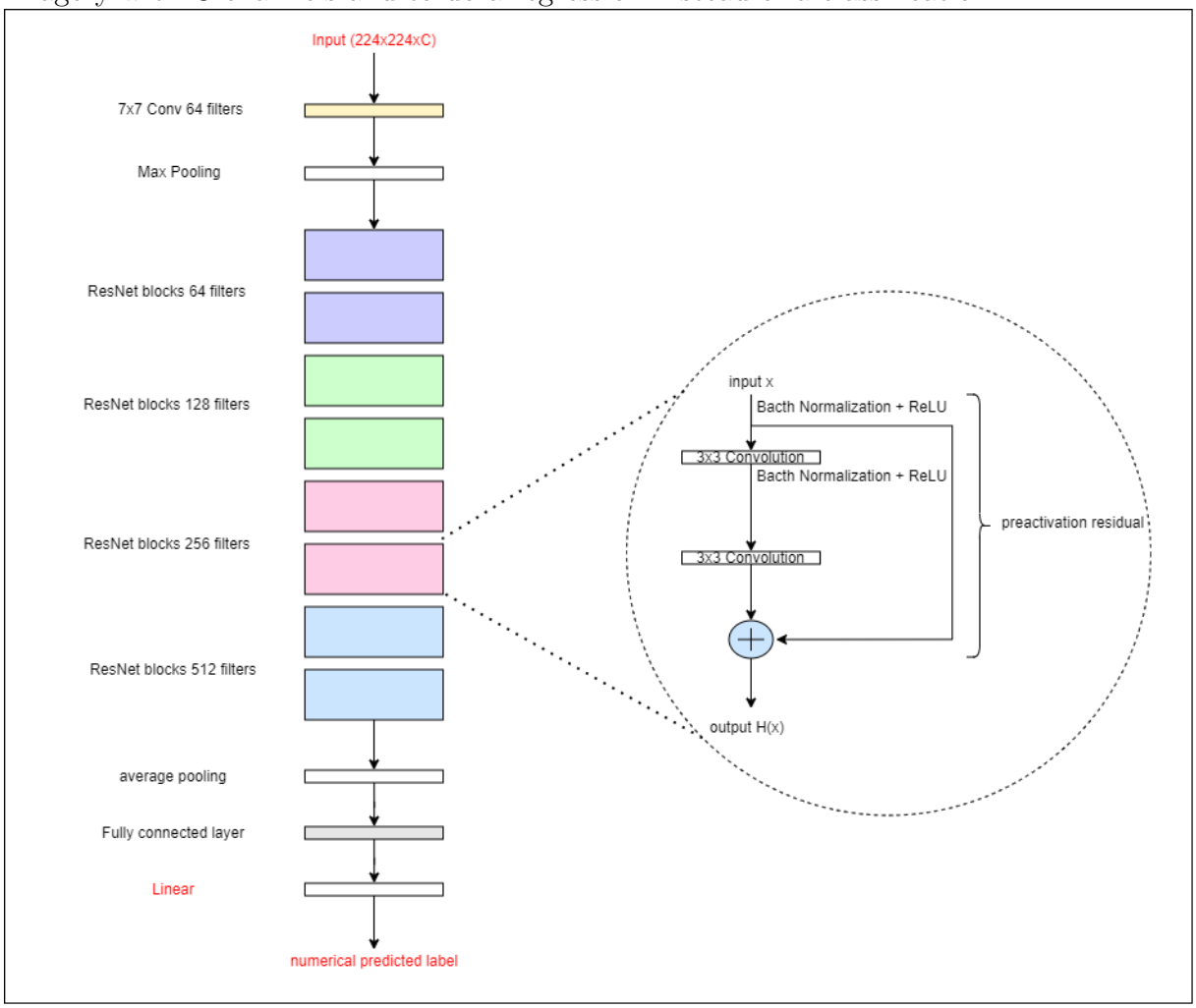
As in the original article, the CNN models were trained using the ResNet-18 architecture with preactivation [8]. Since the ResNet, as most existing CNN models, is designed to work with 3-channel RGB images, the first convolutional layer was modified to accommodate multi-band satellite imagery. This is actually simple to visualize: basically when using C channels, where C is the number of channels, the filters of the first convolutional layer will have depth of C instead of 3, which means that instead of [F, F, 3, 64], the dimensions of the weights become [F, F, C, 64], where F is the filter size of the first convolution layer.

As in the original paper, the weight of the layers were initialized with the pre-trained value of ImageNet [22] with exception of the first and last layers. In the first layer, the weights for the RGB bands are also initialized as the pre-trained values but the weights for the non-RGB bands are initialized as the average of the weights from the RGB channels

and normalized by a factor of 3/C. For the final layer, the weights are initialized randomly from a standard normal distribution truncated at ±2. Finally, for the models trained only on the nightlights bands, the first layer weights were initialized randomly using He's initialization [8] (Gaussian with same overall mean and standard deviation as the RGB channels).

Furthermore, the fully connected layer of the CNN is adapted to do a regression instead of a classification. See Figure 11 for an overall representation of the ResNet18 modified.

Figure 11: ResNet18 with preactivation adapted to accept as input multi-band satellite imagery with C channels and to do a regression instead of a classification.



Source: Author's Compilation

7.1.2 Training

For each of the input band combinations (MS, NL), five separate models were trained using a 5-fold cross-validation procedure. To do so, the data was divided into 5 folds of roughly equal size (same number of clusters). See Table 2 to visualize the folds used for the training. Each model was then trained on 3-folds, validated on a 4th, and tested on a 5th. See Table 3 to visualize the splits used in the cross-validation procedure. Note that to keep the same algorithm used in the experiment of Yeh et al.(2020) [1], where the clusters were aggregated by countries, the cells were divided into 30 groups.

Fold	Groups	clusters
A	group1, group6, group11, group16, group21, group26	1944
B	group2, group7, group12, group17, group22, group27	1944
C	group3, group8, group13, group18, group23, group28	1944
D	group4, group9, group14, group19, group24, group29	1944
E	group5, group10, group15, group20, group25, group30	1972

Table 2: Folds used for training.

Model	Train	Val	Test
1	C, D, E	B	A
2	A, D, E	C	B
3	A, B, E	D	C
4	A, B, C	E	D
5	B, C, D	A	E

Table 3: Split of the 5 folds used for all cross-validated training.

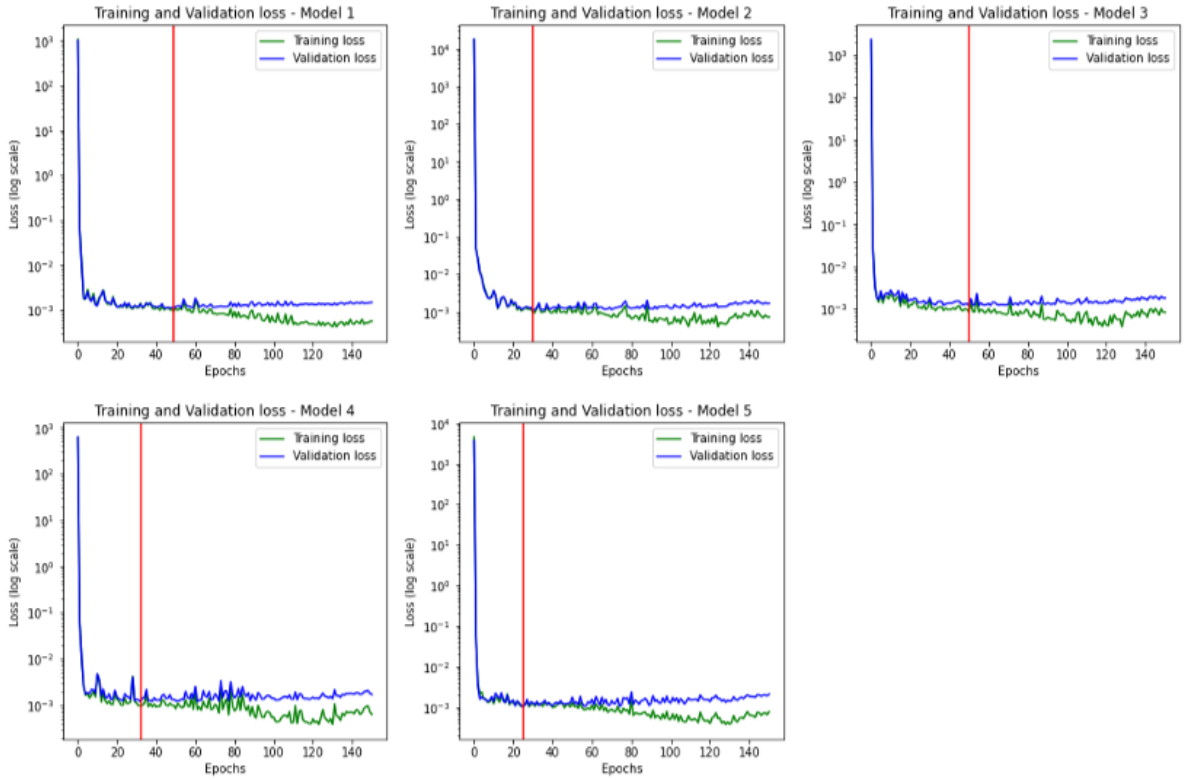
The ResNet-18 models were trained with the Adam optimizer and MSE loss function. The batch size is 64 and the learning rate is decayed by a factor of 0.96 after each epoch. To improve the performance and ability of the model to generalize, data augmentation - random flips and random brightness/contrast on the MS images - and shuffling techniques were used.

The MS models were trained with 150 epochs and a learning rate of 10^{-4} and the NL models were trained with 200 epochs and a learning rate of 10^{-5}. Note that the validation

loss is being used as a metric for early stopping: a checkpoint of the model was saved every time a lowest MSE was achieved.

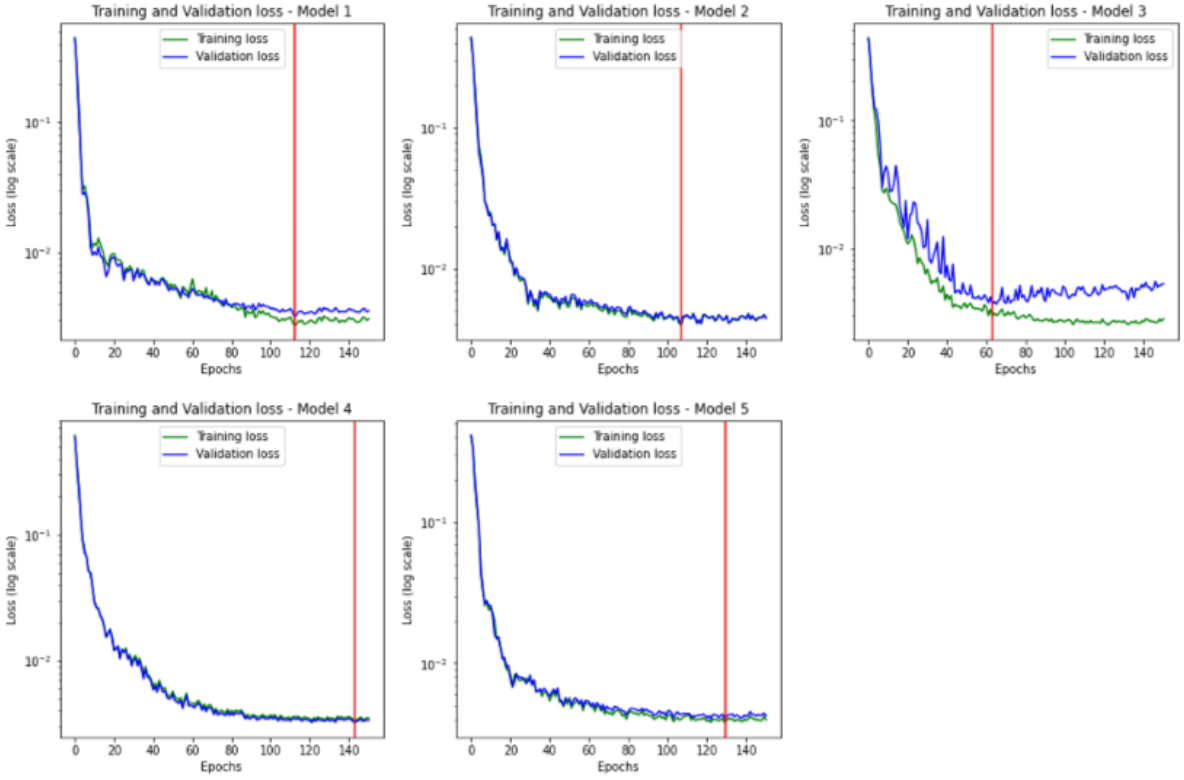
See the figures below for an illustration of the final training and validation loss of each model of the MS and NL bands, respectively.

Figure 12: Training and validation curves for MS models. The red lines represent the checkpoints where the model obtained the lowest MSE.



Author's Compilation.

Figure 13: Training and validation curves for NL models. The red lines represent the checkpoints where the model obtained the lowest MSE.



Author's Compilation.

7.2 Features Concatenation

After training the MS and NL models, their feature maps were extracted and saved as a compressed numpy .npz file. The feature maps contain the pertinent features to classify the images and, after being flattened, they serve as input for the fully connected layers. In the case of the ResNet18, the feature map of each image has a dimension of 1x1x512. To form the combined MS + NL models, the MS and NL feature vectors are concatenated resulting in feature vectors of dimension 1024.

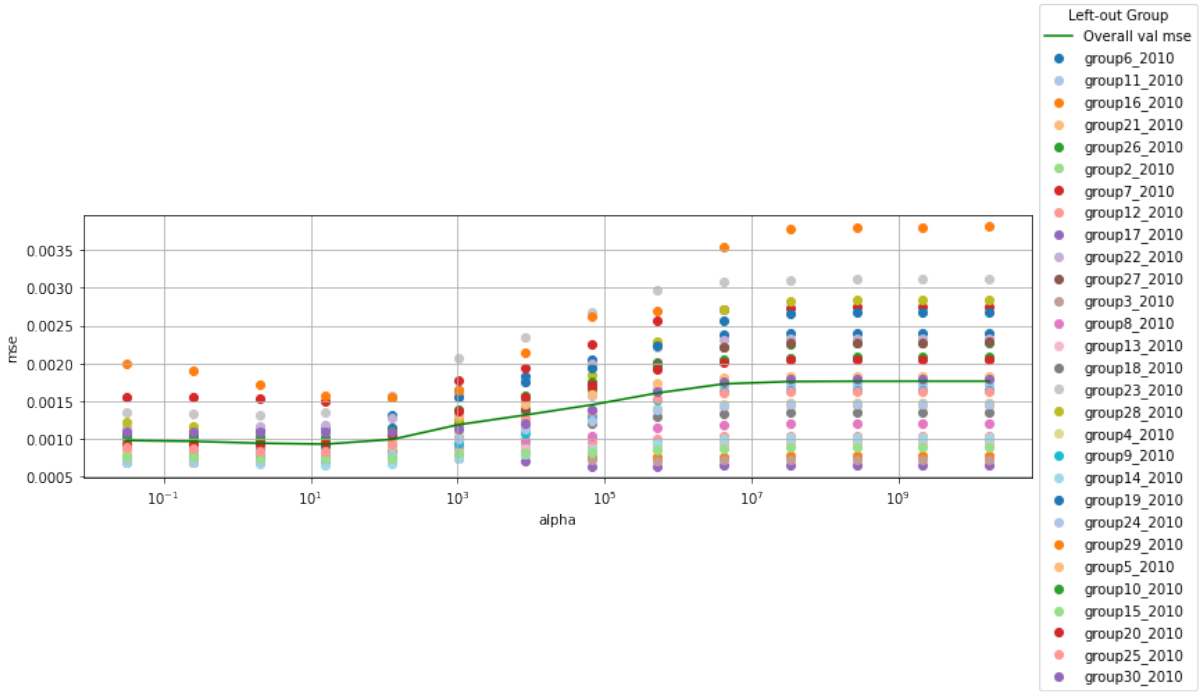
7.3 Ridge Regression

Once the combined feature vectors were obtained, the last fully connected layer of the models was fine-tuned using ridge regression with leave-one-group-out cross-validation.

The L2 norm regularization parameter alpha of the regression is chosen during the cross-validation procedure: the best alpha is the one that yields the lowest mean squared

error between fourteen values with order of magnitude ranging from 10^{-2} to 10^{9}.

Figure 14: Example of leave-one-group-out cross-validation for the concatenated model MS+NL. In this case, lowest MSE = 0.001 and best alpha = 16



Author's Compilation.

8 RESULTS

To analyze the results, the principal metric chosen was the coefficient of determination (R^2). As mentioned in the section 2.4, this coefficient is used to identify the strength of the model, it captures how well the predictions match the observations, or how much of the variation in the observed data is explained by the predictions. Usually the R^2 varies between 0 and 1, being expressed as a percentage (the closer to 1, the better). This metric is very used to measure the performance of a regression method due to its facility of interpretation, being usually more informative than error metrics that have arbitrary ranges, such as MSE.

To evaluate the performance in the same manner as Yeh et al. (2020) [1], the squared Pearson correlation coefficient (r^2), the Spearman's rank correlation coefficient (rank) and the MSE were also used. The r^2 is powerful to find patterns and relationships in data and the closer to 1, the higher the correlation between the variables. While the squared Pearson's correlation (r^2) assesses linear relationships, Spearman's rank correlation assesses monotonic relationships (whether linear or not).

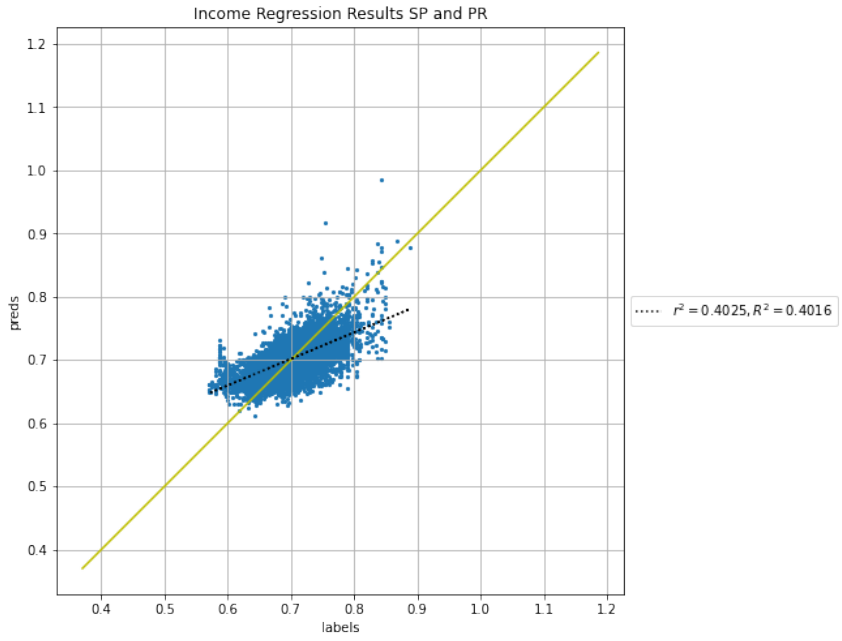
Furthermore, heatmaps with the values of the real and estimated data were made to provide a visual aid of the results.

8.1 Model

The coefficient of determination between the observed and predicted values was $R^2 = 0.4016$. Compared to the article of reference of Yeh et al. (2020) [1], where the authors obtained the best case as $R^2 = 0.70$, the result is not ideal. Nevertheless, from the plot produced it is possible to visualize that the real and predicted values are proportional. Furthermore, the value obtained is higher than the one found by Triñanes et al. (2020) [14] ($R^2 = 0.35$) and than the values found by Tsuru et al. (2021) [23] for the GDP per capita of the states of Alagoas, Paraiba, Rio Grande do Norte and Sergipe.

In addition, compared to the preliminary results obtained in my previous study [18], the performance improved significantly and, with the new data collection methodology, the problem of overlapping between the images has been solved. Therefore, it can be concluded that the algorithm is very promising in the task of predicting socioeconomic indicators from satellite images for the municipalities of SP and PR.

Figure 15: Regression plot for the combined MS+NL model. The dotted line corresponds to the line of best fit.



Source: Author's Compilation.

For comparison, I also ran the ridge regression for the separated MS and NL models, and redid the whole training procedure considering just the RGB bands. See the metrics results in the Figure below.

Figure 16: Final metrics results.

	r2	R2	mse	rank
MS+NL	0.402498	0.401582	0.001038	0.623075
MS	0.361431	0.359598	0.001110	0.579774
NL	0.252676	0.251166	0.001298	0.437445
RGB	0.229353	0.227948	0.001339	0.430193

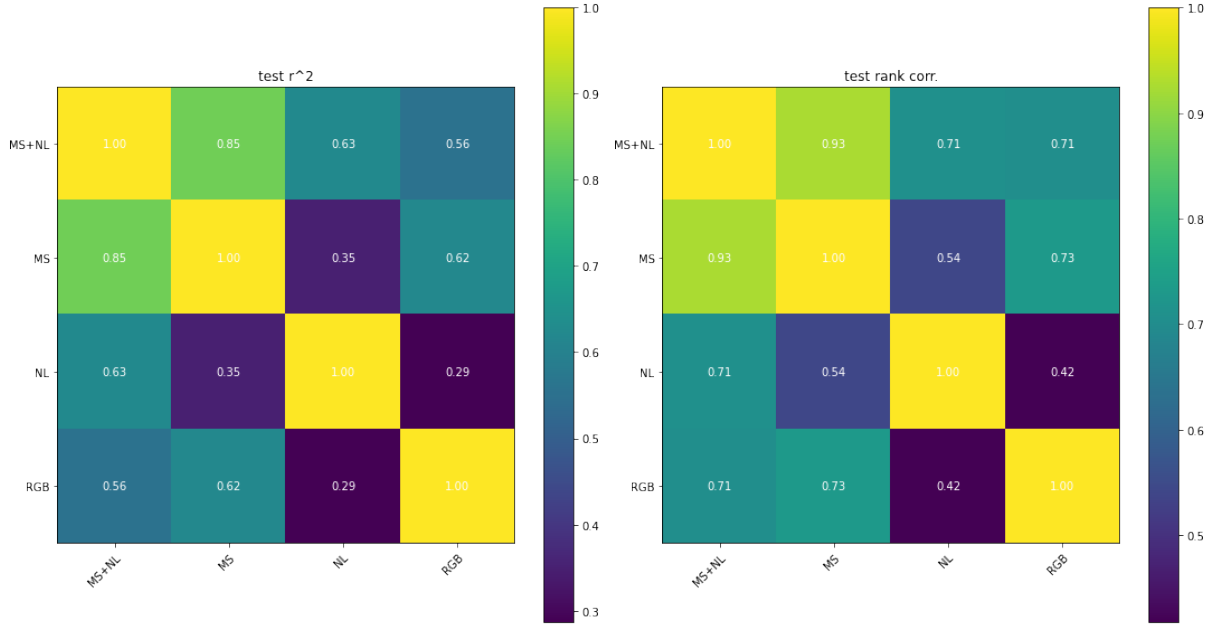
Source: Author's Compilation.

From this, it is possible to draw some conclusions:

- The combined model MS + NL outperformed the other models in all metrics, proving itself as the best model for this task;
- The CNN trained on the RGB bands performed poorly compared to the combined MS + NL model, which is expected and confirms the relevance of using multispectral daytime imagery and nightlight imagery when predicting socioeconomic data.

Furthermore, for complementing the analysis, the correlation matrix for the results between the models were drawn.

Figure 17: Correlation results for the income predictions of all models a) comparison of r^2 metric and b) comparison of rank metric



Source: Author's compilation.

This allows us to make the following conclusions:

- There is a high correlation between the MS and the RGB models, but the MS outperformed the RGB, which again is expected since the MS model includes more bands than the RGB one;
- Even though the NL and RGB models have a similiar perfomance, the correlation between then is low ($r^2 = 0.29$), suggesting that they are good in predicting different features.

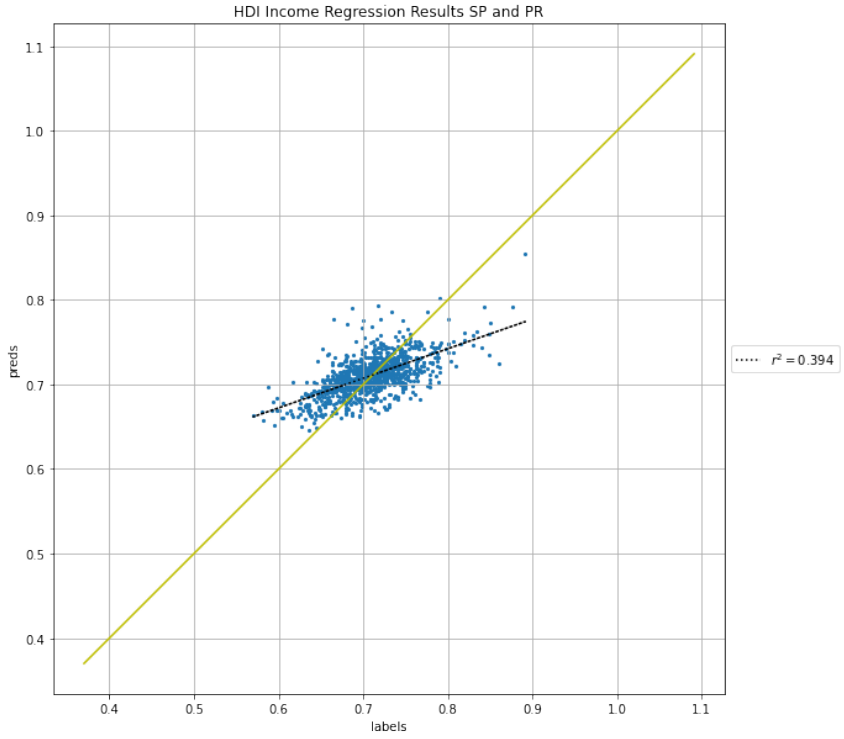
8.2 Results after calculating the income for each municipality

In this section the regression results are analyzed considering the actual values as the income indexes provided by IBGE and the predicted ones as the indexes calculated using weighted average and the values obtained by the model.

8.2.1 São Paulo and Paraná

8.2.1.1 Performance

Figure 18: Regression plot after calculating the income for each municipality. The dotted line corresponds to the line of best fit.

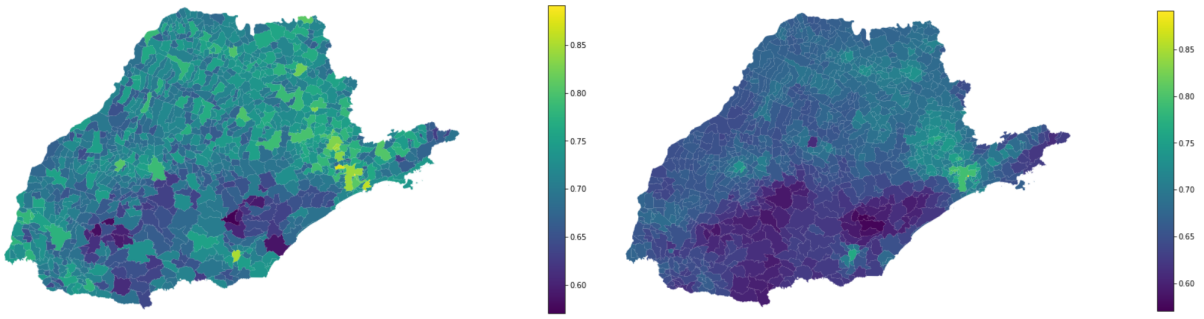


Source: Author's Compilation.

The coefficient of determination for the municipalities of SP and PR is $R^2 = 0.394$. This result is very similar to the one obtained for the model without any recalculations, which leads to the conclusion the methodology adopted for the calculation of the grid's cells average income is well-suited for this project.

8.2.1.2 Visual Analysis

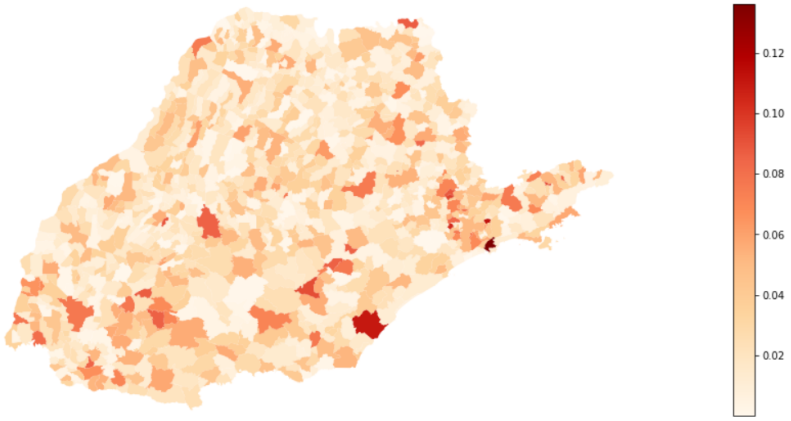
Figure 19: Heatmaps for a) Real HDI Income indicator and b) Predicted HDI Income indicator.



Source: Author's compilation.

To complement this visual analysis, the difference between the real income score y and the predicted ŷ value, defined as $d = \|y - \hat{y}\|$, was also calculated and plotted in the heatmap. From this figure it is possible to visualize that the majority of the municipalities yield a small difference between the real and estimated values, for instance 87,55% of them obtained d equal or lowest to 0.05.

Figure 20: Differences between the real and predicted labels for the municipalities of SP and PR.

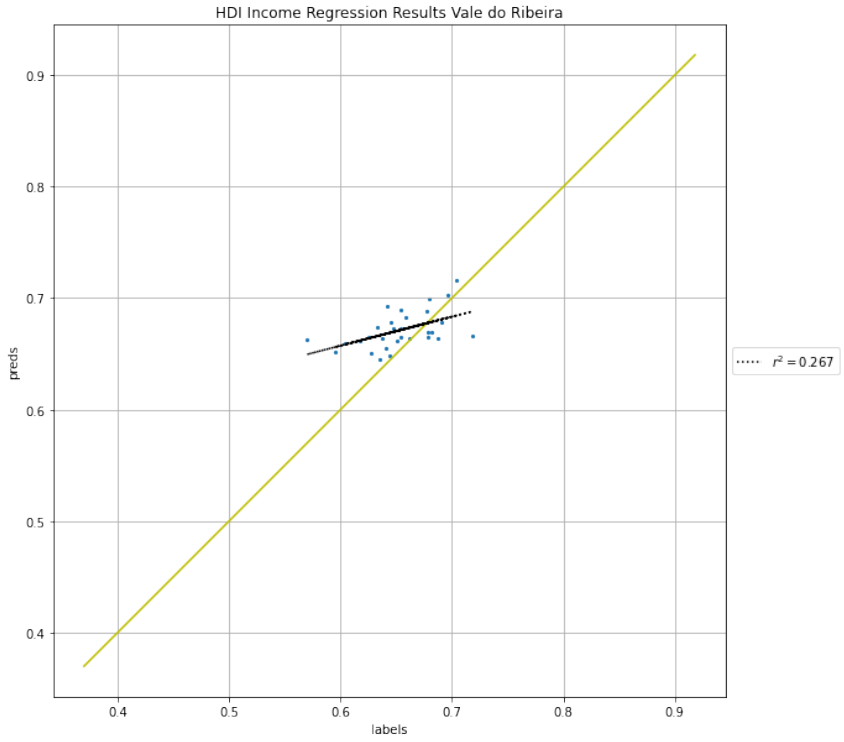


Source: Author's Compilation.

8.2.2 Vale do Ribeira

8.2.2.1 Performance

Figure 21: Regression plot for the municipalities of Vale do Ribeira. The dotted line corresponds to the line of best fit.

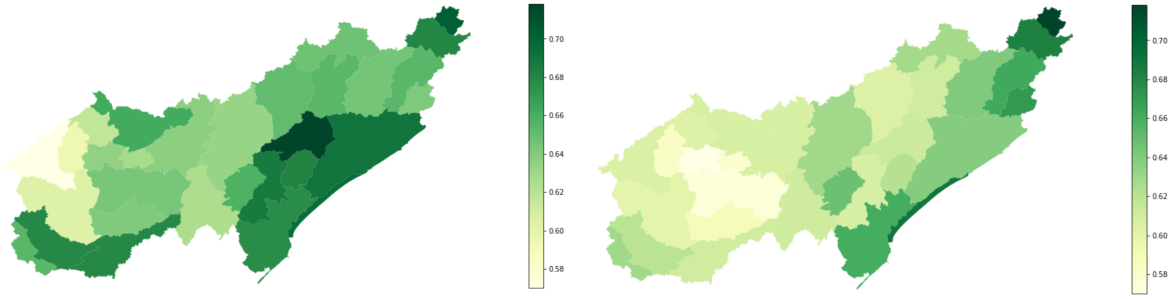


Source: Author's Compilation.

The coefficient of determination for the municipalities of Vale do Ribeira is $R^2 = 0.267$. This result is significantly smaller than the one obtained considering all the municipalities. This is expected and can be justified due to the small number of samples: only 30 municipalities are part of the Vale do Ribeira region.

8.2.2.2 Visual Analysis

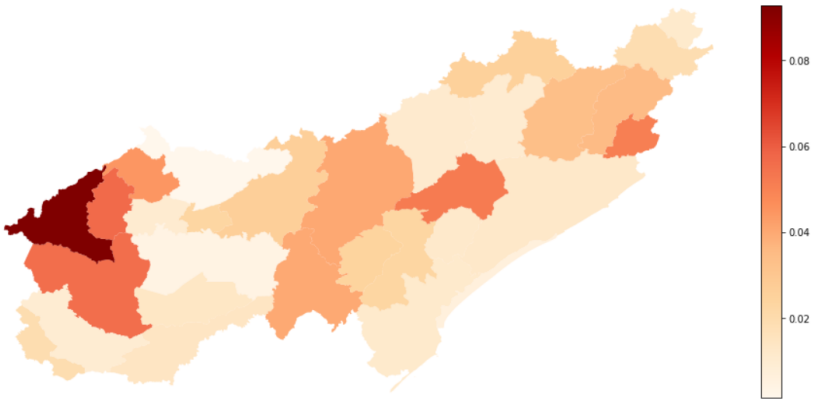
Figure 22: Heatmaps for a) Real HDI Income indicator for VR and b) Predicted HDI Income indicator for VR.



Source: Author's compilation.

To complement this visual analysis, the difference between the real income score y and the predicted ŷ value, defined as $d = \|y - \hat{y}\|$, was also calculated and plotted in the heatmap. From this figure it is possible to visualize that the majority of the municipalities yield a small difference between the real and estimated values, for instance 73,33% of them obtained d equal or lowest to 0.05.

Figure 23: Differences between the real and predicted labels for the municipalities from VR.



Source: Author's Compilation.

PART VI

CONCLUSION

9 CONCLUSION

In this report, public satellite imagery was used to train a combined CNN model to estimate socioeconomic data over space in the states of São Paulo and Paraná, with the goal of particularly analyzing the Vale do Ribeira region due to its economic and environmental importance.

With an R^2 of 0.4016, the model yields a low performance when compared to the one of Yeh et al. (2020) [1], which attempts to replicate ($R^2 = 0.70$) . Nevertheless, it is still promising since the actual and predicted values are clearly proportional. Moreover, when compared to the best results of other studies that have analyzed socioeconomic data using machine learning and satellite imagery in Brazil, it presented a higher R^2 value by approximately 0.05. Finally, it successfully satisfies the goal of improving the performance of $R^2 = 0.289$ obtained in my preliminary study (SOUSA, I et al., 2022) [18].

In conclusion, although not accurate enough to replace field survey data, the model performed well and can be used to help analyze the socioeconomic situation of the Brazilian territory in years when official government survey data are not published, supporting the planning and evaluation of public policies. Furthermore, it is a step towards understanding how convolutional neural networks, daytime and nighttime multispectral imagery can be used in the task of forecasting socioeconomic data, and thus a stimulus for the development of future work in the field of predicting data through satellite imagery.

REFERENCES

[1] Y. et al., "Using publicly available satellite imagery and deep learning to understand economic well-being in africa," *Nat Commun*, vol. 11, no. 2583, 2020. [Online]. Available: https://doi.org/10.1038/s41467-020-16185-w

[2] PARSEC. Building New Tools for Data Sharing and Reuse through a Transnational Investigation of the Socioeconomic Impacts of Protected Areas. [Online]. Available: ⟨http://parsecproject.org/⟩

[3] PATTERSON, J.; GIBSON, A. Deep Learning: A Practitioner's Approach. 1st ed. O'Reilly Media, Inc., 2017. 538 p.

[4] I. C. Education, "What is deep learning?" 2020.

[5] HASTIE, T.; FRIEDMAN, J; TIBSHIRANI, R. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer, 2009. 767 p.

[6] JAMES, G. et al. An Introduction to Statistical Learning with Applications in R. 1st ed. Springer, 2013. 440 p.

[7] S. Cheusheva, "How to do spearman correlation in excel." 2022. [Online]. Available: https://www.ablebits.com/office-addins-blog/spearman-rank-correlation-excel/

[8] HE, K.; ZHANG, X.; REN, S.; SUN, J. (2015). Delving deep into rectifiers: Surpassing human level performance on imagenet classification. In Proc. 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, pp. 1026–1034 (IEEE Computer Society, Washington, 2015). [Online]. Available: https://doi.org/10.1109/ICCV.2015.123

[9] The demographic and health survey program (DHS). [Online]. Available: https://dhsprogram.com/

[10] The world bank. Living Standards measure study (LSMS). [Online]. Available: https://www.worldbank.org/en/programs/lsms

[11] JEAN, N. et al. Combining satellite imagery and machine learning to predict poverty. Science 19 Aug 2016: Vol. 353, Issue 6301, pp. 790-794. [Online]. Available: https://doi.org/10.1126/science.aaf7894

[12] SIMONYAN, K.; ZISSERMAN, A. Very Deep Convolutional Networks for Large-Scale Image Recognition (2015). [Online]. Available: https://arxiv.org/abs/1409.1556

[13] ABREU, M.; OLIVEIRA, J.; ANDRADE, V.; MEIRA, A. (2021). Methodological proposal for spatial calculation and analysis of the intra-urban HDI of Viçosa, Brazil. Revista Brasileira de Estudos de População. 28, pp 169-186.

[14] TRINANES, E.; MACHICAO, J.; CORREA, P. (2020). Application of a deep learning algorithm for predicting socioeconomic data through satellite images in the Vale do Ribeira. [Online]. Available: https://doi.org/10.5281/zenodo.4712815

[15] "Google earth engine." [Online]. Available: https://earthengine.google.com/

[16] BUENO, G. W.; LEONARDO, A. F. G.; MACHADO, L. P.; BRANDE, M. R.; GODOY, E. M.; DAVID, F. S. (2020). Indicadores de sustentabilidade socioambiental de pisciculturas familiares em área de Mata Atlântica, no Vale do Ribeira - SP. Arquivo Brasileiro de Medicina Veterinária e Zootecnia, 72(3), pp. 901-910. [Online]. Available: https://doi.org/10.1590/1678-4162-11389

[17] Machicao, J., Specht, A., Vellenich, D., Meneguzzi, L., David, R., Stall, S., Ferraz, K., Mabile, L., O'Brien, M. and Corrêa, P., 2022. A Deep-Learning Method for the Prediction of Socio-Economic Indicators from Street-View Imagery Using a Case Study from Brazil. Data Science Journal, 21(1), p.6. [Online]. Available: http://doi.org/10.5334/dsj-2022-006

[18] SOUSA, I. A deep learning approach to predict socioeconomic indicators in vale do ribeira from satellite imagery. [Online]. Available: https://zenodo.org/record/6366429#.YxZC13bMK3B

[19] J. P Snyder. Map projections: a working manual. Technical Report 1395, USGS, Washington, D.C., 1987. [Online]. Available: doi:10.3133/pp1395

[20] PNUD, "Índice de desenvolvimento humano municipal - idhm: Metodologia." 2012. [Online]. Available: https://atlasbrasil.org.br/acervo/atlas

[21] Pinkovskiy M., Sala-i-Martin X. (2016), Lights, Camera . . . Income! Illuminating the National Accounts-Household Surveys Debate , The Quarterly Journal of Economics 131 (2), pp 579–631. [Online]. Available: https://doi.org/10.1093/qje/qjw003

[22] IMAGENET. What is ImageNet. [Online]. Available: image-net.org

[23] Megumi Tsuru, Samuel Vieira Ducca, Vitor Dias Souza. Online Platform for inference of indicator data (2021). [Online]. Available: https://pcs.usp.br/pcspf/wp-content/uploads/sites/8/2021/12/Monografia_PCS3860_COOP_2021_Grupo_C07.pdf

[24] GOOGLE. Earth Engine Apps. [Online]. Available: https://www.earthengine.app.

[25] DATAPANE. [Online]. Available: https://datapane.com/