Multimodal Deep Learning from Satellite and Street-Level Imagery for Determining Socioeconomic Indicators in Urban Areas

Economics 2355: Final Project

Arpit Bhate, Ben Ray, and Cyrus Asgari

Harvard College

May 4, 2023

Abstract

In low-resource settings, socioeconomic indicators are usually difficult to gather on a consistent basis, while they are usually only sampled at 5 to 10 year intervals in developed areas - leading to a dearth of information about various regions. Our project attempts to take a first step in solving this by predicting socioeconomic indicators using street-level and satellite imagery. We use a pretrained SWIN model to generate representations of images from London and then train a linear network on these representations to predict economic indicators - income, overcrowding, and living environment deprivation. Our methods are able to achieve a mean absolute error of $\sim$1.7 on certain indicators, approaching the accuracy of our reference paper [Suel et al., 2021] with a much smaller training dataset. These methodology updates show promise for the prediction of economic indicators in low resource settings. Our implementation can be found in this Github repository.

1 Introduction

Measurements of economic development and inequality provide valuable insight for determining changing socioeconomic conditions and guiding public policy. In particular, metrics such as income, overcrowding, and living environment deprivation are valuable for determining the efficacy of existing policies directed towards economic equality and public health.

Data for informing such policies is often limited and sparse. Census data, while informative, offers insight into measuring socioeconomic status only at intervals of 5 to 10 years. Thus, large scale data sources, such as remote sensing images, offer a novel source for measuring such economic indicators at greater frequency, without relying on responses from census collections. This is particularly relevant for areas where collecting census data is more challenging and prone to response bias.

Our project uses a combination of street-level and satellite imagery in order to predict these economic indicators. Using images collected from London, we train a classifier on the feature representations of these images and consequently build our prediction model. Beyond what we are able to accomplish in this paper, future work on this project ideally involves fine-tuning the SWIN model for the task of featurising street-level images and predicting economic indicators, using data from the whole of London and several other cities to train the model, and running inference in areas with varied levels of economic development to test the accuracy of its predictions.

2 Background & Relevant Literature

Past research has applied deep learning methods to either satellite image data or street-level data to obtain various measures of economic equality and development. For instance, numerous studies have applied deep learning methods to satellite data in measuring poverty [Jean et al., 2016] [Steele et al., 2017]. Likewise, street-level imagery has been employed to measure income [Gebru et al., 2017] as well as social and economic inequalities [Suel et al., 2019].

We take a unique approach in simultaneously using both satellite and street-level imagery to measure multiple economic indicators with greater accuracy. Our project is motivated by the findings of a previous paper, published in 2021 in the journal *Remote Sensing of Environment*, which applied multimodal deep learning to satellite and street-level imagery to measure economic indicators such as income, overcrowding, and environmental deprivation. The multimodal approach described in this paper was able to achieve improvements in accuracy ranging from 6 to 11 percent over a similar model using solely street-level data [Suel et al., 2021].

Given the rapid progress of deep learning in recent years - with the advent of vastly improved vision transformers and convolutional neural network models - we modernize the approach taken in this paper by updating its architecture to achieve even more accurate predictions of such economic indicators. In the sections that follow, we outline the data from London that we used, the updated model architectures that we have implemented, and discuss our results and their possible extensions.

3 Methods

3.1 Dataset

3.1.1 Census Data As Labels

London is separated into "Lower Layer Super Output Areas" (LSOAs), which are a convenient geographic unit for reporting small area statistics in England and Wales. In this study, we will use the same LSOA-level outcome data as Suel et al. [2021], including:

- **Income**: mean annual household income estimate, obtained from the Greater London Authority for 2011.
- **Overcrowding**: the percentage of households classified by the Office for National Statistics (ONS) as having at least one fewer room than required based on the number of occupants, obtained from the UK Census 2011.
- **Living environment deprivation**: measured by an index capturing air quality, traffic crash rates, and housing in poor condition, obtained from the Department for Levelling Up, Housing and Communities (formerly known as the Ministry of Housing, Communities & Local Government) in 2013.

For all three outcomes, we will transform the data into deciles, with decile 1 corresponding to the worst-off 10% and decile 10 to the best-off 10%. This enables a more clear comparison of the efficacy of our model across such indicators.

3.1.2 Street-Level Imagery

street-level images have been obtained from Google Maps by querying the Google Street View API for each postcode in London. The API returns the unique identifier for the nearest available panorama image most recently taken by Google, if available. We have replicated previous papers' usage of the panoramic images, cutting them into four segments based on the camera direction (i.e. 0°, 90°, 180°, 270°) to cover a 360° view. However, while our reference paper queried the API for every single postcode in London, we were restricted on the number of free images we could procure for this project. Thus, we randomly selected one postcode from each LSOA and queried the API with that postcode, generating four distinct street-level images. As a result, we obtained 14,520 street-level images, in comparison to our reference paper's 119,238 images. Nonetheless, our approach still allowed us to select a set of images representative of all of London.

The following is an example set of 4 images from our query of the Google Street View API on March 20, 2023:

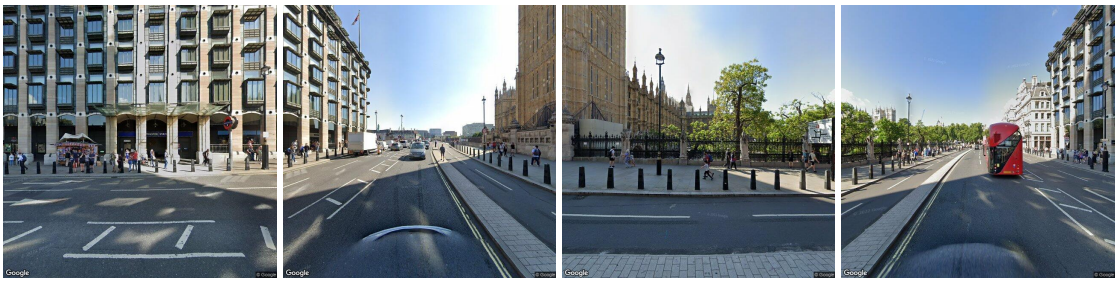

Figure 1: Images from Bridge St, SW1A 2JR (between Westminster Underground Station and Houses of Parliament)

3.1.3 Satellite Imagery

Satellite images have been obtained from Maxar satellites by querying the Mapbox API for each postcode in London, for which we also had a set of four street-level images, using the coordinates at the centroid of that postcode. Maxar has launched its flagship product - Vivid Basic - a global imagery basemap that offers up to 50 cm resolution. Our images are all taken in the years since 2020 and have size 600x600 pixels. This size and resolution of these images allowed us to preserve the quality of the images, while also covering a large enough area, so that our model could account for long-range dependencies not visible in the street-level images.

Figure 2: Satellite image from above Buckingham Palace, SW1A 1AA

3.2 Model Architecture

Our primary contribution to the existing literature has been a modernisation of the architecture employed by Suel et al. [2021], which exclusively uses a VGG network to create feature representations. Instead of the VGG network, our model uses the SWIN Transformer as described below.

The SWIN Transformer uses shifted windows to make several gains relative to standard Vision Transformer models [Liu et al., 2021]. The shifted window approach enables self-attention between pixels in defined windows, while also allowing for cross window interaction. This significantly decreases computational complexity relative to having attention between all pixels of an image. The hierarchical windows also allow the transformer to detect features at various scales - another desirable property.
The overall structure of our network is:

- Begin with a frozen pretrained SWIN transformer model to process the street-level imagery and result in a vector representation of the image.
- Simultaneously use a frozen pretrained SWIN transformer to process the satellite images and give us a vector representation of these images.
- Concatenate the vector representations (which are jointly indexed by their postcode). Note that to make our training consistent, we omitted all postcodes not having four street-level images and one corresponding satellite image.
- Feed these concatenated representations through a linear classifier to get probabilities for each decile of our outcome variable of choice.

As for our model parameters, we use a learning rate of 0.0001 with the Adam Optimizer for training. Our loss function is cross entropy loss in order to be consistent with Suel et al. [2021]. In order to prevent overfitting we also implement a weight decay parameter of 0.0001 (which is also consistent with Suel et al. [2021]. Finally, we use a train-validation-test split of 80-10-10% to randomly split our dataset of (4 street-level images, satellite image, label) tuples.

While the overall scope of our project included fine-tuning the SWIN Transformer model on our training data, we have not implemented this section due to a lack of time and computational resources.

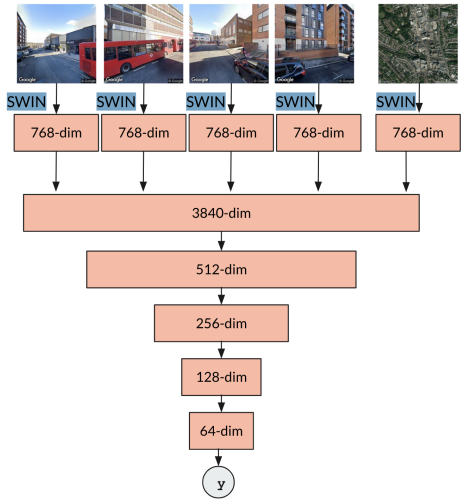

Figure 3: Model Architecture

4 Results

The results of our model on income, overcrowding, and environment deprivation indicators are included in the below table 1. For each indicator, we report the mean absolute error (MAE), which equals the mean absolute difference between the observed data and our predicted data. We include the MAE of our model trained solely on street-level imagery, as well as the MAE of our model trained with both street-level and satellite imagery to demonstrate the improvement in performance that is achieved by incorporating satellite data.

	Income	Overcrowding	Living Environment
SL Model MAE	2.35	2.46	2.21
SL+SAT Model MAE	1.97	1.86	1.67
Percent Improvement	16.2%	24.4%	24.4%

Table 1: Comparison of Model Results

Our results indicate a substantial improvement in accuracy (as measured by MAE) from the incorporation of satellite imagery. In addition, comparing the three economic indicators we analyze, we achieve fairly similar performance, with the living environment deprivation resulting in the smallest mean absolute error.

We further plot our mean absolute error on the validation data for each of these economic indicators, illustrating convergence of our model within 30 epochs of training.

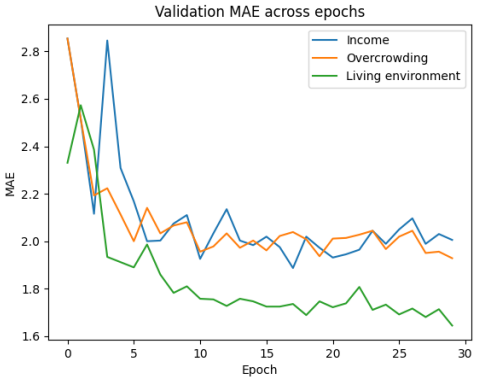

Figure 4: Validation MAE across epochs

To better visualize the results of our model, we utilize the python Geopandas library. We plot maps by postcode of the observed deciles for each of our economic indicators, alongside the predicted deciles determined via our model, as well as the mean absolute error for each indicator.

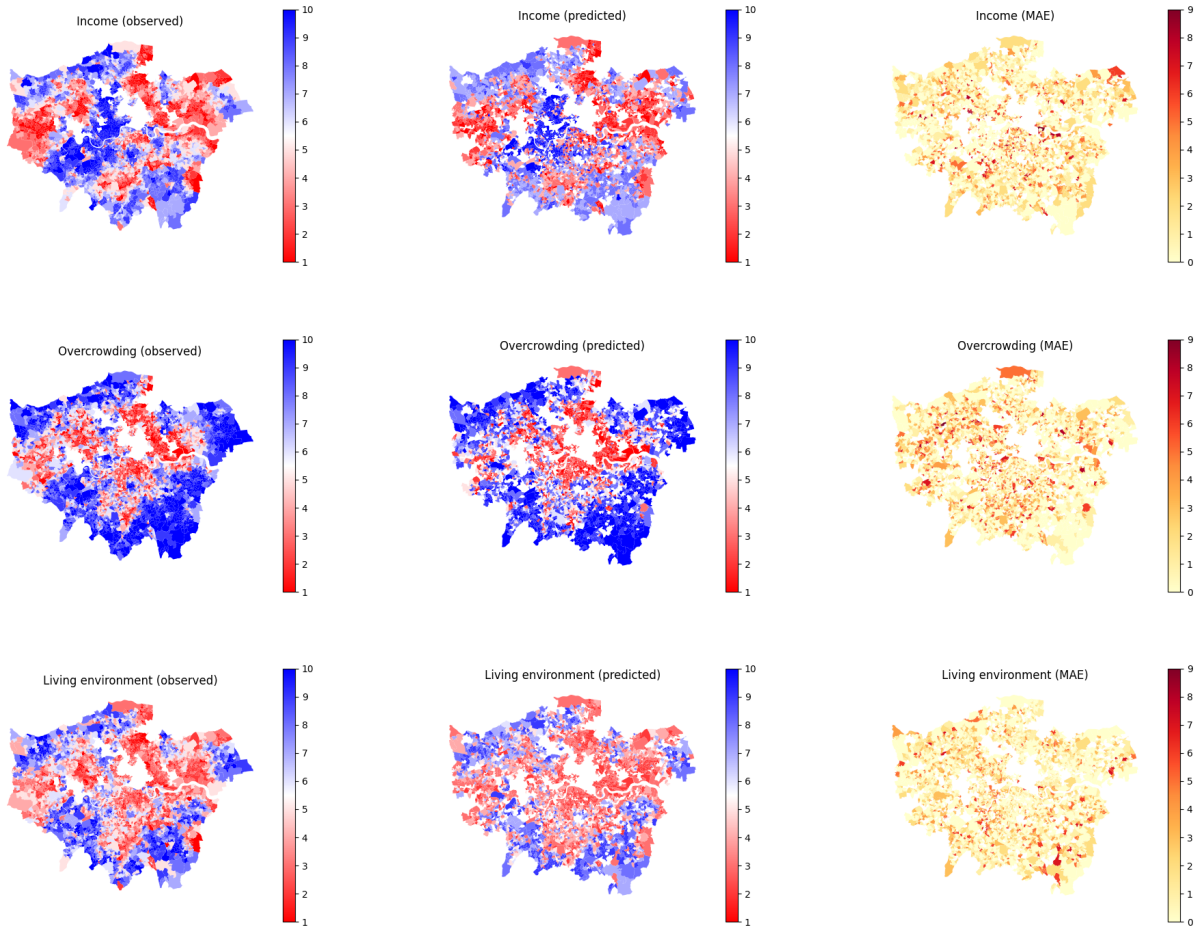

Figure 5: Visualizations of our results, generated using Geopandas

5 Discussions

5.1 Comparison of Economic Indicators and Model Performance

While the performance of our model is somewhat consistent across the economic indicators we have investigated, it is interesting to note that our model consistently performed best on the environmental deprivation data. This perhaps indicates that environmental deprivation is most assessible by visual imagery.

Analyzing the visualizations of predictions in Figure 5, we observe a clear resemblance between our observed and predicted labels across economic indicators. This reflects our model's performance, as we are clearly able to achieve predictions that are in close proximity to the observed values, as shown by our mean absolute errors. When analyzing the visualizations of our mean absolute errors, it is also interesting to note that there is no clear geographic correlation with error. This helps us validate that there are no biases in our model with respect to specific geographic regions.

5.2 Comparison of Results and Methodology to Reference Paper

With their training regimes and architecture, Suel et al. [2021] were able to achieve MAEs of 1.23, 1.30, and 1.17 on income, overcrowding, and living environment deprivation respectively. This is lower than our MAEs of 1.97, 1.86,

and 1.67 on these indicators. Initially, this seems to suggest that our architecture updates are having an adverse effect. However, our methodology differed from theirs in several other key areas that likely contributed to this difference:

1. The size of our training dataset is significantly smaller than Suel et al.'s dataset. While they use 119,238 street-level images, our dataset consists only of 14,520 such images. This gives us significantly less training samples, resulting in a higher mean absolute error.

2. Suel et al. [2021] use stratified random sampling based upon the decile of various indicators when creating their train-test splits, while we simply use random sampling. Through stratified sampling, they are able to make sure that their train and test data contain roughly the same proportion of examples in every decile - improving the inference capabilities of their model.

With greater computational availability and time, we would likely be able to carry out our training with the larger available training set and use stratified sampling on our own dataset - leading to results that are more comparable.

5.3 Applicability of Methodology to Low-Resource Urban Areas

While we are able to get a relatively high performance from this data on London, one should keep in mind that this does not directly translate to performance on street-level and satellite imagery from cities in underdeveloped areas. Depending on the specific features considered by the model, performance in less developed regions may vary substantially.

Therefore, fine tuning and testing this model on cities outside London is necessary to achieve an adequate performance in various regions. Again, the applications of our model are likely even more valuable in lower resource areas where acquiring accurate census data is more challenging. However, training such a model in these areas will be difficult because of the limited access to reliable data.

5.4 Temporal Data Limitations

A clear limitation in our project is due to data availability. As we utilize numerous data sources for extracting socioeconomic indicators in addition to street-level and satellite imagery, ensuring temporal similarity across this data is challenging. It is thus important to acknowledge that our satellite data was obtained from a substantially later date as compared with the census data (2020 vs 2011). Development within these years may skew the results of our findings.

5.5 Potential Extensions

The next step in improving the performance of our model involves the incorporation of more data. We currently utilize data for each LSOA in London, giving us approximately 3630 individual inputs into our model. By expanding our dataset to include inputs for each postcode in London, we will be able to input 4 street-level images and 1 satellite image for 29810 inputs. Expanding our dataset as such is likely to result in substantial performance improvements.

Furthermore, while we experimented with fine-tuning our SWIN model in addition to training our linear classifier, we collected our results using only the frozen pretrained SWIN model with these trained linear layers. This was primarily due to time constraints, as we were unable to determine the optimal hyperparameters to effectively tune the SWIN model to achieve improved performance. However, by incorporating fine-tuning of the SWIN model, we hope to see some improvements in the predictions made by our model.

Another possible consideration is a comparison of a ConvNeXt model architecture, targeted towards feature extraction for the satellite images. Convolutional Neural Networks have inductive biases that may allow them to better capture patterns in satellite imagery. While benchmarks may be misleading, it is possible that CNNs lead benchmarks on satellite classification precisely because of these choices. As such, comparing the performance of ConvNeXt - a modernized convolutional neural network that takes inspiration from vision transformers [Liu et al., 2022] - as an option for satellite imagery processing to that of SWIN could provide improvements to our results.

References

[Gebru et al., 2017] Gebru, T., Krause, J., Wang, Y., and Fei-Fei, L. (2017). Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 9.

[Jean et al., 2016] Jean, N., Burke, M., Xie, M., Lobell, D., and Erman, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353.

[Liu et al., 2021] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

[Liu et al., 2022] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986.

[Steele et al., 2017] Steele, J., Sundsoy, P., Pezzulo, C., Alegana, V., and Bird, T. (2017). Mapping poverty using mobile phone and satellite data. *Journal of the Royal Society*, 14.

[Suel et al., 2021] Suel, E., Bhatt, S., Brauer, M., Flaxman, S., and Ezzati, M. (2021). Multimodal deep learning from satellite and street-level imagery for measuring income, overcrowding, and environmental deprivation in urban areas. *Remote Sensing of Environment*, 257.

[Suel et al., 2019] Suel, E., Polak, J., Bennett, J., and Ezzati, M. (2019). Measuring social, environmental and health inequalities using deep learning and street imagery. *Sci Rep*, 9.