# Predicting Socio-economic Indicator Variations with Satellite Image Time Series and Transformer

Robin JARRY[1]
https://robin-j412.github.io/
Marc CHAUMONT[1,2]
https://www.lirmm.fr/~chaumont/
Laure BERTI-ÉQUILLE[3]
https://laureberti.github.io/website/
Gérard SUBSOL[1]
https://www.lirmm.fr/~subsol/

[1] LIRMM,
Univ. Montpellier, CNRS,
Montpellier, France.
[2] University of Nîmes,
Nîmes France.
[3] ESPACE-DEV,
Univ. Montpellier,
IRD, UA, UG, UR,
Montpellier, France.

**Abstract**

Monitoring local socio-economic variations is essential for tracking progress toward sustainable development goals. However, measuring these variations can be challenging, as it requires data collection at least twice, which is both expensive and time-consuming. To address this issue, researchers have proposed remote sensing and deep learning methods to predict socio-economic indicators. However, subtracting two predicted socio-economic indicators from different dates leads to inaccurate results. We propose a novel method for predicting socio-economic variations using satellite image time series to achieve more reliable predictions. Our method leverages both spatial and temporal information to enhance the final prediction. In our experiments, we observed that it outperforms state-of-the-art methods.

## 1 Introduction

Numerous approaches have been proposed to predict local socio-economic indicators[1] to create a socio-economic map [8, 14]. The most recent ones rely on deep learning and process a satellite image to obtain a scalar prediction for the indicator [2, 11]. The method consists of obtaining a satellite image of a large area, splitting the satellite image into a grid of "patches", and using a deep learning model to predict the socio-economic indicator for each grid cell. The cell dimension thus determines the resolution of the socio-economic map, which is usually around 2 kilometers.

Very often, it is the evolution that is valuable for practical use, to answer the question: is there an increase, decrease, or stagnation of the indicators? Generating two maps at two different dates and computing the difference as a prediction of the variation is misleading due to the level of imprecision (i.e., variance) of the prediction at each date [12, 13, 22].



[1] Example of socio-economic indicators include consumption expenditures, income, asset index, wealth index, ...

A solution to improve the level of precision has been proposed by Yeh *et al*. [22]. The idea is to directly predict the socio-economic indicator variation using a pair of satellite images as input.

In this paper, we propose to extend this idea by considering a *satellite image time serie* (SITS) as input to predict the variation of a socio-economic indicator over the range of dates associated with the sequence of images. To process the SITS, we rely on a video vision Transformer architecture [1].

We compared our approach to Yeh's on the same dataset composed of five African countries without any common "patch" between the training and testing sets. We achieved a 15% improvement in the correlation coefficient, showing that the use of spatio-temporal images, a Transformer model, and ad-hoc pre-training enhances the prediction of the indicator variation.

In Section 2, we review the recent state-of-the-art methods for predicting socio-economic indicators, as well as their variations. We also discuss the recent Transformer approaches dealing with SITS. In Section 3, we present the architecture used and how SITS are processed. In Section 4, we present the results of predicting socio-economic indicator variations across five African countries for various date ranges and discuss the benefits of using pretraining and spatio-temporal series with more than two images. Finally, we conclude in Section 5.

## 2 Related work

Socio-economists, ecologists, remote sensing researchers, computer scientists, and others have been working for a long time to produce dense socio-economic maps [7, 16]. From one or several sparse household socio-economic surveys, the dense map is obtained by predicting missing socio-economic values for the unsurveyed positions using machine learning [4], or deep learning [2, 11], in conjunction with a set of attributes [14, 15], or satellite images acquired during the day or at night (nighttime light images)[22], etc.

Recently, the use of satellite imagery combined with deep learning has led to state-of-the-art results in predicting socio-economic values for specific locations and dates [6, 14]. The pioneering work by Jean *et al*. [11] involved training a CNN to predict socio-economic indicators derived from LSMS[2] or DHS[3] survey data. To address the limited number of surveys used in their experiments, the authors employed pre-training on nighttime light images, followed by fine-tuning for the specific task.

A straightforward extension to address the small size of the training set was proposed by Yeh *et al*. [22], which involves training a CNN on a large dataset using available surveys across Africa. The reported average $r^2$ score is 0.70, which currently serves as the benchmark value using satellite images at a resolution of 30 meters, with 43 surveys in Africa.

These results are interesting and could potentially be improved, for example, by using multi-modalities and multi-domains. However, their use to estimate a *variation* (e.g., by computing the difference in predictions between two dates) is not reliable [12, 22]. Indeed, the level of noise in the predictions and the lack of temporal coherence result in noise that is significantly higher than the value being estimated (i.e., the socio-economic variation). To address this variation prediction issue, Yeh *et al*. [22] proposed to use, as input of their CNN network, the two satellite images associated with the dates forming the time range of interest.

[2] https://www.worldbank.org/en/programs/lsms
[3] https://dhsprogram.com/Data/

This approach has the advantage of including the temporal dimension in the problem and allowing joint spatio-temporal training.

Going further, we propose to use the entire sequence of satellite images composing the time range of interest, instead of only a pair of images. Additionally, we switch to a Transformer neural network, as it is well-suited for sequential data. It is worth noting that Pettersson *et al*. [17] recently proposed an approach that uses SITS for predicting a sequence of socio-economic indicators within a given time range. While their approach demonstrates that using SITS to predict multiple socio-economic indicators at different dates is effective, it lacks an evaluation of the reliability of predicting *variation* between two dates.

As mentioned previously, our solution relies on a Transformer. The use of Transformers has shown significantly better performance compared to convolutional neural networks for image processing [6]. The advantage of such architecture is primarily due to the attention mechanism, which allows the network to better focus on the pixels of interest. By extension, solutions have been proposed to process videos [1]. In that case, the benefit comes from the ability to consider both local and global information when dealing with a series of images.

In the remote sensing community, Transformers were initially adapted at the pixel level, i.e., with time series of reflectance values [8, 18], and then to a small set of pixels [24]. Recently, the spatio-temporal TSViT architecture has been proposed [21]. The particularity of this architecture lies in its application of temporal encoding first, followed by spatial encoding. This order change was made to better align with the task of crop type prediction, as the spatial context might sometimes be irrelevant or misleading for crop type mapping. Our architecture retains the original ordering, with spatial encoding first, followed by temporal encoding.

# 3 A Transformer-based Method

Given two dates $t_0$ and $t_1$ defining a *time range of interest* and thus a series of satellite images, we aim to predict the socio-economic *variation*. In our experimental setup, each date represents a year with one image per year. As mentioned before, we rely on a Transformer, which we briefly recall in Section 3.1. To account for the time dimension, we feed this network with a series of fixed-size (i.e., the series consists of 21 satellite images) and mask the irrelevant images. We describe this masking in Section 3.2. The survey used and the method to define a scalar representing the socio-economic *variation* are described in Section 3.3. We finally describe the SITS in Section 3.4.

## 3.1 ViViT architecture

To process a SITS we chose to use the *video vision Transformer* (ViViT) [1] with a factorized spatio-temporal encoder. The architecture is illustrated in figure 1.

Each image in the series is divided into patches based on a regular $p \times p$ grid, and each patch is projected into a latent space of dimension $d$. A positional encoding is then added to each projected patch to indicate its position within the grid. Additionally, a class token is introduced to represent the entire sequence. Each encoded image, now a sequence of vectors, is processed in parallel by a spatial Transformer module.

The class token from each sequence of vectors is assembled into a new sequence, consisting of the latent projections of each image, along with a new (temporal) class token. Finally, the temporal Transformer processes this sequence of vectors to generate the final prediction in the model's last layer.

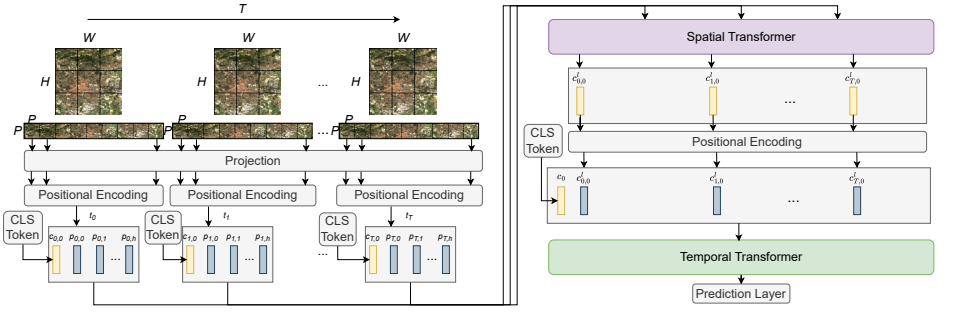


Figure 1: Factorised spatio-temporal Transformer (ViViT).

## 3.2 Masking the irrelevant images of the series

One must inform the ViViT transformer to account for the two dates for which the indicator variation needs to be predicted. A simple and effective solution is to always process a fixed-size sequence of images (in our case, 21 images) and to mask the images outside the time range of interest by making them black. This masking is applied during both training and testing.

An example of such a masked SITS is shown in Figure 2. In this example, the socio-economic indicator variation measured is $-0.2693$, between the time range of interest $t_0 = 2008$ and $t_1 = 2012$. Thus, images before 2008 and after 2012 are masked to let the model predict on the given time range.

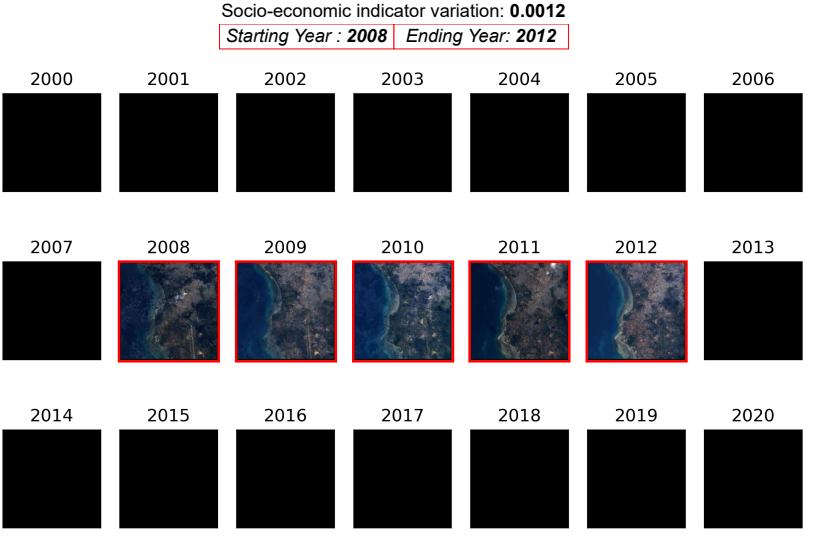


Figure 2: Masking strategy to encode the starting and ending survey dates for the model.

### 3.3 Scalar representation of the socio-economic *variation*

A variation of a socio-economic indicator can only be obtained if surveys have been conducted on two dates at the same location. To the best of our knowledge, LSMS surveys [23], are the only source allowing for the computation of socio-economic indicator variation in Africa. In our experiment, we used the LSMS surveys conducted between 2005 and 2016 in five countries: Nigeria, Ethiopia, Tanzania, Uganda, and Malawi.

A survey involves questioning households with quantifiable questions. The answers are averaged across small sets of households (within each set, households are often close in distance in urban areas and farther apart in rural areas). These sets are referred to as "clusters" in most scientific literature, and the answers from each cluster are stored in tabular form.

To construct the socio-economic indicator variation, Yeh et al. [23] selected a set of asset-related questions from the surveys. For each "cluster", the difference between the two vectors containing the survey answers at the two survey dates is computed. Then, the difference vector is projected onto the principal component obtained from the PCA to derive a scalar (i.e., the socio-economic variation indicator). These "scalars" are available in SustainBench [23], and we directly used them without any additional processing.

Note that before the publication on the SustainBench public website, the associated locations of the surveys were anonymized by randomly altering latitude and longitude within a maximum of 5 km. This spatial granularity (i.e., resolution) implies that the socio-economic maps we produce have a resolution of at best 10 km². We thus use a series of satellite images with a spatial resolution close to a 10 km² surface.

### 3.4 Satellite images dataset

We collected 1-year Landsat-7 median composite[4] series of satellite images from 2000 to 2020 for every location in SustainBench [23]. Each image has six bands: blue, green, red, near-infrared, and short-wave infrared 1 and 2. Note that among various satellites, Landsat-7 is one of the few that has covered the entire Earth for over 20 years. However, its resolution of 30 m limits the detection to large structures only.

The SITS have a 1-year frequency, with no information overlap between two consecutive composite images. Additionally, we believe that using 1-year composite images can better capture variations between different time steps, even if these images are noisier than 3-year composites.

Note that similarly to [22], we choose to work with an image size of 224×224 pixels. This image size, with Landsat-7 images of 30 m of resolution, corresponds to a 6.72 km² surface, so we may miss the exact location of the surveyed households (see Section 3.3).

## 4 Experiments, results, and discussion

### 4.1 Compared models

As mentioned in the introduction, it is misleading, especially given the nature of our data, to use socio-economic spatial predictors to estimate variation with a simple difference. We thus compared our approach to the only available method, which is Yeh's method [22].

[4] A composite image is a combined image created by merging several raw satellite observations to minimize individual errors and improve accuracy and reliability.

We used a ResNet-18 architecture, as in [22], to represent Yeh's approach. It processes a channel-wise stacked pair of images corresponding to the starting and ending years of the survey. The model weights are initialized randomly.

For our approach, we used a ViViT architecture, with out-of-time-range images masked as mentioned in section 3.2. The model weights are initialized randomly.

We also used a ViViT architecture pre-trained to predict nighttime light time series from a large-scale dataset covering Africa and the Middle East, with a similar process than in [10]. Technically, after the pre-training, the last layer which produced 21 predictions was replaced by a layer producing socio-economic variation predictions. It is important to note that the 5 countries in our dataset were excluded from the pre-training dataset, ensuring no overlap between the two datasets. The ViViT model was thus entirely fine-tuned on our socio-economic indicator variation dataset and it also employs the satellite image masking principle.

Note that we evaluated the concept of pre-training, which is widely adopted in many off-the-shelf networks because it usually improves downstream task performance. Additionally, pre-training to predict nighttime light time series is certainly beneficial due to the known correlation between nightlight values and socio-economic indicators [7, 9, 20].

## 4.2 Evaluation and metrics

The final dataset contains 1,665 SITS with their associated scalars representing the socio-economic indicator variations.

We perform 5-fold cross-validation with no spatial or temporal overlap between folds. Some locations may appear multiple times in the dataset, as several survey rounds occurred at the same location but for different dates (e.g. $t_0$, $t_1$, and $t_2$), allowing for the computation of different variations (e.g. $(t_0,t_1)$, $(t_1,t_2)$, and $(t_0,t_2)$). We ensure that such locations belong to the same cross-validation fold to prevent spatial overlap between folds.

For each experiment, we report the average and standard deviation of several scores across the folds, namely the mean absolute error (*MAE*), root mean squared error (*RMSE*), Pearson squared correlation coefficient ($r^2$), and coefficient of determination ($R^2$).

Note that $r^2$ reports the slope of the best-fitted line between observed and predicted values, while the $R^2$ reports the goodness of fit (implicitly with the identity line).

## 4.3 Experimental settings

Each model is trained for 250 epochs, with a batch size of 16 and a learning rate of $5 \times 10^{-4}$. We employed the mean squared error (MSE) as the loss function. Hyperparameters were optimized by minimizing the loss on a validation set, which was a randomly selected subset of the training set. The training set comprised four folds, with the fifth fold reserved for testing the final model.

We did not use any augmentation techniques or advanced hyperparameter tuning, but we conducted several experiments with varying batch sizes and learning rates, finding these values to be well-suited. All models have approximately 10 million learnable parameters.

We trained our models on 4 Nvidia V100 GPUs, with each training session lasting between 2 and 3 hours. The reader should note that there was a bottleneck in data loading, which slowed down the training and inference.

## 4.4 Results

Table 1 reports the four scores of the different models. The model named ViViT is trained from scratch on the adapted SustainBench socio-economic indicator variation dataset. The ResNet-18 model is a reproduction of the model used in [22]. The ViViT-pretrain model is pretrained to predict nighttime light time series and fine-tuned on the socio-economic indicator variation dataset.

| | $MAE \downarrow$ | $RMSE \downarrow$ | $r^2 \uparrow$ | $R^2 \uparrow$ |
|---|---|---|---|---|
| ViViT | $0.482^{\pm 0.015}$ | $0.637^{\pm 0.020}$ | $0.263^{\pm 0.057}$ | $0.245^{\pm 0.054}$ |
| ResNet-18 [22] | $0.528^{\pm 0.019}$ | $0.687^{\pm 0.032}$ | $0.182^{\pm 0.054}$ | $0.122^{\pm 0.061}$ |
| ViViT-pretrain | $0.460^{\pm 0.013}$ | $0.366^{\pm 0.020}$ | $0.328^{\pm 0.063}$ | $0.319^{\pm 0.065}$ |

Table 1: Performance scores of different models on socio-economic indicator variation prediction (average and standard deviation noted as exponent).

The *MAE* and *RMSE* show that our model is better than Yeh's state-of-the-art model [22]. We confirm this superiority with the correlation scores, with a $r^2$ score of 0.263 for our model compared to an $r^2$ score of 0.182 for ResNet-18. We observe an improvement of our model ($r^2 = 0.328$) when pretrained to predict nighttime light time series. Different from the *MAE* and *RMSE* scores, we note a large standard deviation for correlation scores $r^2$ and $R^2$, illustrating an important score difference between folds.

Due to the small size of the dataset, we are unable to perform any generalization tests, leaving us without indications of the model's performance outside the spatial and temporal extent of the dataset. However, previous experiments on nighttime light time series have shown that models are better when predicting long-term evolutions (i.e., 15 to 20 years) [10]. Supported by these results, we illustrate, in figure 3, the socio-economic indicator variation prediction of the ViViT-pretrain model for the period from 2000 to 2020.

To further the analysis, we examined the coherence between our socio-economic indicator variations and the variations of the *international wealth index* (IWI) in several areas. The IWI is obtained from [19] on the GlobalDataLab[5] at the district administrative. The IWI is computed as described in Section 3.3 and quantifies the wealth of a district at a given point in time. To estimate variation, we rely on the differences in the IWI between two dates. For most districts, there are 2 to 3 IWI values at different years, obtained from DHS surveys.[6] Smits *et al.* [19] proposed a linear interpolation at a 1-year frequency to estimate IWI values when data is unavailable for a specific date. Since the values of the IWI and the socio-economic indicator variations predicted by our model are on different scales, we will simply compare the variation intensity of the IWI with our predicted values.

Figure 4 is a zoomed-in view of Figure 3, showing our socio-economic indicator variation predictions from the ViViT-pretrain model for the period 2000 to 2020, focusing on two distinct areas. Left figure 4 displays the West Nile district in Uganda, where our model predicts generally negative values for the socio-economic indicator variation. In this district, the IWI shows a slight increase (rising from 14.7 in 2006 to 19.4 in 2016). Right figure 4 shows an area around Kampala, Uganda. Here, our model predicts values close to zero for

[5] https://globaldatalab.org/areadata/maps/iwi/

[6] It is impossible to calculate variation at the cluster level, as DHS surveys are cross-sectional; new clusters are sampled each time a DHS survey round is conducted in a specific country.

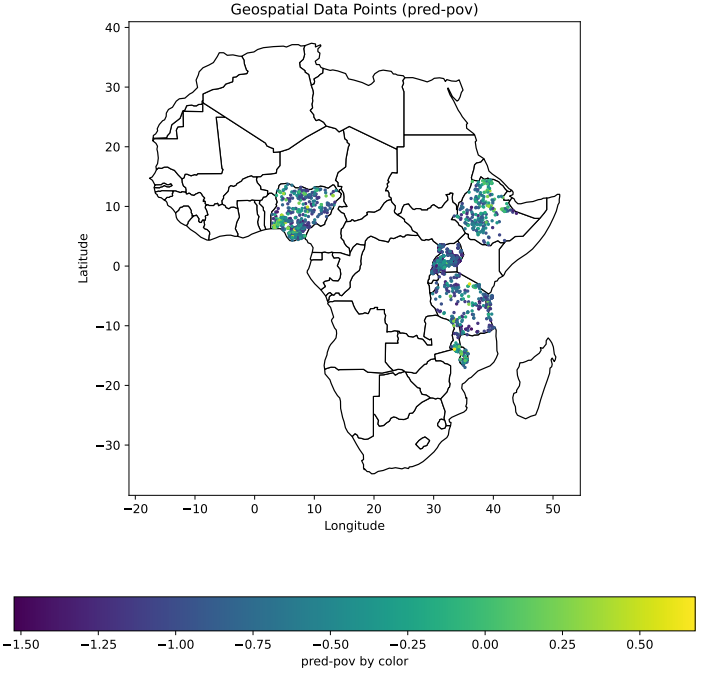


Figure 3: ViViT-pretrain prediction on the time period 2000-2020.

the socio-economic indicator variation, which aligns with the IWI also being close to zero, indicating consistent urban development as noted by [3].

## 4.5 Discussions

These results highlight the advantage of using a Transformer-based spatio-temporal architecture in conjunction with SITS to predict socio-economic indicator variation. We hypothesize that the ViViT architecture benefits from the intermediate images in the SITS taken between the starting and ending dates. These experiments also indicate that 1-year composites can be used instead of 3-year composites, even though they are much noisier. One could argue that 3-year composites may improve the results of the ViViT model as they contain less noise. However, we believe that 3-year composites will excessively smooth the changes between time steps, making it difficult for the model to learn them. Further experiments need to be conducted to confirm this assumption. Using a pretrained version of this model further improves the results, suggesting that the evolution of nighttime light is a good initialization for predicting variation of socio-economic indicators.

We note that all $r^2$ scores are low, indicating that practical use of our model is challenging.

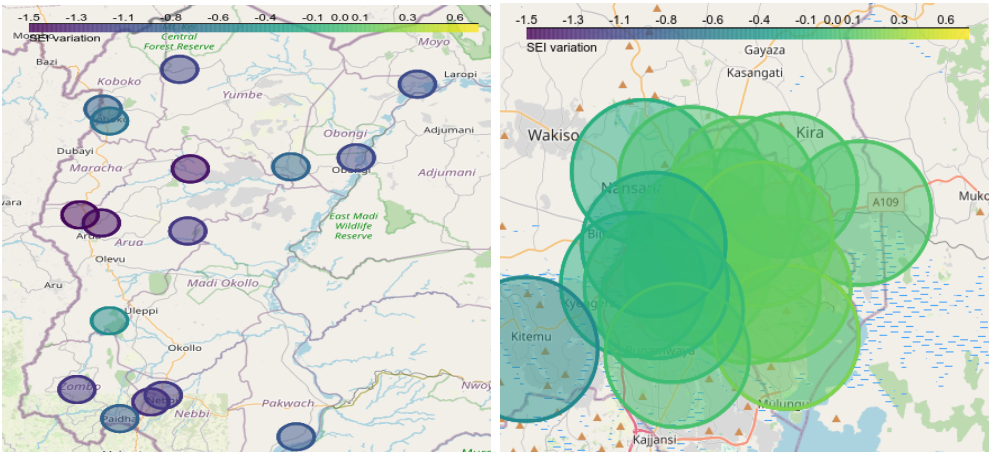


Figure 4: Examples of socio-economic indicator variations predicted on a given area. Left is West Nile district, Uganda. Right is Kampala city, Uganda.

The low $r^2$ values can be attributed to the short time range between two survey rounds, which is insufficient to observe significant changes in the satellite image time series, making the task very difficult for the model. To the best of our knowledge, this is unfortunately the only existing dataset with socio-economic indicator variations.

Due to the small size of the dataset, we are unable to perform any generalization tests, leaving us without hints about the model's performance outside of the spatial and temporal extent of the dataset. However, we hypothesize that pretraining on ILN data may enhance extrapolation capabilities. Moreover, as previously mentioned, a long time frame (i.e., 15 to 20 years) facilitates the observation and prediction of evolutions. Consequently, we believe that over long periods and within neighborhood spatial regions, the performance is likely superior to that reported here

# 5 Conclusion

In this paper, we present a novel approach for predicting variations in socio-economic indicators using satellite image time series. The key innovation of this method is its ability to consider both spatial and temporal contexts, allowing the extraction of relevant information from both dimensions of the data. We employed the Transformer model, which is based on the attention mechanism and excels at learning spatial and temporal dependencies in satellite image time series.

We evaluated the proposed method using a dataset from SustainBench, a collection of benchmark datasets designed to monitor sustainable development goals. We also compared our method with the state-of-the-art approach by [22]. The results demonstrate that taking into account spatio-temporal dependencies significantly improves the prediction of socio-economic variations, outperforming the existing state-of-the-art method.

However, the paper also highlights some limitations. Firstly, the obtained $r^2$ scores are too low for practical application. Secondly, due to the limited size of the dataset, generalization tests could not be conducted, leaving the model's performance in other regions of the world uncertain.

# References


[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, October 2021.

[2] Boris Babenko, Jonathan Hersh, David Newhouse, Anusha Ramakrishnan, and Tom Swartz. Poverty Mapping Using Convolutional Neural Networks Trained on High and Medium Resolution Satellite Images, With an Application in Mexico. In *NeurIPS 2017 Workshop on Machine Learning for the Developing World*, Long Beach, CA, November 2017. doi: 10.48550/arXiv.1711.06323. URL http://arxiv.org/abs/1711.06323.

[3] Fred Bidandi and John J. Williams. Understanding urban land, politics, and planning: A critical appraisal of kampala's urban sprawl. *Cities*, 106:102858, 2020. ISSN 0264-2751. doi: https://doi.org/10.1016/j.cities.2020.102858. URL https://www.sciencedirect.com/science/article/pii/S0264275120312063.

[4] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, November 2015. doi: 10.1126/science.aac4420. URL https://www.science.org/doi/full/10.1126/science.aac4420. Publisher: American Association for the Advancement of Science.

[5] Guanghua Chi, Han Fang, Sourav Chatterjee, and Joshua E. Blumenstock. Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3):e2113658119, January 2022. doi: 10.1073/pnas.2113658119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2113658119.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *In: Proceedings of International Conference on Learning Representation.*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[7] C. D. Elvidge, K. E. Baugh, E. A. Kihn, H. W. Kroehl, E. R. Davis, and C. W. Davis. Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. *International Journal of Remote Sensing*, 18(6):1373–1379, April 1997. ISSN 0143-1161. doi: 10.1080/014311697218485. URL https://doi.org/10.1080/014311697218485.

[8] Vivien Sainte Fare Garnot, Loic Landrieu, Sebastien Giordano, and Nesrine Chehata. Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention, November 2019. URL http://arxiv.org/abs/1911.07757.

[9] Tilottama Ghosh, Sharolyn J. Anderson, Christopher D. Elvidge, and Paul C. Sutton. Using Nighttime Satellite Imagery as a Proxy Measure of Human Well-Being. *Sustainability*, 5(12):4988–5019, December 2013. ISSN 2071-1050. doi: 10.3390/su5124988. URL https://www.mdpi.com/2071-1050/5/12/4988.

[10] Robin Jarry, Marc Chaumont, Laure Berti-Équille, and Gérard Subsol. Comparing spatial and spatio-temporal paradigms to estimate the evolution of socio-economical indicators from satellite images. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 5790–5793, 2023. doi: 10.1109/IGARSS52108.2023.10282306.

[11] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, August 2016. doi: 10.1126/science.aaf7894. URL https://www.science.org/doi/10.1126/science.aaf7894.

[12] Lukas Kondmann and Xiao Zhu. Measuring Changes in Poverty with Deep Learning and Satellite Imagery. In *ICLR Proceedings*, 2020.

[13] Lukas Kondmann, Hannes Taubenböck, and Xiao Xiang Zhu. Blinded by the Light: Monitoring Local Economic Development Over Time With Nightlight Emissions. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 5708–5711, July 2021. doi: 10.1109/IGARSS47720.2021.9554428.

[14] Kamwoo Lee and Jeanine Braithwaite. High-resolution poverty maps in Sub-Saharan Africa. *World Development*, 159:106028, November 2022. ISSN 0305-750X. doi: 10.1016/j.worlddev.2022.106028. URL https://www.sciencedirect.com/science/article/pii/S0305750X22002182.

[15] Jeaneth Machicao, Alison Specht, Danton Vellenich, Leandro Meneguzzi, Romain David, Shelley Stall, Katia Ferraz, Laurence Mabile, Margaret O'Brien, and Pedro Corrêa. A deep-learning method for the prediction of socio-economic indicators from street-view imagery using a case study from brazil. *Data Science Journal*, Feb 2022. doi: 10.5334/dsj-2022-006.

[16] Abdisalan M. Noor, Victor A. Alegana, Peter W. Gething, Andrew J. Tatem, and Robert W. Snow. Using remotely sensed night-time light as a proxy for poverty in Africa. *Population Health Metrics*, 6(1):5, October 2008. ISSN 1478-7954. doi: 10.1186/1478-7954-6-5. URL https://doi.org/10.1186/1478-7954-6-5.

[17] Markus B. Pettersson, Mohammad Kakooei, Julia Ortheden, Fredrik D. Johansson, and Adel Daoud. Time series of satellite imagery improve deep learning estimates of neighborhood-level poverty in africa. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6165–6173. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/684. URL https://doi.org/10.24963/ijcai.2023/684. AI for Good.

[18] Marc Rußwurm and Marco Körner. Self-attention for raw optical Satellite Time Series Classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:421–435, November 2020. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2020.06.006. URL https://www.sciencedirect.com/science/article/pii/S0924271620301647.

[19] Jeroen Smits and Roel Steendijk. The International Wealth Index (IWI). *Social Indicators Research*, 122(1):65–85, May 2015. doi: 10.1007/s11205-014-0683-x.

[20] Paul C Sutton, Christopher D Elvidge, and Tilottama Ghosh. Estimation of Gross Domestic Product at Sub-National Scales using Nighttime Satellite Imagery. *International Journal of Ecological Economics and Statistics*, 8(S07):5–21, 2007.

[21] Michail Tarasiou, Erik Chavez, and Stefanos Zafeiriou. Vits for sits: Vision transformers for satellite image time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10418–10428, June 2023.

[22] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11(1):2583, May 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-16185-w. URL https://www.nature.com/articles/s41467-020-16185-w.

[23] Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David B. Lobell, and Stefano Ermon. SustainBench: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*, 2021.

[24] Yuan Yuan and Lei Lin. Self-Supervised Pretraining of Transformers for Satellite Image Time Series Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:474–487, 2021. ISSN 2151-1535. doi: 10.1109/JSTARS.2020.3036602. Conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.