Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Double Bachelor's Economics and Econometrics

Transfer Learning with CNNs: Predicting Income per Capita and Population Levels in England at Fine Spatial Scales using Satellite Imagery and US-trained Models

Michael van der Merwe (525501)

Abstract

Deep learning has proven to be useful for predicting economic variables at fine spatial scales using satellite imagery and can be used to predict data that is challenging to obtain, coarse, or released infrequently. However, training and tuning these models have substantial computational and time requirements. If transfer learning could be effectively applied to these deep learning methods, this would allow various stakeholders to obtain accurate predictions without the need for training models themselves. This paper explores this possibility, where predictions of income per capita and population levels at fine spatial scales of England are made using satellite images and a convolutional neural network trained on US data. The predictions achieve negative R^2 values, however, this paper further employs a robust model stacking approach using an ordinary least squares regression, which results in R^2 values of 0.051 and 0.460 for income per capita and population levels respectively. Furthermore, this paper finds that the predictions are heteroskedastic, where the residuals increase as the actual values increase. The implications of these results are that transfer learning has the potential to be successfully applied to deep learning methods that predict economic variables, however, further research is required in finding improved model stacking approaches and more adaptable deep learning methods for predicting economic variables in different geographic contexts.

Supervisor:	Andrea Naghi
Second assessor:	Stan Koobs
Date final version:	18 June 2023

1 Introduction

Information on economic variables such as Gross Domestic Product (GDP), inflation, or unemployment is of crucial importance for various stakeholders. Accurate measures of these variables facilitate informed decision-making across different sectors. Governments for instance can utilise this information to predict economic trends and subsequently implement appropriate monetary and/or fiscal policies. When an economic slowdown is predicted, policymakers can employ measures such as lowering interest rates or increasing government spending to stimulate the economy. Similarly, businesses use economic data to mitigate risks associated with economic fluctuations. By monitoring variables such as consumer spending, businesses can make pre-emptive adjustments to inventory levels or operational costs in preparation for a potential decrease in revenue. Moreover, investors rely on economic data to make informed capital allocation decisions. For example, the anticipation of a decrease in interest rates might prompt investors to shift their investments from bond markets to equity markets.

In addition to practical applications, economic data is crucial for academic research. It plays a vital role in identifying patterns and relationships between variables, which contributes to the development of economic theories, as well as empirically evaluating existing theories, determining their validity. For instance, Holly and Jones (1997) hypothesized that real house prices are cointegrated with real income using a dataset from 1939 to 1994 of the UK. However, Gallin (2006) using a more extensive dataset consisting of panel data of 95 metro areas over 23 years found was not able to reject the hypothesis of no cointegration.

Though economic data is crucial, obtaining this data is often challenging due to practical limitations such as accessibility, costs, and time constraints. This leads to some economic data being released infrequently, coarse, or unavailable altogether. In these cases, predictions can be used as a substitute for actual measurements. Predictions are particularly useful in situations where there is a lack of up-to-date or reliable data, such as in developing countries. Predictions could provide valuable insights into economic conditions and trends, even when actual measurements are not available.

Two such economic variables where data is released infrequently are income per capita and population levels at fine spatial scales. This data mostly forms part of census data, which is information collected through a census, a comprehensive survey conducted to gather demographic, social, economic, and housing data about a population. The census aims to provide a detailed snapshot of the population within a specific geographic area, such as a country, region, or city. While census data is crucial for countries to gather comprehensive information about their population, monitor demographic changes, and employ policies, this data is often released at a decadal frequency due to the costly nature of obtaining this data.

Satellite images offer an approach to be able to predict these variables. Satellite imagery can provide detailed, real-time information on economic activities or shocks, that can be cumbersome or expensive to obtain otherwise. This is particularly useful since accessing satellite

imagery has become increasingly accessible to the public, researchers, and organizations due to technological advancements and the availability of various data sources. Although fairly granular data is released annually, such as Eurostat releasing demographic and economic information of each country in the European Union at a NUTS3 level, which are small administrative areas within countries, there are several compelling reasons why various stakeholders may require even more granular data on income and population levels that satellite imagery can provide. Firstly, satellite imagery can be updated more frequently than annual reports. Having access to more real-time data on population movements and economic activity can help governments and businesses make more timely and informed decisions. Additionally, more granular data can help in understanding specific areas within a region that are experiencing rapid growth or decline, which can be crucial for governments in making decisions on where to invest in infrastructure and how to plan urban development. Furthermore, highly granular data can be essential for emergency response planning and resource allocation in times of natural disasters or pandemics. Moreover, companies in the transportation or logistics sectors can benefit from granular data by optimizing their transportation routes. On top of that, businesses understanding the fine-grained economic and population data can be invaluable for market research, where this data can help inform decisions regarding where to open new stores or offer new services. Also, investors could use the granular data to identify areas experiencing population growth, which could signal potential investment opportunities before they are recognized more widely. Lastly, researchers in spatial economics can use granular data to study the economic interactions over space and time, understanding patterns of trade, migration, and the development of economic agglomerations.

Recent research has utilised satellite images as a data source, such as the study conducted by Sutton et al. (2007), which obtains estimates of GDP at sub-national levels for 4 different countries using nighttime satellite images. Elvidge et al. (2009) produced the first global poverty map at a $1km^2$ resolution using nighttime satellite images, which previously was only available at a national level. The poverty estimates were 2.2 billion globally, compared to the 2.6 billion obtained from the World Development Indicators. Furthermore, Ghosh et al. (2009) estimated the value of the informal economy of Mexico using satellite imagery, since informal transactions in an economy serve as a large portion of the means of livelihood of populations, especially in developing countries. This informal economy is often underestimated because of the problems associated with producing these estimates and is thus not considered in the GDP value. Hence, the authors estimate this informal economy by attributing its value to the surplus of the GDP value estimated by spatial patterns of nighttime lights compared to the official estimate of GDP. Moreover, Tingzon et al. (2019) predicted four socioeconomic indicators namely wealth levels, years of education, access to electricity, and access to water through a combination of volunteered geographic information and nighttime satellite imagery, where the best models explained about 63% of the variation in asset-based wealth. This approach of using nighttime satellite imagery is satisfactory over larger areas such as cities, states, and countries, however, it becomes problematic when trying to study smaller areas, as high luminosity in city centers saturates satellite images as well as surface reflectance leading to bleeding of light across space.

To address this issue, research has been conducted using Convolutional Neural Networks (CNN) to predict economic variables from daytime satellite imagery. Jean et al. (2016) used a combination of nighttime maps and high-resolution daytime satellite images to accurately estimate consumption expenditure and asset wealth from five African countries, where the model was able to identify image features that explain up to 75% of the variation in local-level economic outcomes. Another study that used a combination of nighttime and daytime satellite images was done by Burke et al. (2021). This study quantitatively assesses the predictive performances of a variety of approaches that are used to extract information from satellite imagery, in the domains of smallholder agriculture, economic livelihoods, population, and informal settlements. The study found that the predictive performance of the satellite-based approaches is fairly strong. Moreover, Yeh et al. (2020) predicted survey-based estimates of asset wealth over 20000 African villages using multispectral satellite imagery, where the models explained 70% of the variation in wealth. The study also finds that daytime imagery is particularly useful in explaining district-aggregated changes in wealth over time, explaining about 50% of the variation. Furthermore, Engstrom et al. (2017) used daytime high-resolution satellite imagery to estimate poverty rates and consumption of administrative units in Sri Lanka, showing how the features extracted explain 60% of the variation of these variables using a simple linear regression model, compared to 15% when using features extracted from nighttime satellite imagery. A 'task-agnostic' learning approach was developed using daytime satellite imagery by Rolf et al. (2021), which can be used for a diverse set of prediction tasks with accuracy competitive with deep neural networks, all while having far lower computational costs. Predictions for income and population levels achieved R^2 values of 0.42 and 0.75 respectively.

Daytime satellite imagery has so far been proven useful for extracting latent economic information through the use of CNNs, which enabled researchers to accurately estimate or predict a diverse set of economic variables where previous data sources were scarce. However, the training and tuning of these models have substantial computational and time requirements. If transfer learning could be effectively applied to these CNN models, this could prove to be a powerful tool for governments, businesses, investors, and researchers going forward since the need for training these models are not required. This leads to the main research question of this paper:

What is the potential for using transfer learning with CNNs trained on high-resolution daytime satellite imagery to advance economic theory and policy?

In contrast to Rolf et al. (2021), Khachiyan et al. (2022) developed a learning approach using a CNN that was specifically made for predicting income and population at a fine spatial scale of the United States (US) using daytime satellite imagery of the US, with accuracy levels that far exceed existing approaches in both the cross-section and time series. Population and income levels were predicted for the years 2000 and 2010, using both large (2.4km x 2.4km) and small (1.2km x 1.2km) satellite images, achieving R^2 values far greater than 0.75 for both variables. The small 1.2km images are similar in dimension to the 1km images that Piaggesi et al. (2019) and Rolf et al. (2021) used in their approaches. The R^2 values are greater than those obtained by the model created by Rolf et al. (2021), since that model aims for generality rather than

specificity in predicting outcomes from satellite imagery. Population and income changes from 2000 and 2010 were also predicted using large and small images, achieving R^2 values greater than 0.30 and 0.26 respectively. This was the first study to develop a model that predicts changes in local income and population levels at this fine spatial scale, hence there are no estimates in the literature to benchmark these predictions. Khachiyan et al. (2022) acknowledge that there is potential for predicted values having some measurement errors, and to remove any potential correlation between prediction errors and initial conditions, controls are included for local economic characteristics from Census data in their models. Doing this, Khachiyan et al. (2022) find minimal correlations between predictions and initial conditions in their data. Including initial conditions in the model improves R^2 values moderately, with the largest increase being 0.12 when compared to a model without initial conditions. Khachiyan et al. (2022) also trained a CNN on nighttime satellite images to compare with their daytime-trained CNN and found that daytime imagery outperformed nighttime lights significantly. The R^2 values for income levels were 0.33 higher for 2.4km images (0.90 versus 0.57) and 0.35 higher for 1.2km images (0.85 versus 0.50), and similar results were shown for population levels. Furthermore, the difference was even more significant when looking at the predictions for income changes, where the R^2 values were 0.30 higher for 2.4km images (0.40 versus 0.10) and 0.26 higher for 1.2km images (0.32 versus 0.06). Once again, similar results were obtained for population changes. From these results, it is also clear to see that the large 2.4km images achieved consistently higher R^2 values compared to the small 1.2km images. Since the model was trained on images in the 2000 - 2010 period, Khachiyan et al. (2022) predicted population and income levels for the years 2020 and 2017 respectively to see how the model performs out-of-sample which resulted in R^2 being larger than 0.87 and 0.83 respectively. Population and income changes were also predicted out-of-sample, and achieved R^2 values greater than 0.51 and 0.37 for the periods 2000 - 2020 and 2000 - 2017 respectively. However, for the periods 2010 - 2020 and 2010 - 2017, lower R^2 values of 0.17 and -0.1 were obtained for population and income changes respectively. The authors argue that this is due to the model performing best over long time periods and in periods without large business cycle fluctuations such as the Great Recession in the 2007 - 2009 period. Finally, Khachiyan et al. (2022) examined the robustness of their results to changes in the satellite imagery and machine-learning methods used in the analysis. Limiting the satellite images to only the 3 RGB bands, compared to the 7 spectral bands (3 visible, 2 near-infrared, 1 thermal, and 1 mid-infrared) used in the initial analysis, resulted in R^2 values decreasing by 0.04 for both the population and income levels predictions, and a decrease of 0.11 and 0.06 for population and income changes respectively. Furthermore, increasing the resolution of the satellite images to 15 meters, compared to 30 meters used in the initial analysis, only resulted in R^2 values increasing by at most 0.005 for all models.

This paper will replicate part of Khachiyan et al. (2022) results, which will be presented in section 4. The predictions that will be replicated are in-sample population and income level predictions of the US for the years 2000 and 2010 obtained from the CNNs trained on large (2.4km x 2.4km) images, low (30m) resolution, 7 spectral bands, and without initial conditions.

While the model developed by Khachiyan et al. (2022) achieved high predictive accuracy in the context of the US, its performance across different countries remained unexplored. There are several reasons why the CNN model could perform differently in England compared to the US. For instance, England has a larger population density, which could affect how well the model captures population distribution. England's smaller size compared to the US also means that more areas are commutable, which could have implications for income distribution as people have more flexibility in choosing where to live relative to their workplace. Additionally, the sectoral composition of the economy is different, with England having a different mix of industries which could influence both income and population patterns. Khachiyan et al. (2022) has made their trained model publicly available and it would be of interest to see, despite these differences, if this model performs similarly when applied to a new dataset without the need to be trained from the ground up using this new dataset since training and tuning the model is computationally demanding. If the model proves to exhibit similar accuracy in estimating income and population levels for England, it would not only signify the potential for transfer learning to be applied more widely in economic analysis, but also highlight the model's adaptability in accounting for regional differences such as population density, geographic size, and economic structure. This leads to the two research sub-questions of this paper:

1. What is the predictive power of a CNN trained on US satellite images, to estimate income per capita levels of England?

2. What is the predictive power of a CNN trained on US satellite images, to estimate population levels of England?

Answering the two research sub-questions, this paper will deviate from Khachiyan et al. (2022) in the following manner. Firstly, this paper restricts the predictions to population and income levels for the years 2010, 2011, and 2014 and will not consider predictions of population and income changes, since extracting satellite images over multiple periods requires considerable amounts of storage. Secondly, this paper will only consider the model trained on large (2.4km) daytime satellite images, since this model achieved the highest R^2 values for both population and income levels in Khachiyan et al. (2022). The predictions of England will be compared to predictions of California which serves as a benchmark, where the choice of this benchmark is further explained in the methodology section.

The ability to transfer knowledge from one dataset to another has significant implications for research and policymaking. It enables researchers to use the knowledge gained from a specific dataset, such as the US in this case, and apply them to different countries or regions. Not only does this save time and lower computational costs, but it also enables researchers to make accurate estimations and predictions for other countries or regions without the need for costly specialized computers. Transfer learning thus enhances the scalability and generalizability of the predictive models, enabling researchers to investigate a wider range of socioeconomic dynamics and contribute towards understanding global economic phenomena. In addition, the application of transfer learning facilitates cross-country comparisons. By utilising a trained model that

demonstrated high predictive accuracy in one country, researchers can establish meaningful comparisons between economic variables in different countries or regions. This comparative analysis can provide useful insights into the similarities, differences, and underlying factors influencing economic growth and development across different countries. Hence, by investigating the transfer learning ability of the model developed by Khachiyan et al. (2022) to estimate income per capita and population levels in England, this research contributes to the existing literature on transfer learning in economic analysis. This advancement in transfer learning has the potential to shape the future of economic research, policy formulation, and decision-making.

This paper finds that the CNN trained on US data has a lower predictive capacity when applied to England. The R^2 value for predicting income per capita levels is -5.612, whereas California has a value of 0.246. This paper then further addresses potential systematic biases and scaling issues in the predictions through the use of a model stacking approach, and finds improved R^2 values of 0.051 and 0.318 for England and California respectively. This indicates that the model still isn't able to predict income per capita levels for England, while it is able to partially predict income per capita levels for California from which the model was partially trained. However, the model performs better when predicting population levels for both England and California. The R^2 values for predicting population levels are -1.673 and 0.808 for England and California respectively, where the values increase to 0.460 and 0.850 after systematic biases and scaling issues are addressed. This suggests that the model has some predictive capacity for population levels of England given the moderate R^2 value, however, the model performs significantly better for population levels of California where it obtains highly accurate predictions.

Overall, the findings of this paper contribute to the existing literature by providing empirical evidence on the limitations and potential of transfer learning in the domain of economic prediction. By showcasing the varying levels of performance of a CNN trained on US data when applied to England, this paper provides real-world evidence that emphasizes the necessity for careful consideration of regional specificity in transfer learning. Another significant contribution lies in distinguishing between different economic variables, namely income per capita and population levels. This paper demonstrates that transfer learning models may exhibit varied performance depending on the economic variable in question. This introduces a new perspective in the literature, encouraging researchers to not only consider specificity in the geographic context but also the economic variables being predicted.

The rest of the paper is structured as followed: Section 2 describes what data was extracted and used for the analysis, and presents brief summary statistics. Section 3 presents the methodology used for the extraction of images, predicting population and income levels, and evaluating these predictions. Section 4 presents the replication of Khachiyan et al. (2022), then examines the predictive power of a CNN trained on US images when predicting population and income levels of England. Section 5 concludes.

2 Data

To predict income per capita and population levels of England, satellite imagery of England detected by the United States Geological Survey (USGS) Landsat 7 satellite was used and obtained using Google Earth Engine (GEE) (Gorelick et al., 2017) for the year 2010. The images have a spatial resolution of 30m and are constructed of 7 spectral bands (3 visible, 2 near-infrared, 1 thermal, and 1 mid-infrared). To ensure that the images are without irregularities such as persistent clouds or snow, images are taken from the May-August median (summer months). To compare these predictions to actual income per capita and population levels, census data of England was obtained from the Office for National Statistics (ONS) (Office for National Statistics, 2020, 2021). The dataset comprises of two administrative areas, namely Lower Layer Super Output Areas (LSOA), and Middle Layer Super Output Areas (MSOA) for the year 2010. LSOAs consist of 400 to 1200 households and have a resident population between 1000 and 3000 persons, whereas MSOAs are built of multiple LSOAs and consist of 2000 to 6000 households and have a resident population between 5000 and 15000 persons. Population predictions are compared to actual population levels of LSOAs, whereas income predictions are compared to actual income levels of MSOAs since the ONS does not provide income data at the level of LSOAs. A visualisation of the different administrative areas for England is provided in Figure 1.

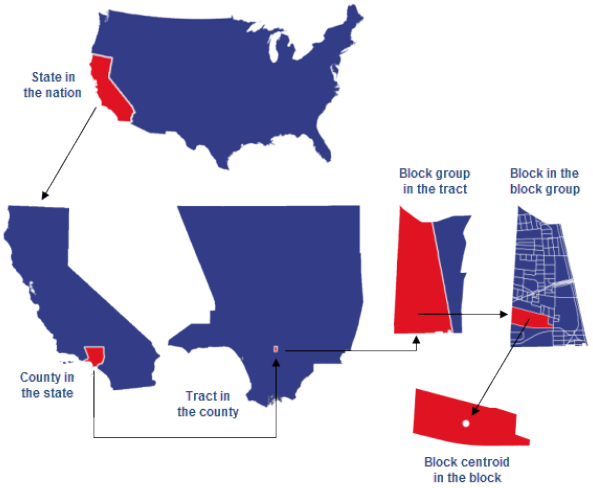
Figure 1: Administrative Areas Visualisation: England

Note. From LSOAs, LEPs and lookups: A beginner's guide to statistical geographies. OCSI (`https://ocsi.uk/2019/03/18/lsoas-leps-and-lookups-a-beginners-guide-to-statistical-geographies/`).

Predicting income and population levels of the US in 2010 was done using the same type of satellite imagery described above, only covering the US in this case. To compare these predictions to actual population and income levels, census data were obtained from the National Historical Geographic Information System (Manson et al., 2022). The dataset comprises of two administrative areas, namely Census Blocks and Census Block Groups for the year 2010. These are the US equivalent of LSOAs and MSOAs in England. Census Blocks are the smallest administrative area for which the Census Bureau collects decennial data. Census Block Groups are above Blocks in the administrative hierarchy and consist of a resident population between 600 and 3000 persons, and in 2000 had a mean of 39 Blocks per Block Group. Once again, population

predictions are compared to actual population levels of Blocks, whereas income predictions are compared to actual income levels of Block Groups since income data are published at the Block Group level. A visualisation of the different administrative areas for the US is provided in Figure 2.

Figure 2: Administrative Areas Visualisation: US



Note. From Documentation. ESRI (`https://learn.arcgis.com/en/related-concepts/united-states-census-geography.htm`).

Table 1 provides descriptive statistics of the income per capita and population levels of England and the US at fine geographic areas. The number of observations is substantially lower for England than the US due to England having a smaller geographic area, as well as having larger granularity of data. The number of observations is also substantially lower for income per capita levels than population levels, given that income per capita levels are provided at a larger geographic area compared to population levels. Furthermore, the mean income per capita in England in 2011 is almost double that of the mean in the US in both 2000 and 2010, which could be due to differences in economic conditions or purchasing power parities. These values have been normalized to 2012 US dollars, which will be further explained in the methodology section. Additionally, the mean population is much larger in England in 2011 compared to the US in 2000 and 2010, which is likely due to the different geographic levels of data being used. Each LSOA in England encompasses a larger population than each Block in the US.

Table 1: Descriptive Comparison of Income per Capita and Population Levels: England (2011) vs. US (2000 & 2010)

	England 2011		US 2000		US 2010	
	Income	Population	Income	Population	Income	Population
Count	7201	34753	208790	8164718	219040	11007990
Mean	61492	1616	31031	34	28523	28
St. Dev.	15979	306	17536	92	16579	78
Min	25224	987	0	0	49	0
25%	49607	1435	21066	0	18505	0
50%	59696	1564	27348	8	24953	3
75%	70626	1738	36214	36	33932	27
Max	145457	8159	666146	23373	366888	19352

Note. Income represents income per capita levels, and is presented in 2012 US dollars. Income per capita levels is at MSOA level and Block Group level for England and US respectively. Population levels is at LSOA level and Block level for England and US respectively.

3 Methodology

3.1 Satellite imagery extraction

This paper predicts income and population levels of England and California from satellite imagery using the trained CNN model of Khachiyan et al. (2022), where California will serve as a benchmark to evaluate the performance of the England predictions. To obtain these predictions of England and California, a satellite imagery extraction process is required. This paper deviates from Khachiyan et al. (2022) with the satellite imagery extraction process, where the satellite data files of Khachiyan et al. (2022) contain label information used in the training and tuning of the CNN model. This paper does not aim to retrain the model and thus the satellite data files do not require the added label information. Thus, to accurately benchmark the predictions done on England satellite imagery, this paper extracts US satellite imagery without the added label information. However, extracting satellite imagery over the entirety of the US at a spatial resolution of 30m requires large amounts of storage and runtime requirements. To limit these requirements, only the state of California is chosen to serve as a benchmark. California was specifically selected since it has the largest population level of all states in the US, which most closely reflects England's population level. In addition, both regions are characterized by significant cultural diversity, with a mix of different ethnicities and backgrounds. Furthermore, when compared to all other U.S. states, California has the largest GDP and is closest to that of England.

To ensure that the image data isn't populated with uninhabited areas, images are only extracted over urban areas. For England, this was done by creating a shapefile (a common geospatial vector data format used in geographic information system software) which contains boundaries for the built-up areas (BUA) of England. BUAs refer to densely developed areas with buildings and infrastructure which include cities, towns, suburban regions, and urbanized regions. To once again limit the amount of storage and runtime requirements, the BUAs were ranked ac-

cording to their shape areas from highest to lowest, and the top 400 were used in the extraction of urban area images for England. For California, an urban area shapefile of the US created by Khachiyan et al. (2022) was used and adjusted to only include the state of California. The original shapefile was created by ranking the Block Groups according to population density and identifying the Block Groups that comprise of 85% of the US population. A 1-mile buffer was then drawn around these Block Groups, leading to the data covering 93% of the US population in the year 2000.

Khachiyan et al. (2022) extracted images with varying characteristics (large/small images, 7 spectral bands, RGB, night-lights, low/high-resolution, with/without initial conditions). It was found that model fit was consistently higher on the sample of larger images (2.4km x 2.4km) than on smaller images (1.2km x 1.2km), thus this paper will only consider the former. The RGB, night-light, and high-resolution images were used as robustness exercises and comparison reasons. Since the 7 spectral bands out-performed the other two satellite images significantly when considering the R^2 values, and the high-resolution (15m) images only moderately improved the R^2 values, this paper will only consider low-resolution (30m) images with 7 spectral bands (3 visible, 2 near-infrared, 1 thermal, 1 mid-infrared). Moreover, Khachiyan et al. (2022) made predictions on images with initial conditions (label data) which performed consistently better than predictions on images without initial conditions, however, these conditions are quite extensive and require a substantial amount of time to obtain, hence this paper will only focus on model predictions without initial conditions. Furthermore, Khachiyan et al. (2022) extracted images over multiple years (2000 to 2019), which was used in predicting population and income levels as well as changes over 10-year periods. This requires substantial storage and runtime requirements, leading to this paper focusing on predicting population and income levels for 3 years, namely 2010, 2011 and 2014 (which is used as part of a robustness exercise).

For the replication of Khachiyan et al. (2022), the satellite images of the urban areas of the US were provided in the replication package and were directly used as input in the prediction model.

3.2 Convolutional Neural Network

Khachiyan et al. (2022) make use of a CNN to predict income and population levels of the US given its ability to process pixel data. This model was trained using satellite imagery of the US obtained from the USGS Landsat 7 satellite. The CNN is a 7-band version of the VGG16 network model created by Simonyan and Zisserman (2014). The model architecture is defined as having an input layer, three convolutional blocks, a "flatten" layer that vectorizes the output of the convolutional blocks, and a fully connected block. The output of the fully connected block is passed through a final linear layer which produces a scalar value that is the predicted output. The weights in all layers in the model are initialized using the Glorot Normal random initialization. The weight of each layer is initialized by drawing random values from a Gaussian distribution with zero mean and a variance that is calculated based on the number of input and output connections to that layer. The purpose of this is to prevent the activations from

exploding or vanishing.

Each convolutional block contains three two-dimensional convolution layers which is then followed by a max-pooling layer. The final convolutional block's output is vectorized, which is used as input to the fully connected block. The number of filters is constant within each convolutional block but doubles for each subsequent block. In this case, the first block has 32 filters, whereas the second and third blocks have 64 and 128 filters respectively. The convolution layers perform convolutions with a stride of 1, a kernel size of 3, and apply the Rectified Linear Unit (ReLU) activation function. The stride refers to the step size at which the convolution kernel moves across the input data. Thus, a stride of 1 means that the kernel moves 1 pixel at a time, ensuring that the kernel covers the entire input spatially without skipping any pixels. The kernel size determines the dimensions of the square-shaped filter that moves across the input data, thus a kernel of size 3 suggests that the filter is a 3×3 matrix. The ReLU activation function applies an element-wise operation, replacing all negative values with zero while keeping positive values unchanged. Furthermore, the convolution kernels are regularized using an L2 norm penalty which helps prevent overfitting, where cross-validation is used to determine the strength of the penalty. The max-pooling layer is used to reduce the spatial dimensions of the feature maps and does this by dividing the feature maps into 2×2 regions, and within each region selecting the maximum value. This downsampling helps reduce computation complexity, extract the most prominent features, and prevent overfitting.

The fully connected block consists of three fully connected layers, each of which is separated by a dropout layer. The fully connected layers also make use of ReLU activations and regularization using an L2 norm penalty. This penalty strength is again determined using cross-validation and grid search. The number of neurons in each fully connected layer is set as $l_i * n$, where $l_1 = l_2 = 16$, $l_3 = 8$, and n is the number of filters in the first convolutional block (in this case $n = 32$). Dropout layers are used as a regularization technique to combat overfitting and achieve this by randomly setting a fraction of the input units to zero during training. The fraction of units dropped is determined by the dropout probability, which is fixed at 0.5.

Since training a CNN is very time intensive and requires specialized computers made for machine learning models, this paper does not attempt to train a CNN on England satellite imagery. Instead, predictions of income per capita and population levels of England will be made using the model parameters obtained from the already trained CNN model of Khachiyan et al. (2022).

3.3 Evaluating Predictions

The predictions are evaluated on an image level and not geographic area level, since the model predicts income and population levels over 2.4 x 2.4 km images. These images potentially contain multiple geographic areas (LSOA/MSOA for England, and Blocks/Block Groups for California). Interpolation of these predictions to the geographic areas would lead to less reliable results, as the predictions would be distributed across the geographic areas based on the percentage area

covered by these areas within an image, and not based on economic information that is latent in the spectral data. Thus, the actual values of images are obtained in the following manner: For population levels, the actual values of the geographic areas within an image will be summed up to obtain the actual value of the image. For example, if a satellite image of England contains 20 LSOAs, the population levels of each LSOA will be summed up to give one population level for the image. For income per capita levels, the average of the actual values of the geographic areas within an image will be computed to obtain the actual value of the image. For example, if a satellite image of England contains 4 MSOAs, the average of the 4 MSOA income per capita levels will be calculated to give one income per capita level for the image. Note that this leads to one of the limitations of this paper since the reported actual population values of images are likely to be larger than the true population values since a geographical area can have a spatial overlap over multiple images, resulting in this geographical area's population value being added multiple times. Furthermore, the reported actual income per capita levels are likely to be different (both larger or smaller) than the true income per capita levels, for the same reason as population levels, resulting in the geographical area's income per capita value being used multiple times for computing averages.

Adjustments for both inflation and exchange rates are needed to compare the income per capita levels predictions with the actual values, since the model predicts these values in 2012 US dollars, whereas the actual values for England are in 2011 British pound sterling, and for California, the actual values are in 2011 US dollars. The adjustments are done using equations (1) and (2) for England and California respectively.

$$y_{12\$} = y_{11£} * (1 + \pi_{11-12}) * x_{12\frac{\$}{£}} \tag{1}$$

$$y_{12\$} = y_{10\$} * (1 + \pi_{10-12}) \tag{2}$$

where $y_{11£}$, $y_{10\$}$ and $y_{12\$}$ are the actual income per capita levels measured in 2011 British pound sterling, 2010 US dollars and 2012 US dollars respectively. π_{11-12} and π_{10-12} are the US dollar inflation rates from 2011 to 2012 and 2010 to 2012 respectively, and $x_{12\frac{\$}{£}}$ is the average 2012 British pound sterling to US dollar exchange rate.

R^2 values will be calculated to assess the performance of the predicted values from the CNN using equation (3).

$$R^2 = 1 - \frac{SST}{TSS} \tag{3}$$

with:

$$SST = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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

where y_i and $\bar{y}$ are the observed values and mean of observed values of the dependent variable

respectively, n is the number of observations, and $\hat{y}_i$ is the predicted values of the dependent variable.

Furthermore, regression analysis is used to fit the predictions from the CNN with the actual values using Ordinary Least Squares (OLS) to address potential model misspecifications. The presence of heteroskedasticity is tested using the Breusch-Pagan test. Under the null hypothesis, homoskedasticity is present, while under the alternative hypothesis, heteroskedasticity is present. The test is conducted by first fitting a regression model, then the squared residuals are calculated and used as the dependent variable in a new regression model with the original regressors. The Chi-Square test statistic χ^2 is then calculated as $n * R^2_{new}$, where n is the total number of observations and R^2_{new} is the R^2 value of the regression model with the squared residual as the dependent variable. The null hypothesis is rejected when the p-value corresponding to the χ^2 test statistic with p degrees of freedom is less than a 5% significance level, where p corresponds to the number of regressors.

4 Results

The results section is divided into three subsections, where the first subsection presents partial replicated results of Khachiyan et al. (2022) analysing the predictive performance of the CNN for the US. The second subsection presents the extension to Khachiyan et al. (2022), where the predictive performance of the CNN is analysed when using a dataset from which the model is not trained, namely England. The final subsection presents the results of a robustness exercise.

4.1 Replication

R^2 values are presented in Table 2 for the large 2.4km x 2.4km, 7 spectral band model predictions of income per capita, population, log-income per capita, and log-population levels for the years 2000, 2010, and 2000 & 2010 combined.

Table 2: Comparison of R^2 Values for CNN Model Predictions: US (2000, 2010, and 2000 & 2010 Combined)

	2000	2010	2000 & 2010
Income per Capita	0.381	0.424	0.405
Population	0.773	0.701	0.737
Log Income per Capita	0.447	0.454	0.455
Log Population	0.889	0.891	0.891

Note. The number of observations for income per capita and population is 112932 for both the years 2000 and 2010, and 225864 for the years combined.

Looking at Table 2, the R^2 values of income per capita predictions are 0.381, 0.424, and 0.405 for the years 2000, 2010, and 2000 & 2010 combined respectively. The log-transformed income predictions have slightly improved values, with values of 0.447, 0.454, and 0.455 for the three

periods respectively. This suggests that the model fits the data somewhat well, where income levels might have a moderate linear or exponential relationship with the features extracted from the satellite images. On the other hand, the R^2 values of population predictions are significantly higher when compared to income predictions, with values of 0.773, 0.701, and 0.737 for the years 2000, 2010, and 2000 & 2010 combined respectively. A strong linear relationship between population levels and the features extracted could be argued, though transforming the population levels to a log scale substantially increases the R^2 values to 0.889, 0.891, and 0.891 for the years 2000, 2010, and 2000 & 2010 combined respectively. Due to this, the relationship between the features and population levels might be exponential rather than linear, and transforming the population levels to a log scale linearizes this relationship.

Figure 3 shows scatter plots of the actual values of income per capita and population levels against the model-predicted values. For illustration purposes, the 20 biggest outliers are excluded from both scatter plots.

Figure 3: Scatter Plots of US Predictions and Actual Values with OLS Fitted Lines (Years 2000 & 2010 Combined)

Note. The scatter plots are presented with the 20 biggest outliers excluded for illustration purposes.

The plot of income per capita levels exhibits a positive correlation between predicted and actual income levels, which is consistent with the R^2 values presented in Table 2. This positive correlation indicates that as the actual income per capita increases, the predicted values generally also increase, suggesting that the model is capturing some of the underlying trends. However, it is worth noting that there is a presence of heteroskedasticity, as evidenced by the fanning out of the scatter plot for higher income levels. This implies that the model's accuracy varies depending on the income level, with predictions becoming less reliable as income levels increase. This can also be seen from the OLS line fitted in the scatter plot, where the deviations from the line are larger for large income levels. Another observation is that the model did not provide predictions lower than 17076.53, suggesting that the model fails to predict low income levels. The plot of population levels exhibits a larger positive correlation between predicted and actual values compared to the plot of income levels, which is in line with the higher R^2 values presented in Table 2. Thus, the model is able to better capture the underlying trends in population levels than income levels. The model also performed better in predicting low population levels compared

to predicting low income levels, which can be seen from the observations being tightly grouped around the origin in the population scatter plot. As with income levels, the model's predictions become less reliable for larger population levels. This is due to the presence of heteroskedasticity in the model, as evidenced by the deviations from the fitted OLS line increasing as actual population levels rise.

4.2 Extension

R^2 values are presented in Table 3 for the large 2.4km x 2.4km, 7 spectral band model predictions of income per capita, population, log-income per capita, and log-population levels for the year 2010 for California, and 2011 for England.

Table 3: Comparison of R^2 Values for CNN Model Predictions: England (2011) & California (2010)

	England 2011	California 2010
Income per Capita	-5.612	0.246
Population	-1.673	0.808
Log Income per Capita	-12.248	0.383
Log Population	-13.135	0.674

Note. The number of observations for income per capita is 3680 for England and 10963 for California. For population, there are 3746 observations for England and 20803 for California.

Table 3 shows that the R^2 values for predictions in England are negative for both income per capita and population levels, with values of -5.612 and -1.673 respectively. These negative values mean that a horizontal line at the mean of the actual income is a better predictor than the model trying to predict income. The log-transformed versions of income and population levels also show negative R^2 values, with values of -12.248 and -13.135 respectively, suggesting that the model struggles regardless of the scale of the target variable. In contrast, the R^2 values for predictions in California are positive for both income and population levels, with corresponding values of 0.246 and 0.808. This indicates that the model performed relatively well on California data, which is to be expected since California is part of the US on which the model is trained on. Interestingly, the model performs slightly worse for the log-transformed population levels, with a value of 0.674, which is in contrast with the findings from Table 2, where the log-transformed predictions performed better than the standard predictions for the US. This is not the case for log-transformed income levels, which see an increase to 0.383 and is in line with the findings from Table 2.

These results indicate that the model, which was trained on US data, does not generalize well to England. This could be due to differences in geographical and socio-economic features between the US and England, which the model has not been trained to recognize. One difference between the US and England is the population density, where England is more densely populated than most parts of the US. The trained model might not have learned features that are relevant to

highly dense areas if it was predominantly trained on less dense areas. Furthermore, the size and scale of the buildings and plots might be different between the two countries, causing the model to misinterpret the data. Moreover, the satellite images for England may have different spectral characteristics than the US due to the different geographic regions, for example, differences in vegetation, soil, and water bodies might affect the features that the model has learned to associate with income or population levels. Additionally, the urban infrastructure is different in the US and England, where the US tends to have more sprawled-out suburban areas, whereas England has more compact cities. This can cause the model to misinterpret land-use features, for instance, in the US, large houses with big lawns may be an indicator of wealth, whereas in England wealth may be less directly correlated with property size. Finally, the model might be trained on a certain scale of economic disparities in the US, which might not translate well when trying to predict income levels in England.

In light of the geographical and socio-economic disparities between the US and England identified above, it is evident that adjustments are needed for the model to be effective in the context of England. One such measure is to employ an OLS regression in a model stacking approach, where the predictions of the initial model are used as inputs to the OLS regression for calibration against the actual values. Model stacking, in this case, serves the purpose of bias correction and scaling adjustments. Specifically, by employing an OLS regression, it is possible to address the systematic biases that might arise due to the differences between the regions, by correcting the intercept (adding or subtracting a constant term), and to handle scaling issues by adjusting the slope (applying a scaling factor). However, while this approach can enhance the model's performance, there is potential for overfitting, which could lead to a model that performs well on the calibration data but does not generalize well to new, unseen data. Table 4 presents the R^2 values obtained after the OLS model is fitted as part of the model stacking process.

Table 4: Comparison of R^2 Values for OLS-Calibrated CNN Model Predictions: England (2011) & California (2010)

	England 2011	California 2010
Income per Capita	0.051	0.318
Population	0.460	0.850
Log Income per Capita	0.060	0.402
Log Population	0.530	0.717

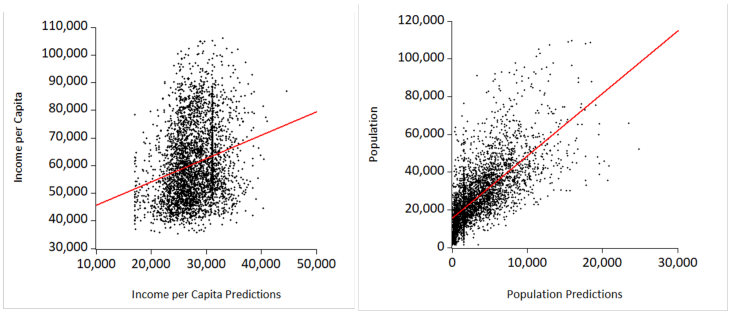
Note. The number of observations for income per capita is 3680 for England and 10963 for California. For population, there are 3746 observations for England and 20803 for California.

Upon fitting the OLS regression to the model's predictions, there is a significant improvement in the R^2 values for England compared to the initial values before the OLS fitting (as shown in Table 3). In Table 4, the values for both income per capita and population levels in England are now positive, with values of 0.051 and 0.460 respectively. The same improvement is present for the log-transformed income and population levels, which now have values of 0.060 and 0.530

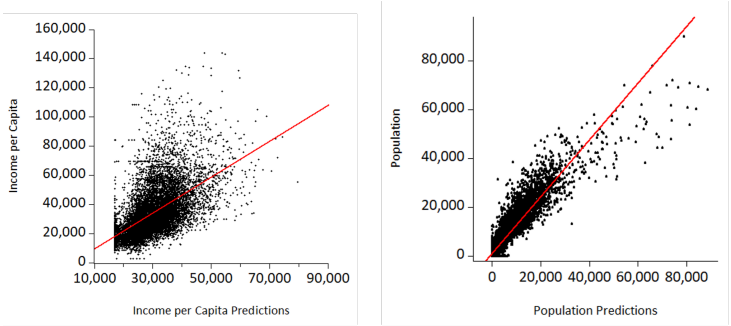
respectively. These positive values indicate that the OLS fitting has corrected some of the systematic biases and scaling issues that were present in the original predictions. For population levels, there is now a moderate linear as well as exponential relationship between the features extracted from the images and the actual population levels, suggesting that the model has some predictive capacity for England. On the other hand, the model still struggles to predict income levels as well as log-transformed income levels for England given the low R^2 values. For California, the R^2 values in Panel A also exhibit improvements, however, these improvements are only slight since the model already performed relatively well and did not have the same systematic biases or scaling issues in predictions.

Figure 4 shows scatter plots of the actual values of income per capita and population levels against the original model predicted values for both England and California, as well as the fitted OLS lines, where the top 20 outliers are excluded for illustration purposes.

Figure 4: Scatter Plots of England (2011) & California (2010) Predictions and Actual Values with OLS Fitted Lines



(a) England 2011



(b) California 2010

Note. The scatter plots are presented with the 20 biggest outliers excluded for illustration purposes.

The income plot for England shows a very slight positive correlation but with a wide spread of observations. This is consistent with the small R^2 values reported in Table 4 and confirms that the model isn't effectively capturing income trends in England. On the other hand, the population plot for England shows a stronger positive correlation, where the observations are more tightly grouped around the OLS line compared to the income plot. Given that the two plots' observations do not start from the origin and that there is a difference in scale between the two

axes, suggests that the OLS calibration accounted for this by adding a constant term and having a coefficient value different than 1, addressing some of the systematic bias and scaling issues which caused the negative R^2 values reported in Table 3. The presence of heteroskedasticity is observed through the higher spread in predictions for higher values, indicating that the model's reliability varies with the level of population. For California, the positive correlation in the income plot is stronger than in England, however, there is still a presence of heteroskedasticity, meaning that the model's reliability is inconsistent across different income levels. Interestingly, the model once again did not predict values lower than 17076.54 for both England and California, which was also the case for the whole US in Figure 3. The population plot for California shows a large positive correlation and observations are tightly grouped around the OLS line. This indicates that the model is highly effective in predicting population levels in California, however, there is a slightly larger spread of observations for larger values of population levels. The fact that observations start from the origin and the scales of the axes are the same suggests that there were fewer systematic bias and scaling issues in the predictions for California compared to England. This is to be expected since the model was trained to predict these values and is consistent with the slight increase in R^2 values in Table 4 after an OLS regression was fitted.

The heteroskedasticity observed in the scatter plots in Figure 4 is formally tested using the Breusch-Pagan test, where the results are presented in Table 5. In the scatter plots, it was observed that for England there was a higher spread in predictions for higher population levels, and for California, there was a higher spread in predictions for higher income and population levels. These visual observations are consistent with the statistical evidence provided by the Breusch-Pagan test. The χ^2 values are significantly high and the p-values are very close to zero for income per capita and population levels, for both England and California. This leads to a rejection of the null hypothesis of homoskedasticity with a 1% significance level. The presence of heteroskedasticity suggests that the model's prediction errors are not constant across different levels of income or population. In practice, it means that the model may not be equally reliable for different ranges of income and population levels. This information is important for researchers and policymakers who might consider using this model for decision-making purposes, as it indicates areas where the model's predictions may be less reliable.

Table 5: Breusch-Pagan Test for OLS-Calibrated CNN Model Predictions: England (2011) & California (2010)

	England 2011		California 2010	
	χ^2	p-value	χ^2	p-value
Income per Capita	14.702	0.000***	215.100	0.000***
Population	234.029	0.000***	3745.180	0.000***

Note. *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively.

To gain some insight into the systematic biases and scaling issues previously mentioned, Table 6 presents the regressions of income per capita and population predictions fitted to the actual values using OLS, for both England and California. Analysing the constant terms and coefficients leads to a further understanding of the nature of corrections required to improve the model's predictions. It's important to note that White standard errors are presented in parentheses, which is appropriate given the presence of heteroskedasticity, as confirmed in Table 5. These standard errors are robust to heteroskedasticity and thus provide a more reliable statistical inference.

Table 6: Regression Analysis of Income per Capita and Population CNN Predictions Using OLS: England (2011) & California (2010)

	England 2011	California 2010
Panel A: Income per Capita		
Constant	37574.020***	-2975.590***
	(1564.664)	(560.569)
Income per Capita prediction	0.838***	1.240***
	(0.056)	(0.020)
R^2	0.051	0.318
Observations	3680	10963
Panel B: Population		
Constant	15291.110***	931.502***
	(309.234)	(46.538)
Population prediction	3.461***	1.138***
	(0.096)	(0.021)
R^2	0.460	0.850
Observations	3746	20803

Note. *, **, *** indicates significance at the 90%, 95%, and 99% level, respectively. White standard errors are presented in parenthesis.

Table 6 shows that there are stark differences in the regression results between England and California. Beginning with Panel A, which focuses on income per capita as the dependent variable, the constant term for England is positive and highly significant with a value of 37574.020, whereas for California the constant term is negative and also significant with a value of -2975.590. This indicates that there is a systematic bias in both regions, although this bias is significantly greater for England than for California. The positive constant term for England implies that the model tends to underpredict income per capita, requiring an upward adjustment to align the predictions with the actual values. Conversely, the negative constant term for California suggests that the model overpredicts income per capita, requiring a downward adjustment. The difference in magnitude of the constant terms suggests that the systematic biases vary signific-

antly between the two regions, with England needing a more substantial correction. Continuing with the coefficients corresponding to the income predictions for England and California, contrasting behaviour is observed. The coefficient for England has a value of 0.838, which is smaller than California's coefficient, which has a value of 1.240. For England, this indicates an over-scaling issue, where the model overestimates the effect of changes in income. For California, this indicates a slight under-scaling issue. Additionally, it is important to consider the R^2 value associated with the regression for income per capita in England, which equals 0.051. The R^2 value is notably small, which implies that despite the adjustments made through the OLS regression, the model still explains only a tiny fraction of the variability in income per capita in England. This suggests that the regression results for England may not be very informative in practice and that the model may lack the ability to capture the underlying dynamics of income per capita in this region. Moreover, the model also struggles somewhat in predicting income per capita accurately for California, from which the model is partially trained, as evidenced by the need for downward adjustments, scaling corrections, and relatively low R^2 value.

Moving to Panel B, which focuses on population as the dependent variable, the constant term for England, much like in Panel A, is positive and significantly large with a value of 15291.110. In contrast, California's constant term is significantly smaller at 931.502. These values, again, indicate the presence of a systematic bias in the model's predictions for both regions, albeit to different extents. The large positive constant for England suggests that the model generally underpredicts population, requiring an upward adjustment. The smaller positive constant for California implies that the adjustment needed is considerably lower in magnitude. Regarding the coefficients for population predictions, England has a value of 3.461, whereas California's coefficient has a value of 1.138. England's coefficient is significantly larger than 1, which indicates an under-scaling issue for population predictions in England. On the other hand, California's coefficient is closer to 1, indicating that the model's predictions are more proportionate to actual values, though still slightly under-scaled. In Panel B, the R^2 value for population in England is moderate, suggesting that the adjusted model captures some of the variability in the population, but not all. This indicates that the model has some predictive capacity for population in England, although not highly accurate. For California, the R^2 value is large, suggesting that the model performs much better in predicting population for the region it was trained on.

4.3 Robustness Exercise

To examine the robustness of the OLS fitted results of England, income per capita and population levels of England are predicted for the year 2014 and calibrated using the England 2011 OLS regression constant terms and coefficients found in Table 6. Both the R^2 values for original and calibrated predictions are presented in Table 7.

Applying the OLS regression calibration to the predictions made by the model for the year 2014 demonstrates a notable level of robustness. This is evidenced by the significant improvement in the R^2 values for the year 2014 when comparing the values obtained from original model

Table 7: Comparison of R^2 Values for 2011 OLS-Calibrated CNN Model Predictions: England (2011 & 2014)

	England 2011	England 2014
Panel A: Original predictions		
Income per Capita	-5.612	-5.940
Population	-1.673	-1.734
Panel B: Calibrated predictions		
Income per Capita	0.051	0.013
Population	0.460	0.407

Note. The number of observations for income per capita is 3680 and 3992 for the years 2011 and 2014 respectively. For population, there are 3746 and 4058 for the years 2011 and 2014 respectively.

predictions in Panel A, to the values obtained from OLS-calibrated predictions in Panel B. The R^2 values increased from -5.940 and -1.734 to 0.013 and 0.407 for income per capita and population levels respectively. These improvements signify that the OLS calibration method was effective in correcting systematic biases and scaling issues in a manner that generalized to a different time period. Since the calibrated model's performance remained relatively stable between 2011 and 2014, this indicated that concerns regarding overfitting have been addressed. By successfully enhancing the model's performance on an independent 2014 dataset without further tuning, suggests that the OLS calibration is capturing underlying relationships rather than fitting noise specific to the 2011 dataset. This supports the validity of employing an OLS calibration being a moderately robust method for enhancing the performance of the CNN when predicting income per capita and population levels for regions outside of the US.

5 Conclusion

This paper has been an exploration of the potential of using transfer learning with CNNs. Specifically, it investigated the application and limitations of transfer learning in predicting economic variables, including income per capita and population levels, using a CNN model trained on US data from Khachiyan et al. (2022). A fundamental research question that this paper addressed was: What is the potential for using transfer learning with CNNs trained on high-resolution daytime satellite imagery to advance economic theory and policy?

To help answer this research question, two sub-questions were presented in this paper. The first sub-question was: What is the predictive power of a CNN trained on US satellite images, to estimate income per capita levels of England? The second sub-question was: What is the predictive power of a CNN trained on US satellite images, to estimate population levels of England? The investigation of this paper centered around answering these two sub-questions to assess how the CNN model performs when applied to a new geographical context, in this case, England, where California served as a benchmark. This examination was vital in understanding the ad-

aptability and limitations of CNN models in making predictions across different geographical contexts.

A finding of this paper was that the CNN model had a poor predictive performance for both income per capita and population levels in England. However, with the integration of an OLS regression in a model stacking approach, there was a noticeable improvement in predictive performance for population levels, though the model still struggled with income per capita levels predictions. This improvement in predictive performance could have potentially been caused by overfitting. Due to this, the robustness of the model stacking approach was tested by applying the OLS calibration of England for the year 2011 to the model's predictions of England for the year 2014. There was a significant improvement in the R^2 values for both income per capita and population levels in 2014, as compared to the original model predictions, which suggests that the model stacking approach was fairly robust and that the improvements are not a result of overfitting. These results lead to answering the two research sub-questions of this paper. For the first sub-question, the model failed to estimate income per capita levels of England, even after the predictions were calibrated using OLS. For the second sub-question, the model initially was unable to estimate population levels, however after the predictions were calibrated using OLS, the model was able to moderately estimate population levels of England.

Additionally the analysis revealed the presence of heteroskedasticity in model predictions for both England and California, through scatter plots and using the Breusch-Pagan test. This indicates a limitation of the model, as the reliability decreases for larger values of income per capita and population levels for both England and California. This highlights the need for further refinement of the model to account for the varying distribution and spread of data across the range of predicted values.

Furthermore, the contrast between the constant terms and coefficients in the OLS regression results for England and California provides further insights into the systematic biases and scaling issues inherent in the model's predictions. The large positive constant terms for England meant that the model notably underestimated both income per capita and population levels. Moreover, the size of the coefficient for income per capita predictions for England was slightly lower than 1, indicating that there was a small over-scaling issue, while for the population predictions, the coefficient was particularly larger than 1, demonstrating a large under-scaling issue.

Practically, these findings hold implications for governments, businesses, investors, and researchers alike. For governments, understanding the potential and limitations of transfer learning in predicting socio-economic indicators like income and population levels is crucial for making informed decisions and formulating policies. The CNN model alongside the OLS calibrations, as demonstrated by its robustness across different time periods, can be a valuable tool for governments to utilise in allocating resources more efficiently, monitoring urbanization patterns, and designing targeted interventions for economic development. Nonetheless, governments should also recognize the model's limitations in predicting income per capita levels. For businesses and investors, predictive models using transfer learning can be useful for market analysis and

making investment decisions. For instance, businesses looking to expand or investors looking for opportunities might use such models to identify areas with rapid population growth or increasing income levels. However, this paper's findings imply that they need to exercise caution and critically evaluate the reliability of such models, especially when they are being applied to different regions. For researchers working in the field of transfer learning and predictive modeling, these findings offer a path to explore methods that further minimize biases and improve model accuracy across different geographic contexts. This includes addressing the remaining challenges in predicting income per capita levels and refining calibration techniques. Future research should focus on developing methodologies that are more adaptive in accounting for regional differences in economic variables, which may include employing more sophisticated model stacking approaches that can better account for the diversity in geographical and socio-economic characteristics.

This paper has made strides toward answering the research question by examining the capabilities and limitations of transfer learning using CNN models in predicting economic variables with satellite imagery. The findings suggest that while there is potential, especially in predicting population levels, there are also inherent challenges and limitations that need to be addressed. This paper paves the way for future research to build upon these findings, aiming toward the development of more robust, adaptable, and reliable predictive models for economic analysis using satellite imagery.

References

Burke, M., Driscoll, A., Lobell, D. B. & Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. *Science*, *371*(6535), eabe8628.

Elvidge, C. D., Sutton, P. C., Ghosh, T., Tuttle, B. T., Baugh, K. E., Bhaduri, B. & Bright, E. (2009). A global poverty map derived from satellite data. *Computers & Geosciences*, *35*(8), 1652–1660.

Engstrom, R., Hersh, J. S. & Newhouse, D. L. (2017). Poverty from space: using high-resolution satellite imagery for estimating economic well-being. *World Bank Policy Research Working Paper*(8284).

Gallin, J. (2006). The long-run relationship between house prices and income: Evidence from local housing markets. *Real Estate Economics*, *34*(3), 417–438. doi: 10.1111/j.1540-6229.2006.00172.x

Ghosh, T., Anderson, S., Powell, R. L., Sutton, P. C. & Elvidge, C. D. (2009). Estimation of mexico's informal economy and remittances using nighttime imagery. *Remote Sensing*, *1*(3), 418–444.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. & Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*. Retrieved from `https://doi.org/10.1016/j.rse.2017.06.031` doi: 10.1016/j.rse.2017.06.031

Holly, S. & Jones, N. (1997). House prices since the 1940s: Cointegration, demography and asymmetries. *Economic Modelling*, *14*(4), 549–565. doi: 10.1016/s0264-9993(97)00009-6

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794. doi: 10.1126/science.aaf7894

Khachiyan, A., Thomas, A., Zhou, H., Hanson, G., Cloninger, A., Rosing, T. & Khandelwal, A. K. (2022). Using neural networks to predict microspatial economic growth. *American Economic Review: Insights*, *4*(4), 491–506. doi: 10.1257/aeri.20210422

Manson, S., Schroeder, J., Van Riper, D., Kugler, T. & Ruggles, S. (2022). *IPUMS National Historical Geographic Information System: Version 17.0.* [dataset]. Minneapolis, MN: IPUMS. Retrieved from `http://doi.org/10.18128/D050.V17.0`

Office for National Statistics. (2020). *Income estimates for small areas, england and wales.* Retrieved from `https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/smallareaincomeestimatesformiddlelayersuperoutputareasenglandandwales`.

Office for National Statistics. (2021). *Lower layer super output area population density (national statistics).* Retrieved from `https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/lowersuperoutputareapopulationdensity`.

Piaggesi, S., Gauvin, L., Tizzoni, M., Cattuto, C., Adler, N., Verhulst, S., ... Panisson, A. (2019, June). Predicting city poverty using satellite imagery. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr) workshops.*

Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., ... Hsiang, S. (2021).

A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*, *12*(1). doi: 10.1038/s41467-021-24638-z

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sutton, P. C., Elvidge, C. D., Ghosh, T. et al. (2007). Estimation of gross domestic product at sub-national scales using nighttime satellite imagery.

Tingzon, I., Orden, A., Go, K., Sy, S., Sekara, V., Weber, I., ... Kim, D. (2019). Mapping poverty in the philippines using machine learning, satellite imagery, and crowd-sourced geospatial information. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., ... Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, *11*(1), 2583.