Human-in-the-loop solution for scoring economic development from geospatial data

Anonymous Author(s)

Abstract

Reliable and timely measurements of economic activities are fundamental for understanding economic development and for delivering humanitarian aid and disaster relief to where needed. However, many developing countries still lack reliable data. This paper introduces a novel approach for measuring economic development from high-resolution satellite images in the absence of ground truth statistics. Our method's novelty is that we break down a computationally hard problem into sub-tasks, which involves a human-in-the-loop solution. With the combination of unsupervised learning and the partial orders of dozens of urban versus rural clusters, our method can estimate the economic development scores of over 10,000 satellite grids with less human labor than other baseline approaches (Spearman correlation of 0.851). We demonstrate how to apply our method to both developed and developing economies.

1 Introduction

Collecting data on socioeconomic activities is crucial for designing sound government policies to promote and sustain economic development. Up-to-date information about economic activities at fine-grained levels can also be critical for the efficient provision of humanitarian and disaster relief in the right place at the right time. Not surprisingly, most developed countries and international organizations, such as the World Bank, deploy significant amounts of financial and human resources to conduct censuses and surveys. For instance, there are more than 130 ongoing surveys in the U.S. on demographics and economic activities currently conducted by the United States Census Bureau.

Unlike rich countries, however, developing countries often lack resources for conducting expensive surveys and suffer from data reliability issues. To solve these limitations and improve the quality of economic measures, researchers from various fields have attempted to use alternative data sources. Remote sensing data can be an attractive source for economic measurements because of its extensive geographic coverage, timeliness, and high granularity. For example, high-resolution satellite images allow social scientists to analyze the local impact of air pollution, changes in land use or crop choice, and fluctuations in retail demand predicted by counting cars [2].

Despite the rapid development of deep learning techniques in analyzing satellite imagery to construct economic development proxies, there remain two considerable limitations to use them in practice. First, existing classification or regression models cannot be applied to regions without ground truth data, where these techniques are most needed, because these models require a massive amount of labels for training. Second, deep learning remains as a black-box function. Its lack of interpretability hinders practitioners from applying such techniques to real-world problems.

This paper proposes a novel approach to overcome such limitations. We introduce a model that learns from high-resolution satellite images to *rank relative scores* of economic development, for which we use urbanization as a comprehensive proxy, following the economics literature [10]. Our deep neural network model clusters images based on visual features and defines paired sets of clusters, i.e., a partial order graph (POG). The POG is an essential element in our approach addressing the limitations

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

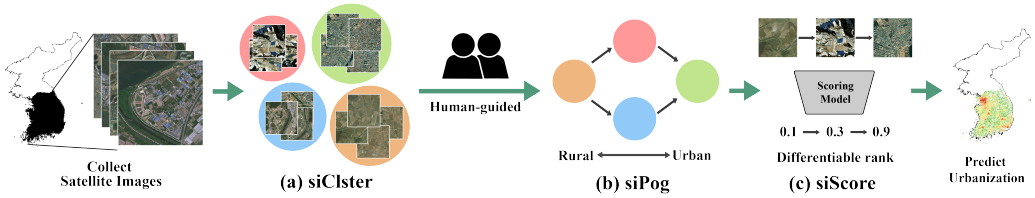

Figure 1: The overall architecture of the proposed model, composed of (a) siCluster for clustering satellite images, (b) siPog for generating partial order graph (POG), and (c) siScore for training the scoring model with POG.

of the existing methods: it is an interpretable input to the scoring model that estimates the economic development, which can be generated either by readily available data or light human annotation. Human-guided POG leads to a human-in-the-loop solution, whose generation process is simple but makes it possible to see which human activity patterns in satellite imagery depict a more advanced economic development level. Our proposed computational framework to measure sub-district level economic development from satellite imagery without the guide of any partial ground-truth data is novel and shows remarkable performance gain over existing baselines. These geo-sensing techniques have broad applications for continuous monitoring of economic activities.

2 Method

2.1 Overview

Problem definition: Assume that $\mathcal{I} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is satellite images set for a specific area and economic ground-truth value y_i of each image $\mathbf{x}_i$ is unknown. The main purpose of the proposed method is to train model f in order to compute a score $\hat{y}_i$ for each image $\mathbf{x}_i$ (i.e., $\hat{y}_i = f(\mathbf{x}_i)$) that is aligned well with the economic ground truth y_i. In this process, a label that depends on humans is indispensable because we do not have any ground truth value. However, the existing method, like pairwise [6], is human-intensive. To mitigate this problem, we propose a novel human in the loop approach to minimize human labor. The method consists of three stages: (1) identifying satellite grid clusters of similar urbanization traits via a deep learning architecture — siCluster, (2) generating the human in the loop POG of the found clusters — siPog, and (3) computing differentiable scores from the learned POG structure — siScore. The intermediate POG structure helps solve the intensive human problem efficiently and also makes the result interpretable. These steps are visualized in Fig. 1.

2.2 siCluster - Clustering satellite imagery

To generate scores that represent the urbanization development, one needs to know what kinds of human activities capture such values. Satellite images are a promising source of the sophisticated prediction because each Satellite cell contains detailed landcover information, including vegetation and built-up. To recognize diverse human activities, we apply the deep learning-based clustering algorithm to efficiently reduce the computational cost via hierarchical architectures [5, 8]. This work employs DeepCluster [1], which is a novel end-to-end training process in an unsupervised manner.

Primitive DeepCluster has two weaknesses. One is the initial state randomness; the model depends on the initial weights to propagate through the training process. Another is the lack of consistency in the class assignment; the model relies on the pseudo-labels generated from its k-means clustering, which is not robust to noise in data. As a result, DeepCluster is not directly applicable to our problem. In a nutshell, our clustering algorithm builds upon DeepCluster with two new improvements.

Improvement #1 from transfer learning: To ensure an organized initial state for the encoder, we adopt transfer learning in a weak supervision fashion. Given a small dataset with proxy labels that include one thousand satellite images with three labels: urban, rural, and uninhabited, we then introduce a semi-supervised learning technique to leverage massive amounts of unlabeled data

effectively. The Mean Teacher [9] model, which uses a temporal ensemble by minimizing the inconsistent predictions between the teacher and student, is used.

Improvement #2 from consistency preserving: We added new loss terms to prevent the model from learning trivial features. Suppose a given satellite grid $\mathbf{x}_i$ and its corresponding encoded vector $\mathbf{v}_i$, i.e., $\mathbf{v}_i = h_W(\mathbf{x}_i)$. We then augment $\mathbf{x}_i$ via common techniques such as rotation, gray-scale, and flipping that do not deform the original visual context. Let us call the augmented versions $\hat{\mathbf{x}}_i$. Then, the distance between $\mathbf{x}_i$ and its augmentations $\hat{\mathbf{x}}_i$ in the embedding space should be close enough, compared with the distance between $\mathbf{x}_i$ to other data points. We define the *consistency preserving loss* to represent this invariant feature characteristic against data augmentation in the embedding space. siCluster is trained by jointly optimizing the negative log-likelihood loss and reducing the Euclidean distance between the input and its augmentations on embedding space (Eq. 1).

$$L_{ecp} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} || \, h_W(\mathbf{x}_i) - h_W(\hat{\mathbf{x}}_i) \, ||_2 \tag{1}$$

2.3 siPog - Generating partial order graph

Images within each cluster share similar visual contexts that likely represent a similar level of economic development. The second step of the algorithm aims at ordering these identified clusters. The partial order graph (POG) is an efficient representation showing the order across different clusters while ignoring any within-cluster difference. We generated a POG in the order of economic development. Here, development refers to an economic transition from agriculture to manufacturing and service industries, which tend to cluster in more urbanized areas [10]. When two clusters showed a similar development level, they were placed at the same level without any strict ordering between them, as illustrated in Fig. 1(b). Below are two strategies of siPog under different conditions.

Human-guided.
We first considered the human-in-the-loop design and asked human annotators to sort clusters manually. Both experts and laypeople participated in this ordering task. Annotators compared clusters and identified relative orders of clusters by examining the provided grid images. Clusters were ordered and connected as a graph by their presumed economic development level. Cluster pairs whose development levels were judged to be indifferent were placed at the same level within the POG. The strength of this method is its lower cost than the full comparison of images.

Data-guided.
POG can also be generated without human guidance. While grid-level demographics are costly to obtain, ample resources can be used as a proxy, such as Internet search results. Proxy data are aggregated at a high-level (e.g., city or province) or are not accurate. Below we demonstrate one example, *nightlight luminosity*. Nightlight luminosity is the light intensity measured in nighttime satellite imagery. This publicly free data is only available at low resolution. We first extrapolate the nighttime images to match the size of the daytime satellite grids. Once we identify all the extrapolated nighttime grids corresponding to each cluster, nightlight intensity was averaged. We perform a two-sample t-test with a threshold 0.01 to detect any significant intensity difference between every two clusters and consequently create an edge between them if two clusters are sufficiently comparable.

2.4 siScore - Computing scores using POG

Given the relative orders of clusters in the POG, the next task is to assign a score between 0 and 1 to every cluster using the CNN-based scoring model f. The model automatically detects which satellite imagery (belonging to clusters) determines the urbanization score via supervised learning. This scoring is identical to the metric learning framework, as the model determines semantic scores that can measure urbanization. We adopt the list-wise metric learning method with our unique structures for the scoring model, siScore. We limit the range of values of the scoring model from 0 to 1 by clamping smaller or larger values during training.

List-wise metric learning: The only knowledge from POG is the orders of clusters, but not the orders of individuals images. The third step, siScore, use the POG structure in learning scores of every satellite grid in the following way. We first extract every ordered path from the POG. After choosing one path from them, an equal batch size of n_s images are sampled from each cluster C_k

along the selected path. Since the selected path is already ordered, the actual rank of clusters in the path is accordingly aligned. The scoring model f trains to match two ranks: generated rank from the model score and actual rank based on POG. The Spearman correlation evaluates how well given two variables are related as a monotonic function in terms of measuring rank correlation. Accordingly, we calculate and maximize the estimated ranks' Spearman correlation against the actual rank to train the model.

However, the Spearman correlation is not differentiable and thereby unsuitable due to the back-propagation nature of deep learning. Based on recent advances in computing ranking losses, we use a simple approximation method suggested in [3] to mimic the sorting algorithm and make the algorithm differentiable to use the Spearman correlation as a loss directly.

Variance regularization: Finally, we added a loss to regularize the score distribution of each cluster to satisfy small score variance within each cluster. With small variances in score distributions, the overlapping part between two adjacent score distributions, where the flipped results are brought, would be reduced. The average variance of score distributions of every cluster in the selected path P_j is minimized as a regularization loss:

$$L_{var} = \frac{1}{|P_j|} \sum_{C_i \in P_j} Var(f(\mathbf{X}_{C_i})), \tag{2}$$

where $\mathbf{X}_{C_i}$ indicates the batch images in cluster C_i and Var denotes the function that calculates the variance of the given score list.

Finally, loss for maximizing Spearman correlation (L_s) and loss for variance regularization (L_{var}) are concurrently optimized to train siScore with the weight parameter α as in Eq. 3.

$$L_{score} = L_s + \alpha \times L_{var} \tag{3}$$

3 Experiments

3.1 Datasets

Satellite imagery: ESRI, a GIS software company, provides satellite-based remote sensing data at various zoom levels. The zoom level (Z) refers to the spatial resolution. We chose the $Z = 15$ images with 4.7m-resolution, which can distinguish individual buildings and other artifacts such as roads. Each tile contains three spectral bands (RGB), and the images are cloud-free for most of the area. We consider data from three countries: South Korea, Malawi, and Vietnam, where images are between 2015 and 2017. Nighttime luminosity is available for public use from the NASA Earth Observing System Data and Information System with the best resolution at $Z = 9$ due to its blurring effect.

Evaluation dataset: the grid-level ground truth labels were prepared in two forms: **Population** and **Gross Floor Area**. Facebook has contributed to making the most precise population map of the world, and they open the data that covers most of the Asian and African countries online [4]. The estimation is at the resolution of an arcsecond-by-arcsecond scale. Since this unit area is smaller than our grid size, we sum up each unit's estimated population density located inside the grid for evaluation. The Gross Floor Area, which is the total amount of floor space or construction in buildings, can refer to the degree of economic development in a given area. Such information can be computed from GIS data and detailed information on building shapefiles. The local government can collect such information; we utilize the official data released in South Korea.

3.2 Performance evaluation

We evaluated our model performance on the South Korea dataset. In South Korea, the cluster's optimal number was found to be 21 based on grid search. The POG with discovered clusters was generated as follows. **Human-guided** method involved annotations from five experts and five locals, where both the average and the maximum performance are reported. For data-guided POG, we utilized grid-level nightlight intensity data (**Nightlight-guided**). Table 1 reports the correlation values between the estimated economic development and two kinds of ground truth labels: Gross Floor Area and Population. Both Spearman and Pearson correlations were calculated on the log-scaled ground truth values. All models produced solid correlation (i.e., above 0.7) with ground truth labels,

Table 1: Comparison of prediction models for South Korea. Results are evaluated for two grid-level statistics: Gross Floor Area and Population Count.

Method		Gross Floor Area		Population	
		Spearman	Pearson	Spearman	Pearson
A	Human-guided (Avg)	0.825	0.787	0.764	0.766
	Human-guided (Max)	**0.851**	0.800	**0.795**	0.778
	Nightlight-guided	0.846	0.801	0.794	0.789
B	Nightlight-only	0.664	0.655	0.728	0.731
	Pairwise (Human)	0.651	0.610	0.300	0.302
C	K-means	0.434	0.587	0.451	0.557
	DeepCluster	0.618	0.559	0.532	0.551
D	Triplet (POG)	0.807	0.754	0.768	0.726
	Pairwise (POG)	0.825	0.759	0.767	0.739
	w/o Score model	0.737	0.675	0.678	0.673

A : Our model, B : Baselines, C : siCluster ablation, D : siScore ablation

even when such information was not available during training. The best performance comes from the human-guided model, reaching 0.851 and 0.795 in Spearman correlations.

Seven baselines were implemented for comparison, which used ResNet-18 as the backbone network. **(1) Nightlight-only** uses the nightlight intensity for measuring economic development. We also experimented with human-annotated labels **(2) Pairwise (Human)** that indicated the relative rank of four thousand random image pairs. Three annotators with domain knowledge of target countries were asked to choose which image in pair showed higher economic development, and their decisions were aggregated. The pairwise loss was then used for training. This model is one kind of naive methods that directly learn from the human-annotated orderings. These baselines were not as effective as our models.

The next two baselines were ablations for siCluster. We replaced this module by the conventional **(3) K-means** clustering algorithm and the original **(4) DeepCluster** algorithm. The remaining three were ablations for siScore. The labels for **(5) Triplet (POG)** and **(6) Pairwise (POG)** were generated by an identical POG instead of human annotation. When generating labels from the POG, cluster pairs were randomly selected, and images from each cluster formed pairs. The order of chosen clusters was considered as a label for these pairs. Triplet labels included anchor, positive, and negative samples. The model was trained to generate a similar score between anchor and positive while producing different and order-preserving scores between anchors and negatives. We also proposed baseline **(7) without the score model**, which gives a scalar value that preserves POG's orders to each cluster instead of a deep learning-based score model. Nightlight-guided POG is used for these baselines.

Ablation studies demonstrate that every tested component of the model is critical to performance. Mostly, human-guided approach performance does not lag behind a data-driven approach. Our SiPog approach outperforms than a pairwise strategy, which is labor-intensive.

3.3 Application to developing economies

We conducted additional experiments on two developing economies, Malawi and Vietnam, with a total of 64,303 and 226,305 satellite images for each country, respectively. All models were trained in the same manner as mentioned earlier, except for the cluster count n_t in siCluster. The optimal n_t was found to be 7 for Malawi and 11 for Vietnam based on grid search, respectively.

Figure 2 compares the performance of the models, evaluated by the grid-level Facebook population data [4]. Our model repeatedly outperforms the conventional nightlight model for developing economies. In the case of Malawi (i.e., the poorer of the two), our models improve the correlation to the Facebook data from 'weak' to 'strong.' Our model's advantage is attributed to the use of daytime imagery, which overcomes the light saturation effect in nighttime satellite images. Moreover, nighttime satellite imagery is known to be erroneous for extreme poverty areas since light intensity is very low and varies little in these areas [7].

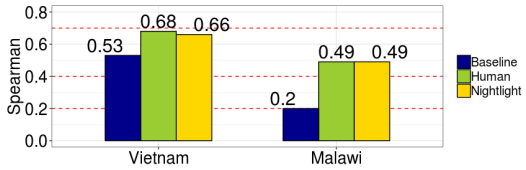

Figure 2: Results for Malawi and Vietnam. Two models (Human and Nightlight) are compared against the conventional nightlight model (Baseline). The red lines indicate the boundaries for 'weak,' 'moderate,' 'strong,' and 'very strong' correlations from the bottom.

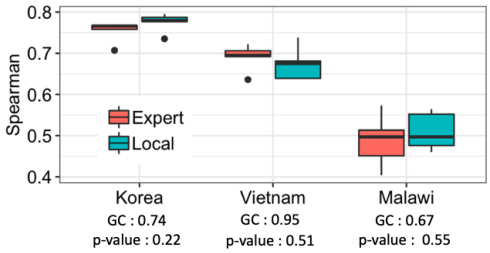

Figure 3: Spearman correlation of POGs generated by expert versus local annotators. GC refers to the graph consistency. The p-value of the t-test indicates there is no significant difference between experts and locals.

3.4 Robust test for human in loop approach

POG robustness: To validate the intermediate POG designs generated by human-guided and data-guided siPog approaches, Table 2 reports that Spearman and Pearson correlation for the generated POGs and the actual orders of images according to the population data. We assigned the rank to every image following the POG, and images in the same cluster are considered to have the same rank. As a result, shown in Table 2, Both designs' correlations (Human and Nightlight data-guided POG) are near 0.7, and this high correlation infers that our POGs with three models well mimic the actual order of images.

Table 2: Measured consistency between generated POG from our proposed designs and ground-truth population dataset. Spearman and Pearson correlations are used for evaluation.

	Human-guided	Nightlight-guided
Spearman	0.68	0.71
Pearson	0.65	0.70

Sensitivity of human annotator type: Does constructing a POG require advanced knowledge? We examined whether the expertise of annotators has any effect on the quality of POG. Our expert annotators were knowledgeable in economics and geography, whereas the local annotators had an in-depth knowledge of the target country. For different countries, we recruited different local annotators. All annotators were given the same instruction and data, and their POGs are compared in Figure 3. This small scale comparison shows there is no evidence of the quality difference in the generated POGs between experts and locals, demonstrated by the p-values of the two-sample t-tests. Still, the views of local annotators are very important for countries with few modern buildings. In Malawi's case, we observed that the locals' POG performs slightly better spearman correlation with lower variance. We verified that locals have a better understanding of the target country's cultural background from the personal interview. For example, most economics experts did not recognize the dot-like objects in the rural side as residential buildings. However, they are small dwellings which are widely spread in Malawi. Furthermore, high graph consistency between the two groups, which is calculated by measuring the Spearman correlation of two partial orders, shows their POGs are similar (i.e., the value of graph consistency is near or over 0.7). These findings imply that our POG generation process is simple, and thereby anyone who wants to train the model can quickly implement by themselves.

4 Conclusion

We presented a new method that measures economic development from high-resolution satellite imagery using deep learning without any labeled data. This idea of utilizing the high representation power of deep learning in the demographic analysis could be useful for many existing humans in loop projects. The model applies to developing economies where labeled data is limited and developed economies where grid-level census data are not gathered frequently. Generating a human-aided

partial order graph helps interpret what kinds of land cover patterns are linked to high urbanization scores. Our approach can be used in various scenarios, including change detection under the outbreak of economic crisis or natural disasters; repeatedly computing the economic development scores over newly collected satellite imagery can measure changes in fast-moving areas over time.

References

[1] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *proc. of the ECCV*, pages 132–149, 2018.

[2] Dave Donaldson and Adam Storeygard. The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4):171–98, 2016.

[3] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Sodeep: a sorting deep net to learn ranking loss surrogates. In *proc. of the IEEE CVPR*, pages 10792–10801, 2019.

[4] Facebook. *Data for Good program*, 2020. Available at `https://data.humdata.org/organization/facebook`. Date accessed 29 Jan 2020.

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[6] Eyke Hüllermeier, Johannes Fürnkranz, Weiwei Cheng, and Klaus Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16-17):1897–1916, 2008.

[7] 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, 2016.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *proc. of the NIPS*, pages 1097–1105, 2012.

[9] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *proc. of the NIPS*, 2017.

[10] Henderson J. Vernon. Urbanization and economic development. *Annals of Economics and Finance*, 4(2):275–341, 2003.