COMPARING SPATIAL AND SPATIO-TEMPORAL PARADIGMS TO ESTIMATE THE EVOLUTION OF SOCIO-ECONOMICAL INDICATORS FROM SATELLITE IMAGES

Robin Jarry[1], Marc Chaumont[1, 2], Laure Berti-Équille[3], Gérard Subsol[1]

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

ABSTRACT

In remote sensing, deep spatio-temporal models, *i.e.,* deep learning models that estimate information based on Satellite Image Time Series obtain successful results in Land Use/Land Cover classification or change detection. Nevertheless, for socioeconomic applications such as poverty estimation, only deep spatial models have been proposed. In this paper, we propose a test-bed to compare spatial and spatio-temporal paradigms to estimate the evolution of Nighttime Light (NTL), a standard proxy for socioeconomic indicators. We applied the test-bed in the area of Zanzibar, Tanzania for 21 years. We observe that (1) both models obtain roughly equivalent performances when predicting the NTL value at a given time, but (2) the spatio-temporal model is significantly more efficient when predicting the NTL evolution.

1. INTRODUCTION

Satellite devices are designed to orbit for several years, sometimes up to one or two decades. They can provide a long-term view of the Earth's surface offering a rich source of information for monitoring socioeconomic activities.

In particular, these last years, many research works about inferring poverty indicators based on satellite images have been proposed (see e.g. [1, 2]). Methods based on deep-learning techniques give interesting results, allowing to estimate poverty with a quite good accuracy (R^2 around 0.7 to 0.8 w.r.t. the ground truth values) in countries where organizing on-site surveys or studies is complicated and sometimes unfeasible. Nevertheless, recent advances pointed out the difficulties of estimating *poverty evolution*, that is concluding if poverty is increasing or decreasing over a period of time [2, 3, 4].

We hypothesize that this is due to the imprecision in poverty estimation at a given time which prevents from just subtracting values between two times to conclude over the period. All the proposed methods are trained only on spatial data, that is a set of images defined by their coordinates with an associated poverty value which was measured on-site around the area. One idea to improve these results could be to take into account the temporal dependency. For example, in [2], the authors try to learn and estimate an index of poverty evolution between two times according to the pair of corresponding images. We can extend this concept by training methods on *Satellite Image Time Series* (SITS), composed of a set of images at the same position but at several consecutive times with the associated socioeconomic indicators measured at the same times. This idea is supported by the fact that for other applications such as Land Use/Land Cover or change detection, the line of works on spatio-temporal methods [5, 6] shows noticeably successful results.

In this article, we aim to assess if the spatio-temporal estimation paradigm could be better than the spatial one for estimating socioeconomic evolution from satellite images. In Section 2, we describe a test-bed where we use Transformer-based techniques and where we remove the problem of the sparsity of socioeconomic indicators by using the NTL, a standard proxy. Then, we present experiments and discuss the results in Section 3.

2. A TEST-BED TO COMPARE SPATIAL AND SPATIO-TEMPORAL PARADIGMS

We aim to set up a general and flexible test-bed. We decided to work with the Transformer architecture, as it is a state-of-the-art methodology that has been adapted for both spatial and spatio-temporal data. We choose to work with Landsat optical images which cover with an acceptable resolution a wide time window. As labels, we use NTL data as they are also available for a wide time window and are considered as a standard proxy for socioeconomic indicators [7].

2.1. Transformer-Based Models

Transformer models come from Natural Language Processing research and are designed to process sequences of group of letters. They have been adapted to process images, then sequences of images. In the remote sensing area, [8] is a first attempt to use Transformer for SITS classification. Then, this work was extended at the image level by [6] and shows promising results. In this paper, the authors studied the

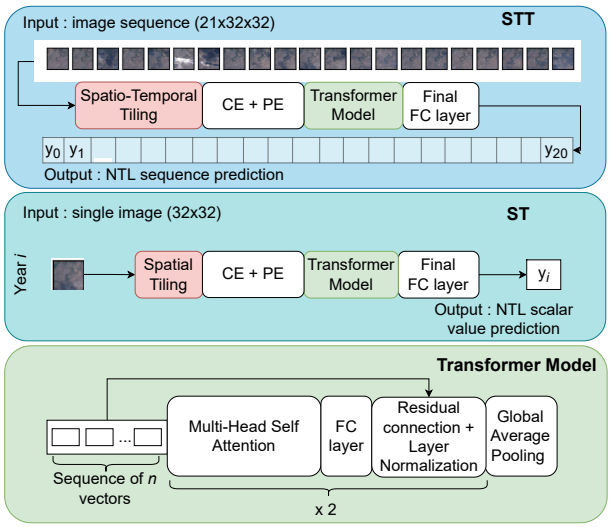

Fig. 1. Spatio-Temporal Transformer (STT), Spatial Transformer (ST), Transformer backbone architecture. CE and PE stands for convolutional embedding and positional encoding, respectively. Both ST and STT models share the same Transformer architecture, highlighted in green.

land use classification in California in 2019 with Sentinel-2 images and proposed a SITS Transformer pre-trained on a pretext task and fine-tuned on a crop classification task. Our work rely on the same transformer architecture (without the pre-training stage), but is performed on time series of satellite images that covers a wider time period (2 decades) on which we assume that the socioeconomic situation is smoothly evolving. The architecture of our models is depicted in Figure 1. Both models process the same images: the Spatial Transformer (ST) treats each image and NTL value independently, while the Spatio-Temporal Transformer (STT) maps one SITS to one NTL sequence. We detail below how both models process the data.

Spatial Transformer. To make sequential data from satellite images, we follow the approach of [9], a spatial tiling with 2 steps. First, the image is divided into small (8×8) non-overlapping tiles, then, the tiles are concatenated according to their order in the image: the first (resp. last) tile corresponds to the tile in the upper left (resp. lower right) corner of the image. Finally, each tile in the sequence is processed by a convolutional layer and a positional encoding layer.

Spatio-Temporal Transformer. To embed a SITS into a sequence of vectors, we follow the straightforward method as in [10] referred to spatio-temporal tiling in Figure 1. Each image in the SITS is divided into small (8×8) tiles and rearranged into a global spatio-temporal tile sequence. Then, this sequence is fed to a convolutional layer and a positional encoding layer. As our SITS have n time steps with images of $h \times w$ pixels each, this results in a sequence of $n \times \frac{h}{8} \times \frac{w}{8}$ vectors.

Then, in the ST model, the sequence of vectors is a representation of a single image and the weights of the transformer model are computed only considering spatial dependencies. The sequence processed by the STT model is made with an entire image sequence, then the attention weights focus on spatial and temporal patterns at the same time. Finally, the output sequence of vectors is averaged to produce a single vector representation, and fed to a fully connected layer. In the end, the ST and the STT models have respectively 280k and 285k parameters.

2.2. Satellite Image Time Series & Nighttime Lights

Satellite Image Time Series. There exist only few sensors that can deliver long-term SITS. Among them, we choose to use Landsat-7 images for several reasons. It covers a 25-year time period, capturing the full Earth's surface every 16 days, making it possible to create cloud-free composites per each year, almost everywhere on Earth. Additionally, several tools exist[1] to download and preprocess the data. We collect SITS of Landsat-7 images that overlap the study area from 2000 to 2020 (see section 3.1). We rely on the Google Simple Composite algorithm with default parameters to make a cloud-free composite of Landsat-7 images for each year. This algorithm performs top-of-atmosphere conversion, computes a cloud-score per pixel and deliver the median pixel values of the less cloudy pixels. We choose to use 6 spectral channels, which are blue, green, red, NIR, SWIR-1, and SWIR-2 as it is a classical choice when working with Landsat data.

Nighttime Lights. There are mainly two sources that collect NTL covering a long-term period over the Earth's surface. The former is data acquired by the DMSP/OLS sensors. It consists of several satellites that cover the period from 1992 to 2013. The latter is data acquired with the VIIRS sensor that covers years from 2013 to today. As the types of sensors capture NTL with very different modalities, a cross-sensor calibration is needed. Various methods exist, here we rely on the work of [11]. The authors used an auto-encoder model to convert the DMSP NTL to a VIIRS-like NTL. To train their model, they first intercalibrated the DMSP/OLS data, then used both NTL data to learn the transformation DMPS NTL to VIIRS NTL. Moreover, the dataset is publicly available and consists of 21 world maps of VIIRS-like NTL.

3. EXPERIMENTS

3.1. Study Area and Patching Strategy

Figure 2 depicts our data collection and preprocessing. We choose to study a large neighborhood of Zanzibar, an island beyond the Tanzanian coastline, as it is a sample of the study area of [2]. The time period is 2000 to 2020 (included) as it is the largest we can obtain considering that we used the NTLs produced by [11]. We collected one single Landsat-7

[1] website: `https://earthengine.google.com/`

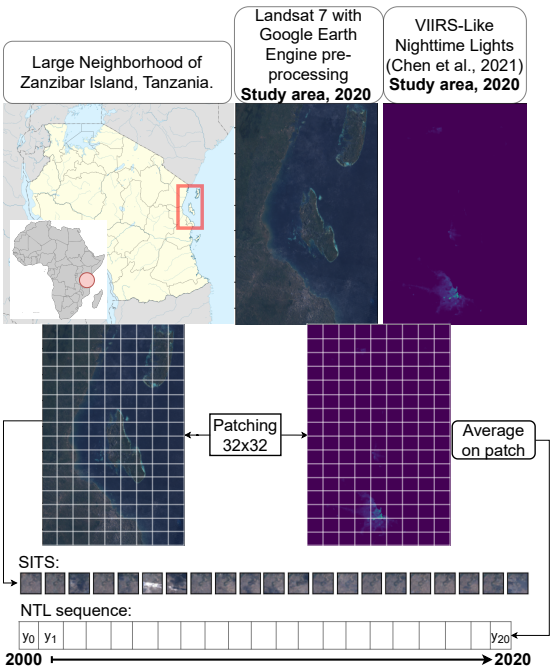

Fig. 2. Data collection and preprocessing pipeline. The grid is much larger on the figure, for visualization purposes.

composite for each year, leading to $n = 21$ images (8,692 $\times$ 5,505 pixels each) spanning the study area, to fit the time sampling in [11]. Then, the Landsat images are patched according to a regular grid into sequences of sub-images of size $h \times w = 32 \times 32$ pixels that cover approximately a 1km^2 area, as the resolution of Landsat images is 30 meters. To make a one-to-one correspondence between a patch sequence and the corresponding NTL sequence, we extract the average NTL sequence measured on the same spatial area covered by the patch sequence. Then, we exclude patch sequences that overlap the Zanzibar island (but kept them for illustration). Other patch sequences and NTL sequences that are outside Zanzibar are split into 6 folds, for cross-validation.

3.2. Experimental Details

We performed a 6-fold cross-validation with a fixed validation fold use to select the best model configuration. In the end, it gives 5 training and testing phases. Our dataset is imbalanced as most of the NTL values are zeros. This causes the model falling into a trivial solution at training time. Thus, we filter the training dataset such that, all the pairs containing at least a non-null NTL in the sequence are kept. Additionally, 500 randomly selected examples with a null NTL sequence are added for each fold. We use the Mean Squared Error (MSE) as the loss function. We choose to set the number of epochs to 200 and the batch size to 64 for both models. The training converge for both models, ensuring their best performances. The learning rate is set to 5×10^{-4} and the embedding dimension is set to 72. We use the *coefficient of determination* R^2 for the comparison. First, we compute the R^2 for each year individually and average the results. Then, we evaluate how both models predict the NTL evolution (= difference) between two time steps spaced by $\Delta t \in \{1, 10, 15\}$ years. For a given Δt, we average the R^2 scores of all evolutions in the time period, and report the results in Table 1.

R^2	year	$\Delta t = 1$	$\Delta t = 10$	$\Delta t = 15$
TRF	0.38 [0.06]	0.08 [0.12]	0.20 [0.11]	0.30 [0.03]
ST	0.59 [0.14]	-3.38 [2.56]	-0.36 [0.68]	-0.05 [0.31]
STT	0.71 [0.04]	0.15 [0.14]	0.33 [0.17]	0.46 [0.09]

Table 1. R^2 score for STT and ST models. TRF stands for temporal RF model, trained at the pixel level. R^2 scores are first averaged over the 5 folds for each possible two time steps spaced by Δt years. Then R^2 scores are averaged again over the possible time steps. We compute the standard deviation over all the possible time steps. We preferred to proceed that way in order to emphasize the score variation for different evolution period. (*e.g.* the evolution over 2000 and 2010 might not give the same R^2 as the evolution over 2008-2018.)

We use a random forest model, denoted as temporal random forest (TRF) model with 100 estimators, as a baseline model. To train it, we fed it the sequence of the average pixel value on each patch in a SITS. It has to estimate the nighttime light evolution, similarly to the STT model.

3.3. Results

In the first column of Table 1, we first notice that both models reach quite equivalent R^2 scores when predicting the NTL value for a given year, as $R^2 = 0.71$ for the STT model and $R^2 = 0.59$ for the ST one. The baseline TRF model reaches lower performances as $R^2 = 0.38$. In [2], the result obtained on the same task is higher, but the dataset and methodology are different. Through different years, the standard deviation for the STT model is 0.04 while it reaches 0.14 for the ST model, indicating that the STT model gives more stable performances.

In Table 1, we observe that for $\Delta t = 1$, *i.e.,* a 1-year evolution, all scores are either negative or with a high standard deviation, meaning that short-term dependencies are difficult to capture in both spatial and temporal domains. But, as Δt increases, the STT model makes better evolution prediction with $R^2 = 0.46$ for $\Delta t = 15$. On the other hand, the ST model still performs poorly as $R^2 < 0$ for all Δt. Individually, both models tend to perform better when Δt increases. The TRF model succeeds to predict evolution as R^2 is significantly positive, so it performs better than the ST model and, as expected, it is less performant than the STT model.

We illustrate these results in Figure 3. We used both models to predict the NTL evolution on Zanzibar Island, which was excluded from the training set. The first row shows the NTL evolution between 2000 and 2020 in the area of Zanz-

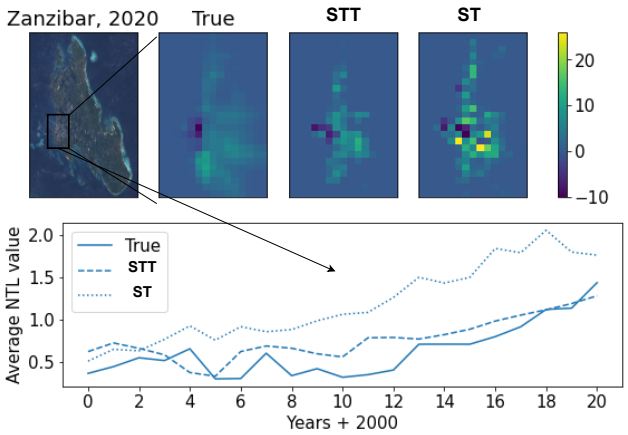

Fig. 3. First row: NTL evolution (ground truth, ST and STT) between 2000 and 2020 on the highlighted area. Second row: yearly average NTL estimation over the 2000-2020 period.

ibar city. We notice that the ST predictions are much more heterogeneous compared to the ground truth, which confirms the results obtained by [3]. We can see that the STT model reduces the heterogeneity over space, and is closer to the ground truth. The second row shows the average NTL evolution from 2000 to 2020 in the highlighted area. The STT model is much closer to the ground truth evolution and shows less variation than the ST model.

4. CONCLUSION AND FUTURE RESEARCH

We set a test-bed to compare spatial and spatio-temporal paradigms. Both models process 32×32 pixels patches of Landsat-7 composites covering an area of 1km^2, but the STT model aims to estimate a NTL sequence according to a patch sequence, while the ST model estimates a NTL scalar value according to a single patch. We observed that both models obtain equivalent performances when predicting a NTL value for a given year, but the STT model gives more reliable evolution predictions. Also, both models are more precise as the time period is long (10 to 15 years). We illustrate this behavior for Zanzibar city and observe that the STT model fits better with the ground truth data while reducing spatial heterogeneity. However, we perform our analysis on one single area and further experiments have to be done to confirm the results at a larger scale.

Acknowledgments - Funded by MPA-Poverty ANR project `https://anr.fr/Project-ANR-19-CE03-0005`

5. REFERENCES

[1] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, August 2016.

[2] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke, "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa," *Nature Communications*, vol. 11, no. 1, pp. 2583, May 2020.

[3] L. Kondmann and X. X. Zhu, "Measuring changes in poverty with deep learning and satellite images," *ICLR, Practical ML for Developing Countries*, p. 6, 2020.

[4] M. Burke, A. Driscoll, D. B. Lobell, and S. Ermon, "Using satellite imagery to understand and promote sustainable development," *Science*, vol. 371, no. 6535, pp. 8628, March 2021.

[5] C. F. Brown, S. P. Brumby, B. Guzder-Williams, T. Birch, S. B. Hyde, J. Mazzariello, W. Czerwinski, V. J. Pasquarella, R. Haertel, S. Ilyushchenko, K. Schwehr, M. Weisse, F. Stolle, C. Hanson, O. Guinan, R. Moore, and A. M. Tait, "Dynamic World, Near real-time global 10 m land use land cover mapping," *Scientific Data*, vol. 9, no. 1, pp. 251, June 2022.

[6] Y. Yuan, L. Lin, Q. Liu, R. Hang, and Z. Zhou, "SITS-Former: A pre-trained spatio-spectral-temporal representation model for Sentinel-2 time series classification," *Intl Journal of Applied Earth Observation and Geoinformation*, vol. 106, pp. 102651, Feb. 2022.

[7] A. M. Noor, V. A. Alegana, P. W. Gething, A. J. Tatem, and R. W. Snow, "Using remotely sensed night-time light as a proxy for poverty in Africa," *Population Health Metrics*, vol. 6, no. 1, pp. 5, October 2008.

[8] V. S. F. Garnot, L. Landrieu, S. Giordano, and N. Chehata, "Satellite image time series classification with pixel-set encoders and temporal self-attention," in *CVPR*, June 2020, pp. 12325–12334.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *ICLR*, 2021.

[10] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *ICCV*, October 2021, pp. 6836–6846.

[11] Z. Chen, B. Yu, C. Yang, Y. Zhou, S. Yao, X. Qian, C. Wang, B. Wu, and J. Wu, "An extended time series (2000–2018) of global NPP-VIIRS-like nighttime light data from a cross-sensor calibration," *Earth System Science Data*, vol. 13, no. 3, pp. 889–906, March 2021.