Measuring economic development in North Korea: Human-machine collaboration using satellite imagery

Version as of May 27, 2021

North Korea has long been a black box with no official data for outsiders to assess its economic development. The Lack of ground truth labels poses a profound challenge in applying inference models based on the country's few resources, such as satellite imagery for economic measurement. To overcome these constraints, we develop a human-machine collaborative algorithm that incorporates lightweight human annotations on economic development in the machine-learning process. We present two main findings from the grid-level estimates of the economic development in North Korea by applying our model on daytime satellite images for the period from 2016 to 2019. First, the evaluation confirms that our human-machine collaborative algorithm outperforms machine-only learning approaches based on nightlight intensity or land cover classification. Second, the use of our measure as a proxy of economic development indicates that amid rising pressure from economic sanctions the centrally planned economy has been directing more resources towards its capital and regions with highly publicized state-led development projects. Our model can be applied to other developing countries with insufficient data and provide a reliable and inexpensive indicators of economic development on a granular level.

Introduction

Reliable data on economic activity remain insufficient in developing countries, limiting research and policy analysis on them (*1*). North Korea is an extreme case; its economy has been under a veil of secrecy for several decades. Little is known about its conditions after a series of economic sanctions since 2017 despite the great interest of the international community. The last official statistics were produced by the United Nations in 2008, covering the county-level population (*2*). While organizations such as UNICEF have been allowed to conduct province-level surveys in the country (*3*), no statistics at the sub-provincial level have been released. To overcome the data deficiency, researchers have compiled data by other means, including conducting interviews with North Korean defectors (*4*), examining news articles published by the North Korean news outlets (*5*), and utilizing luminosity data from nightlight satellite imagery (*6, 7*).

Meanwhile, recent computer models have shown surprisingly high predictive power in analyzing satellite imagery and explaining socioeconomic statuses such as consumption and assets in Sub-Saharan Africa (*8, 9*) and Southeast Asia (*10, 11*). Predictions have become more elaborate when combined with alternative data sources, such as the geotagged information on Wikipedia (*12*) and the audience estimates on Facebook's advertisement platform (*13*). However, ground-truth data still remain the backbone of these extant machine learning approaches; computer models must be "supervised" by considerable quantities of ground-truth data labels corresponding to each observed region. Ironically, countries that can benefit the most from remote sensing technology tend to lack reliable ground-truth labels (*1*).

We design a human-machine collaborative algorithm that converts satellite images to a grid-level proxy of economic development. Unlike previous approaches, our data generating process does not require millions of ground-truth data labels corresponding to each satellite image. Therefore, we can apply computer vision techniques even to the world's most unveiled country, North Korea. The algorithm comprises three stages: a machine first groups similar-looking images into a manageable size of clusters; human annotators then rank the clusters based on their perceived degree of economic development;finally, the algorithm computes each image's score based on the image features and on human annotations in the absence of ground-truth labels.

We present two main findings from analyzing publicly available satellite images of North Korea for the period from 2016 to 2019. First, the multi-faceted evaluation shows that the human-machine collaborative algorithm can outperform machine-only learning approaches based on the nightlight intensity or land cover classification data. Moreover, our model's performance is comparable to what has been achieved by state-of-the-art supervised learning models that utilize an extensive set of labels and have access to high-resolution satellite imagery (*8, 14*). Second, a quantitative analysis of temporal changes using our measure as a proxy of economic development, a quantitative analysis on temporal changes suggests that North Korea has been directing more resources towards its capital city and urban areas, as well as regions with highly publicized state-led development projects. Economic development in the vast majority of the county, where most of the population resides, appears stagnant. This prediction helps us better understand the unseen regional development in North Korea under the heavy economic sanctions experienced since 2017.

Method

This paper utilizes a human-machine collaborative model that learns visual features from satellite images to rank the relative scores of economic development without the use of conventional labeled data (see Fig. 1). The inputs are 256×256 pixel-sized Sentinel-2 satellite images taken at 10 m/pixel resolution, which is the finest resolution available for North Korea among public resources.[1] Satellite imagery at this resolution can indicate structures of buildings, roads, and other human artifacts, such as crop fields. The proposed model relies on multiple deep learning-based computer vision techniques to identify image clusters with similar visual features. The algorithm lacks interpretation, and human knowledge guides the ranking process by reading patterns of civil artifacts from satellite images in a light fashion. We define the ranked sets of clusters as a partial order graph (POG). A POG contains essential information on the relative ranking of clusterwise economic development. The machine is trained to compute grid-level prediction scores that tally with the order of satellite images in the POG.

While a POG can be generated either by readily available data (i.e., data-guided POG or machine-only learning algorithm[2]) or by humans (i.e., human-guided POG or human-machine collaborative algorithm), the human-guided approach is significant as it does not require expensive ground truth labels. This approach makes our model applicable to predicting the economy

[1] The grid size is determined by the satellites' resolution, in units of a zoom level (Z) of 0-20. We use a 10 m resolution, equivalent to a zoom level of 14 or 9.547 meters per pixel. Each image covers a 2.5×2.5 km^2 area.

[2] One can extrapolate the nationwide raster data (e.g., nightlight, land cover classification) to match the size of the satellite grids. Once all grids that correspond to each cluster are identified, the average value (e.g., nightlight intensity, urbanization ratio) for each cluster can be used to rank clusters in the POG.

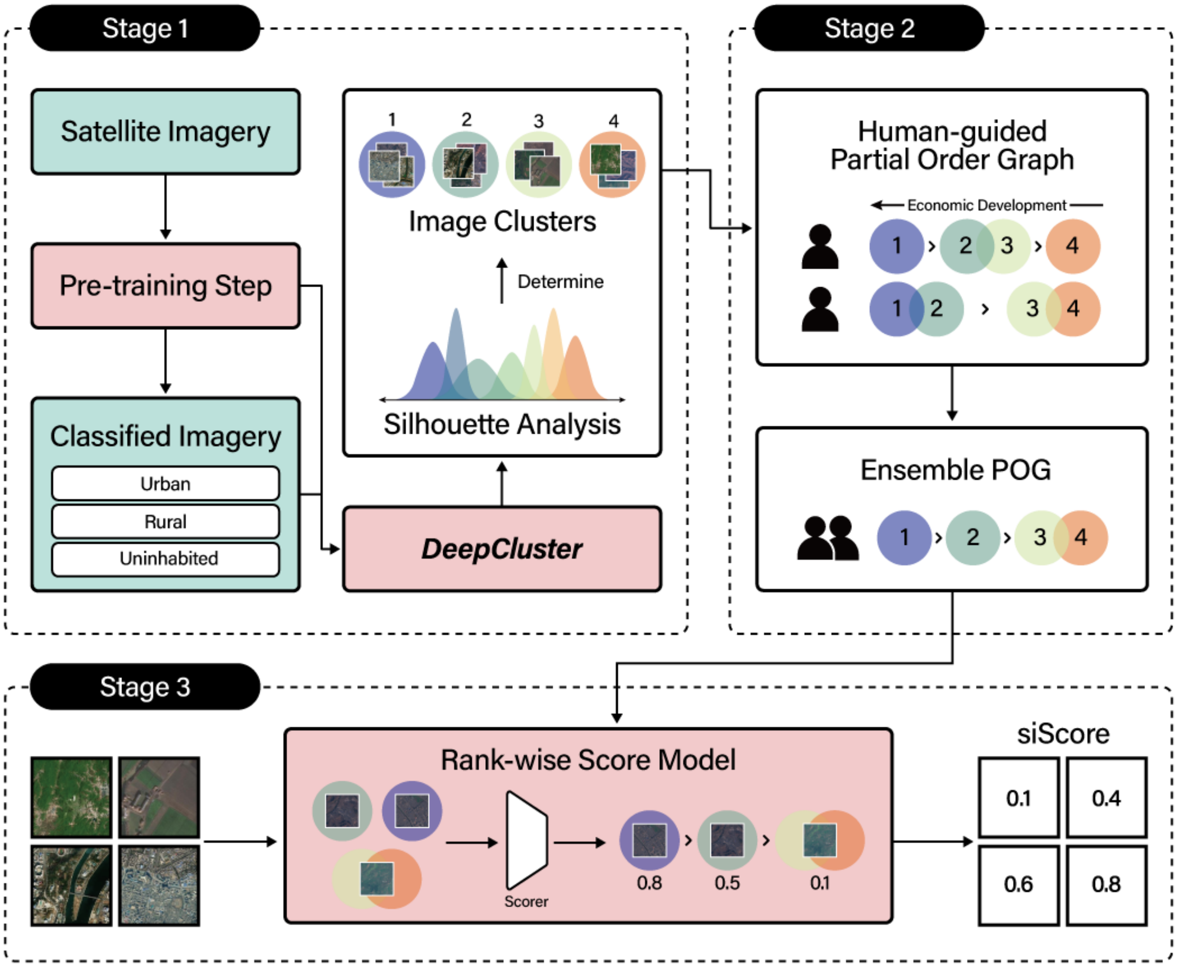


Figure 1: Illustration of the proposed model, which is composed of (1) machine-driven clustering of satellite images, (2) human guidance on the *partial order graph (POG)* of image clusters, and (3) a machine-driven rankwise scoring model. A POG contains information on the relative ranking of each cluster's development, perceived and judged by each participating human. Knowledge from multiple POGs is summarized into a single representative POG by means of an ensemble process and is then used by the scoring model.

of unknown regions where labels are restricted or scarce. Below, we describe the three stages of the new human-machine collaborative algorithm.

Stage 1: Clustering satellite images. Deep learning-based clustering can discriminate distinctive visual patterns and group images with similar traits. This work employs DeepCluster (*15*), an unsupervised deep learning algorithm, to analyze satellite images. Before the clustering process, we first separate uninhabited regions via a pretraining step and regard them as a single cluster. This adjustment enables the clustering algorithm to concentrate on the images that need critical comparison, thereby improving the computational efficiency. The uninhabited areas constitute a large proportion of terrain (e.g., an estimated 75% of North Korean territory is uninhabited). Then, to determine the optimal number of clusters, we apply silhouette analysis over the clusters' embedding space to measure how similar the instances are within a cluster. In the case of North Korea, the final data contain 23 clusters, including the uninhabited cluster.

Stage 2: Sorting clusters by human guides. Humans were used to order the 23 image clusters by subjectively evaluating the satellite images assigned to each cluster. The outcome of this human-guided process is the relative rank of clusters, called the POG. Note that a POG represents the ordinal relationship among clusters, where each cluster comprises visually indifferent images. We hired three types of human annotators to generate POGs for North Korea's economic development—economists, geographic information system (GIS) experts, and North Korean defectors. The main advantage of ranking clusters, not individual satellite images, is consideration of the geocultural knowledge of humans. Scaling down the ranking task size to nearly two dozen clusters made the task feasible for human participants[3]. The POGs contributed by each human can be summarized into a single representative POG by means of an ensemble rank process (see Supplementary Materials for details).

Stage 3: Training a rank-wise score model. The final stage is designing and training a convolutional neural network (CNN) to assign a numeric score between 0.0 and 1.0 for every satellite grid image, which we call *siScore*. The scores assigned to tens of thousands of satellite grid images should align with the human-guided POG obtained from the previous step. This mapping is nontrivial since POG clusters and grids differ in size by several orders of magnitude (e.g., 23 clusters versus 32,578 grid images for North Korea). The objective of preserving the POG ordering is as follows. For every ordered path that connects the least and the most developed cluster pair in the POG, we train a ranker f that assigns scores to images in the corresponding clusters such that their ranks are as similar as their cluster ranks in the POG. This objective is identical to maximizing the Spearman correlation between the model's scores and the order in the POG path. The problem, however, is challenging to solve with existing deep learning models

[3] All participants completed the ranking task within 1-2 hours.

since ranks are nondifferentiable. We use a heuristic algorithm to approximate a differentiable ranking function described in detail in the Supplementary Material.

Results

We applied our human-machine collaborative model to North Korea and generated a spectrum of scores uncovering the country's regional development at the grid-level. The grid-level map (Fig. 2 A) shows the average scores over four years between 2016 and 2019. This period is of particular interest to researchers and policymakers because North Korea was subject to a series of economic sanctions to suppress its capacity for nuclear and missile programs. The map depicts several distinctive development patterns; the western plains and eastern coastal port areas appear developed, whereas the vast central and northern regions with high altitude mountains have low development scores. The model also provides a richer picture than the nightlight-based image shown in Fig. 2(B). This difference is more evident in the zoom-in view of, for example, Sepho County in Fig. 2(D) and (E), where the model predictions capture more refined variations in economic development across its urban, rural, and mountainous areas. Notably, our model results are comparable to the maps of land cover classification or building footprints constructed by the South Korean government on a decennial basis, as shown in Fig. 2(C and F). However, unlike our method, generating land cover classification and building footprints requires many resource, such as proprietary satellite and aerial images and extensive human inspection.

Fig. 3 compares the performance of our human-machine collaborative model with that of data-guided models based on nightlight and land cover datasets. The evaluation presented uses two manually constructed partial datasets of North Korea as the ground truth. The first is the building footprint dataset, which covers 70% of the country and contains each building's outline marked by GIS experts in 2014 (Fig. 2(F)). We calculate the total floor area of buildings and then use it to compute the built-up ratio as a proxy for economic development.[4] The first and second rows of Fig. 3 present the evaluation results at the grid and district levels, respectively. Our second evaluation measure used the district-level population density from the country's most recent population census in 2008. The evaluation results are shown in the bottom row of Fig. 3.

[4] The building footprint dataset does not include height information, which restricts us to compute the floor area instead of the gross floor area. Moreover, the dataset only covers 69.4% of North Korea, hence leaving the remaining 30.6% of all regions (or 9,959 grids) incomparable. This limited coverage is a critical caveat of using ground truth data based on human labor. In contrast, our model is capable of a full-scale prediction of for *all* regions.

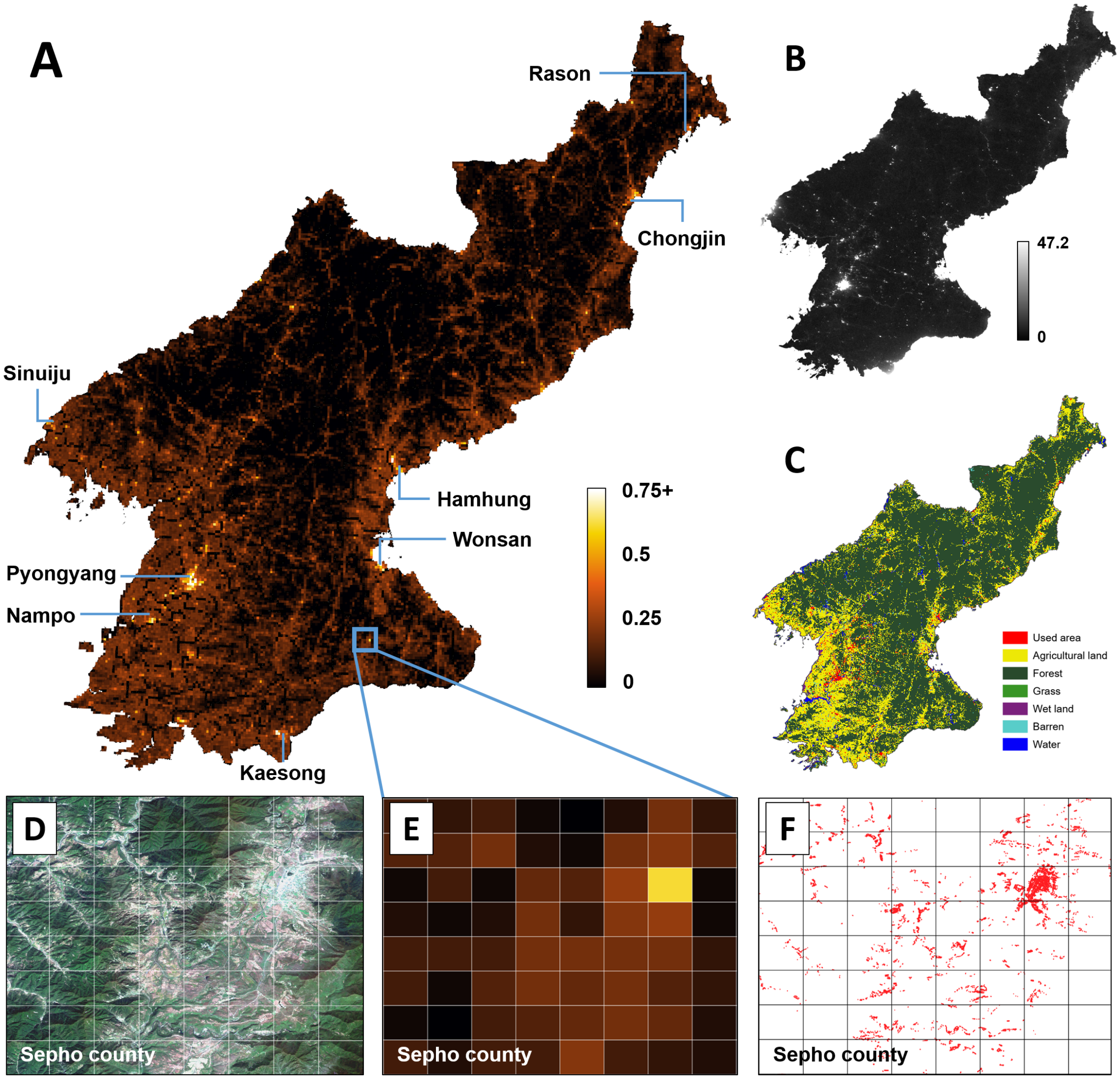


Figure 2: Visualization of economic development levels predicted by our human-machine collaboration model. (A) Prediction scores over grid images averaged over four years from 2016 to 2019, (B) shows the yearly aggregated VIIRS nightlight data in 2019, and (C) shows the land cover classification map released by the South Korean Government in 2019. The zoomed-in views in (D–F) compare predictions for Sepho County in the Kangwon region. From left to right are the Sentinel-2 satellite images taken in 2019, model predictions and manually verified buildings (colored red) from the building footprint data in 2014.

(Supplementary Materials contain more evaluation results comparing manually identified market locations and industry listings of North Korea.)

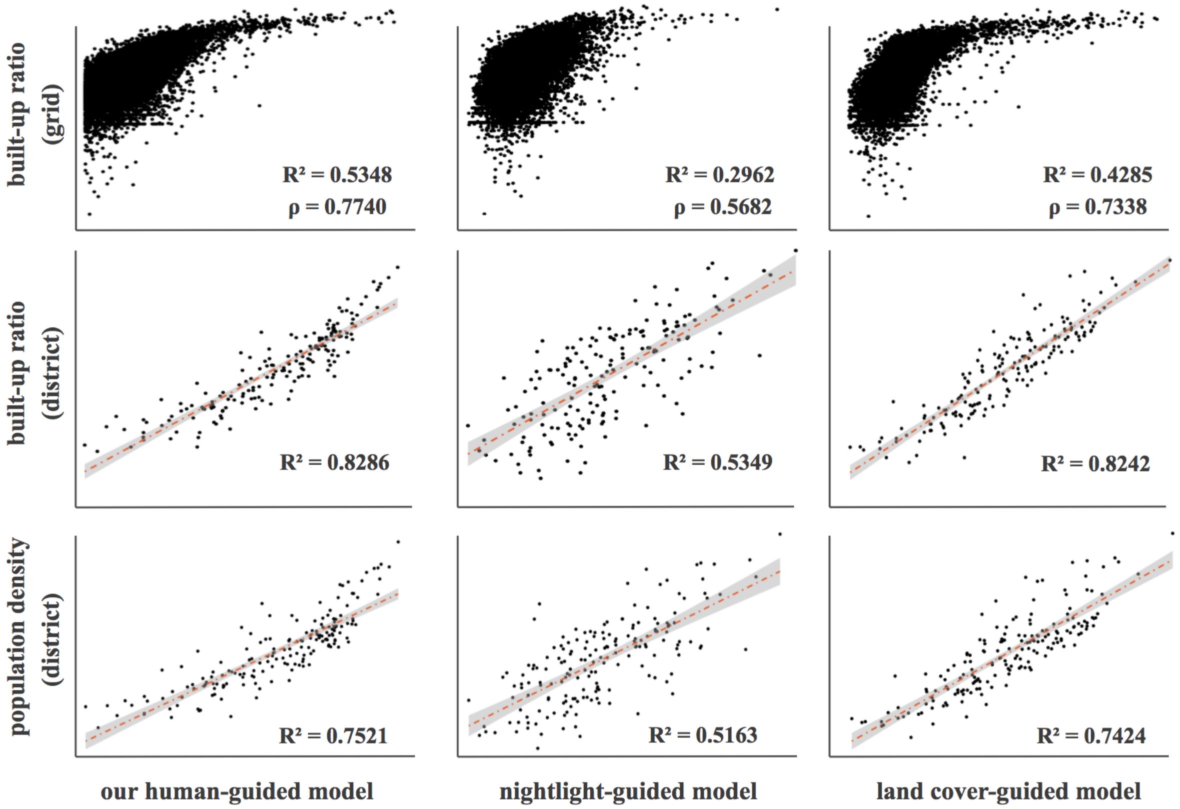


Figure 3: : Comparison of the model performance for the human-guided POG and two data-guided POGs against partial ground-truth datasets of the built-up ratio from 2014 and the population density from 2008. We use the nightlight and land cover datasets to construct data. Tguided POGs. Evaluations based on the built-up ratio in the first two rows were performed for available areas (70% of the country) in the building footprint dataset from 2014. Null values or areas missing from the comparison dataset were excluded.

Our human-machine collaborative algorithm generates scores highly predictive of these proxies of economic development both at the grid-level (R^2=0.53, Pearson's ρ=0.73, Spearman's ρ=0.77)[5] and at the district-level (R^2=0.75–0.83). The human-guided model outperforms the nightlight-guided model (grid: Spearman's ρ=0.57, district: R^2=0.52–0.53) and shows comparable performance to the data-rich land cover-guided model (grid: Spearman's ρ=0.73, district: R^2=0.74–0.82). The ability to generate high-quality predictions in the absence of ground truth labels from publicly available satellite imagery lends our model to challenging remote sensing

[5] Note that the Spearman's rank correlation coefficient is more appropriate to evaluate our model as the model is targeted to predict rankwised scores.

tasks, as presented in the current study. Existing state-of-the-art supervised learning models based on a rich set of labels and high-resolution satellite images have reported Pearson correlations of 0.64 (*8*) and 0.74 (*14*).

Figure 4 shows satellite images taken in 2016 and 2019, along with model predictions represented as a heatmap, for two locations: a construction site in the tourist areas in Kalma district (top images) and an industrial development area in Wiwon county (bottom images). The top pictures reveal more vivid changes due to new buildings and roads compared to the bottom pictures. The changes in scores in the heatmap reflect the difference; any increment (or decrement) in the model score represents an increase (or decrease) in the relative economic development seen in the picture.

Having assessed the model, we next use the yearly predictions at the grid level to examine changes in regional development from 2016 to 2019. In the context of a planned economy, we conjecture that development differences across regions may indicate the central government's political and economic strategy to cope with pressure from sanctions. One of the long-standing questions regarding sanctions against North Korea has been on understanding what measures the regime takes and how it distributes resources for survival and stability. For example, what changes have sanctions brought to the capital Pyongyang and North Korea's state-led development projects (*16, 17*)?

We conduct a quantitative analysis by examining changes in *siScore* throughout North Korea. Similar to our visual exercise, we compare development scores in 2016 (before sanctions) and 2019 (after sanctions). We employ a simple regression framework to analyze the relationship between region-specific features and economic development, predicted by our model. As an alternative measure of economic development, we also use nightlight intensity, which is commonly utilized in economics and social sciences (*18*). We focus on four major features that are considered to be crucial to the regime as potential determinants of regional economic development: proximity to economic and political hubs, whether an area is designated as an economic development zone (EDZ, also known as *Gyeongje-gaebalgu*), the number of major mineral mining sites, and containment of nuclear-related facilities. Proximity is measured as the Euclidean distance from a grid's center to the center of each hub. For EDZ features, we assign a value of one to grids that are located inside the zones and a value of zero otherwise. For other site-specific features, such as mines and nuclear test sites, we assign a value of one to grids that contain such sites or that are immediately adjacent to a grid containing them. To

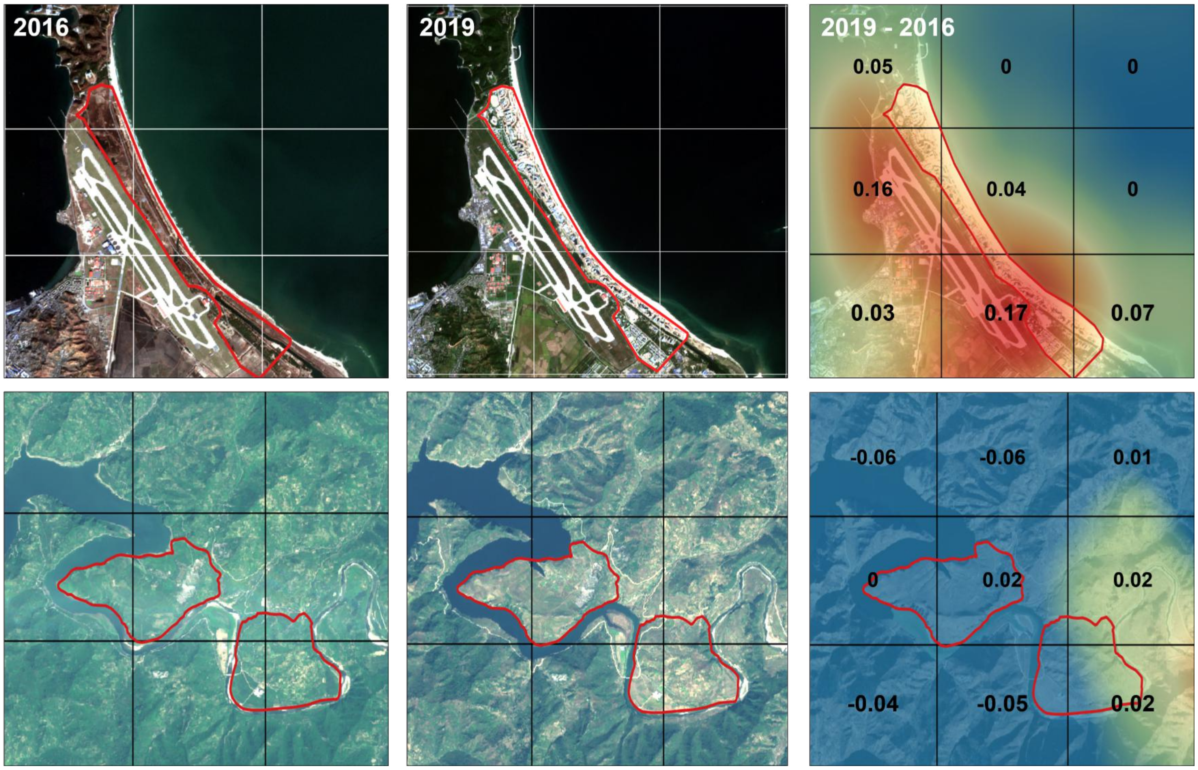


Figure 4: The Sentinel-2 images and the heatmap to show changes in model predictions between 2016 and 2019. The top pictures present the recently constructed Kalma tourist project of Wonsan city. The bottom pictures present industrial development areas in Wiwon county. The boundaries of these development projects are drawn as red lines.

account for other region-specific determinants of economic development, we include province indicator variables (e.g., province fixed effects) and the district's population and surface area in the regression.

Regression coefficients from ordinary least squares estimation are reported in Table 1. Columns (1) and (2) use differences in the log of *siScore* and nightlight as outcome variables to capture percentage changes in these proxies from 2016 to 2019, respectively. Column (1) suggests that areas more distant from the country's capital, Pyongyang, or the nearest city, including provincial capitals, are associated with fewer development activities during 2016-2019. We also find that EDZ regions designated for agriculture or tourism show increased development scores relative to EDZ regions with industrial or export processing sites and non-EDZ regions. Coefficient estimates of the major mining sites are either economically or statistically nonsignificant. Interestingly, while we do not observe development activities around previously known nuclear test sites, we detect substantial development in provinces with uranium mine sites.

In contrast, as shown in column (2), using nightlight as the outcome variable does not indicate the same development patterns as those observed with *siScore*. Specifically, distance to the capital or EDZ regions with agriculture or tourism development is no longer correlated with changes in nightlight intensity. However, nightlight data suggest a decrease in intensity in regions developed for the export processing industry, which is plausible given that economic sanctions targeted restricting North Korea's exports. In columns (3) and (4), we report coefficient estimates from using an indicator for positive changes in *siScore* and nightlight intensity, respectively, as the outcome measure. The results for *siScore* in column (3) are qualitatively similar to those in column (1), whereas the estimates using nightlight intensity in column (4) are inconsistent with those in column (2).

We believe there are at least two reasons for the discrepancy between *siScore* and nightlight. First, these measures may capture different aspects of economic development. For example, our model score is effectively identifies structural or sectoral changes in regions such as barren lands being transformed into agricultural fields or rice paddies into factory buildings or infrastructure sites. On the other hand, nightlight intensity can effectively detect the utilization of facilities and buildings. Second, nightlight intensity varies little in less developed regions with deficient nightlight luminosity levels. In the context of North Korea, the median grid's nightlight luminosity is zero, which does not necessarily mean that there is zero economic activity in that region. Moreover, given the constant shortage of electricity supply in rural regions, industrial and construction activities are likely to occur during the daytime and, as a result, are not well represented in data that detect nighttime light. Overall, our human-machine collaborative model's predictions using daytime satellite images can serve as a valuable complement to existing remotely sensed measures, such as nightlight intensity, and can help researchers and policymakers better understand the process of economic development in countries that lack sufficient labels.

Table 1: Grid-level regression estimates (2016-2019)

	(1)	(2)	(3)	(4)
	Δ ln(siScore)	Δ ln(NL)	$\mathbb{1}\{\Delta\text{siScore} > 0\}$	$\mathbb{1}\{\Delta\text{NL} > 0\}$
Proximity to economic and political hubs				
Log distance to NK-China and Russia border	0.089	0.058*	0.002	-0.055
	(0.071)	(0.028)	(0.031)	(0.043)
Log distance to Pyongyang	-0.256***	0.175	-0.142***	0.223
	(0.069)	(0.098)	(0.040)	(0.147)
Log distance to nearest city	-0.105**	0.083**	-0.034*	-0.090**
	(0.033)	(0.032)	(0.017)	(0.037)
Log distance to nearest major port	0.132**	0.022	0.044**	-0.018
	(0.048)	(0.024)	(0.015)	(0.054)
Economic Development Zone (EDZ)				
Agriculture development	0.243***	0.070	0.236***	0.307**
	(0.041)	(0.081)	(0.022)	(0.118)
Tourism development	0.297***	0.391	0.182***	0.643**
	(0.083)	(0.281)	(0.032)	(0.198)
Industrial development	-0.058	-0.083	-0.005	0.153*
	(0.128)	(0.165)	(0.130)	(0.077)
Export processing	-0.063	-0.378***	-0.038	0.199
	(0.108)	(0.047)	(0.143)	(0.127)
Mining site of key minerals				
Gold mine	0.083**	-0.024	0.005	0.020
	(0.031)	(0.028)	(0.019)	(0.038)
Coal mine	0.074*	0.003	0.039*	0.117**
	(0.034)	(0.024)	(0.022)	(0.049)
Copper mine	0.090	0.076*	-0.000	-0.018
	(0.050)	(0.040)	(0.030)	(0.049)
Iron mine	0.116	-0.126*	0.113**	-0.042
	(0.072)	(0.059)	(0.037)	(0.042)
Nuclear-related site				
Nuclear test site	-0.033	-0.098*	0.048	-0.011
	(0.087)	(0.046)	(0.047)	(0.113)
Uranium mine	0.375**	0.077	0.197***	0.110
	(0.156)	(0.072)	(0.042)	(0.078)
Province FE	Yes	Yes	Yes	Yes
Mean of outcome variable	-0.09	3.10	0.45	0.30
Observations	32578	32578	32578	32578

Notes: This table reports ordinary least squares (OLS) regression estimates. The outcome variable in columns (1) and (2) is the difference of logarithmized values between 2016 and 2019. Columns (3) and (4) use an indicator for positive change as the outcome variable. All specifications include province fixed effects, log of district population in 2008, and log of district area. Standard errors are clustered at province level and reported in parentheses. * denotes statistical significance at 0.10, ** at 0.05, and *** at 0.01.

Discussion

This paper presents a new approach to measuring economic activity in the absence of ground-truth labels to allow researchers to investigate low-income countries at scale. Our first contribution is the development of a human-machine collaborative algorithm to overcome the data challenges associated with machine-only approaches. While earlier computer vision techniques have significantly expanded the scope for measuring vital statistical information from satellite images, these models' applicability and performance have been largely conditional on access to high-resolution data or considerable quantities of ground-truth labels (*9, 14, 19*). This paper illustrates that our human-guided model can equal the performance of, or even outperform, data-driven alternatives using nightlight intensity and land cover classification data to predict local development activity. Thus, our approach complements these existing efforts by showcasing an approach that requires only publicly available satellite imagery combined with human knowledge on the underlying economics of the geography embodied in such images.

Our second contribution is to provide a reliable indicator of local economic development for North Korea, a country that releases almost no socioeconomic labels to the outside world. Our model suggests a concentration of development amid sanctions in regions closer to cities, including the country's capital, relative to regions farther away. We also find rapid regional development in state-led projects in agriculture and tourism development and areas with uranium mines. Admittedly, our model prediction cannot speak for other critical economic development dimensions, such as better education, improved health, and higher income. Nonetheless, our measure provides information on how the centrally planned economy allocates its resources across different regions to cope with mounting pressure from economic sanctions.

Our approach is the first of its kind to be applied to restricted regions such as North Korea. However, there remains considerable potential for applicability and improvement in future practice. First, the predictions can be enriched with an alternative source of remote sensing or other geolocated data. While we utilized publicly available satellite images, our model can readily be applied to other proprietary satellite images and aerial photographs to improve the prediction quality. Second, the model training can also be improved. Due to the multistage structure, noise in the initial clustering stage can propagate throughout training, degrading the performance. The clustering step, for example, can be designed more cohesively with the human guidance step in the form of active learning. Last, the applicability of our model is not limited to North Korea but extends to a broad set of developing countries that may benefit, as we show for Cambodia and Nepal in the Supplementary Materials. By further developing our model to generate reliable and inexpensive indicators based on hyperlocal socioeconomic data, our research can potentially support the international community in solving humanitarian and economic challenges.

References

1. M. Burke, A. Driscoll, D. B. Lobell, S. Ermon, *Science* **371** (2021).

2. Central Bureau of Statistics of the DPR Korea, *DPR Korea 2008 Population Census: National Report* (Central Bureau of Statistics of the DPR Korea, Pyongyang, 2009).

3. C. B. of Statistics of the DPR Korea, UNICEF, *DPR Korea Multiple Indicator Cluster Survey 2017, Survey Findings Report* (Central Bureau of Statistics of the DPR Korea and UNICEF, Pyongyang, 2017).

4. M. Hong, M. S. Cha, E.-l. Joung, H. Kim, *North Korea National Market Information: Focus on Status of Formal Markets* (Korea Institute for National Unification, Seoul, 2016).

5. S. K. Lee, S. Y. Lee, *Current Status of North Korean Companies in the 2000s: Focus on the Analysis of Official Media* (Korea Institute for Industrial Economics and Trade, Seoul, 2014).

6. Y. S. Lee, *Journal of Urban Economics* **103**, 34 (2018).

7. J. Crespo Cuaresma, *et al.*, *Palgrave Communication* **6** (2020).

8. N. Jean, *et al.*, *Science* **353**, 790 (2016).

9. J. L. Abitbol, M. Karsai, *Nature Machine Intelligence* **2**, 684 (2020).

10. S. Han, *et al.*, *proc. of the AAAI* (2020).

11. I. Tingzon, *et al.*, *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XLII-4/W19**, 425 (2019).

12. E. Sheehan, *et al.*, *proc. of the ACM SIGKDD* (2019), pp. 2698–2706.

13. M. Fatehkia, R. Kashyap, I. Weber, *World Development* **107**, 189 (2018).

14. K. Ayush, B. Uzkent, M. Burke, D. Lobell, S. Ermon, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20* (2020), pp. 4410–4416.

15. M. Caron, P. Bojanowski, A. Joulin, M. Douze, *proc. of the ECCV* (2018), pp. 132–149.

16. K.-s. Lee, S.-j. Hong, Y. S. Jang, C.-w. Jang, *North Korea knowledge dictionary* (Institute for Unification Education, Seoul, 2016).

17. M.-c. Cha, *Economic Zones of the DPR Korea* (Oegungmunchulpansa, Pyongyang, 2018).

18. J. V. Henderson, A. Storeygard, D. N. Weil, *American Economic Review* **102**, 994 (2012).

19. C. Yeh, *et al.*, *Nature Communications* **11** (2020).

20. National Spatial Data Infrastructure Portal, Digital map of Korea (2018).

21. V. Cha, L. Collins, The markets: Private economy and capitalism in north korea? (2018).

22. S. K. Lee, C. Kim, H. J. Bing, S. Y. Lee, *North Korean Companies: Manufacturing and Energy Companies Handbook* (Korea Institute for Industrial Economics and Trade, Sejong, 2014).

23. W.-s. Shim, S. K. Lee, S. Y. Lee, H. J. Bing, C. Kim, *Establishment of DB contents for North Korean companies (mining industry, electricity) and trends in industries and companies* (Korea Institute for Industrial Economics and Trade, Sejong, 2015).

24. Ministry of Environment of Republic of Korea, *Directive No.1317: Guidelines for Land Cover Maps* (2018).

25. B. M. Ahn, S. C. Kim, *Analysis of North Korean Port Status* (The Korea Transport Institute, Goyang, 2012).

26. J. W. Seo, S.-w. Noh, *Northeast Asia and North Korea Transport Logistics webzine* **2014**, 1 (2014).

27. J. Bermudez, V. Cha, Punggye-ri nuclear test site: Imagery supports rok and u.s. government reservations about permanent disablement (2019).

28. J. Bermudez, V. Cha, D. Kim, Recent activity at the pyongsan uranium concentrate plant (nam-chon chemical complex) and january industrial mine (2021).

29. J. Bermudez, Yongbyon declassified part ii: Progress on building irt-2000 reactor (2018).

30. S. Michalopoulos, E. Papaioannou, *Annual Review of Economics* **10**, 383 (2018).

31. S.-J. Kim, J. H. Hong, North korea's standard of living from cross-country studies, *Policy Research Series 19-03*, Korea Institute for National Unification (2019).

32. World Bank, World Bank country and lending groups. Accessed: 2021-03-04.

33. S. M. Mun, *Understanding North Korea's Economy from Statistics* (Bank of Korea, 2014), chap. Introducing North Korea's National Income Statistics and Comparing Income Levels.

Acknowledgments

This work was supported by the Institute for Basic Science (IBS-R029-C2). Sangyoon acknowledges research funding supported by Hong Kong's Research Grants Council (GRF 17504920). We appreciate Cho Tae-hyung, Kim Hyun-ok, Kim Jinseok, Kim Kyoo-chul, Kim Min-jung, Lee Jongkyu, Park Hyanggi and Park Kyung Ryul for their precious comments. We acknowledge Lee Seok-gi for sharing data of North Korea. We appreciate POG makers: Bahk Junbeom and Kang Mijin.

Supplementary materials

Materials and Methods
Supplementary Text
Figs. S1 to S3
Tables S1 to S4
References *(4-10)*

Supplementary Materials

1 Socioeconomic Data of North Korea

1.1 Official Data Release

Data published by North Korea are scarce. The 2008 Population Census conducted by the Central Bureau of Statistics of North Korea and the United Nations offers the last county-level population statistics (*2*). While external organizations such as UNICEF have been allowed to conduct province-level surveys in the country (*3*), no official statistics at the subprovincial level have been released since 2008. The country's total population was 23,349,859 in 2008, with the top five cities listed as Pyongyang (2,708,648), Hamhung (668,557), Chongjin (667,929), Nampo (366,815), and Wonsan (363,127).

1.2 Administrative Boundaries

Administrative boundaries change over time. To be consistent with the census data, we utilize the distinction made from 2008. This dataset contains boundary information for 178 districts. The original dataset was constructed by the North Korean government and was then shared with the Institute for Peace Affairs (IPA) in South Korea, who released it publicly at `http://www.cybernk.net/`. Among 178 administrative districts, 27 are designated counties (called *'si'* in Korean) and the others as counties (called *'gun'*). The 27 cities are Pyongyang, Anju, Chongjin, Haeju, Hamhung, Hoeryong, Huichon, Hyesan, Jongju, Kaechon, Kaesong, Kanggye, Kim Chaek, Kusong, Manpo, Munchon, Nampo, Pyongsong, Rason, Sariwon, Sinpo, Sinuiju, Songrim, Sunchon, Tanchon, Tokchon, and Wonsan.

1.3 Building the Footprints Dataset without Heights

Building footprints can be used as a proxy of economic development in a given region. Modern GIS programs can compute the built-up ratio or the ratio of land occupied by buildings based on such data. Two relevant statistics can be computed from building footprint data: *floor area* (FA, area of floor space) and *gross floor area* (GFA, the total area of all floors, when the height information is given). Some research has utilized the geospatial footprints from crowd-generated OpenStreetMap (OSM) as an alternative to the ground truth. However, OSM has low coverage in North Korea and is unreliable, as it calls for the voluntary participation of individuals who have local knowledge.

Instead, we utilized the digital map of North Korea released by the National Geographic Information Institute (NGII) in South Korea (*20*) (Fig. 5(B)). This digital map contains building footprints in GIS format (without any height information) for 2008, 2014, and 2017. These maps have a large coverage of the nation. The 2014 version covers 3.6 million buildings in 70% of the North Korean land. The 2017 data has only been partially released and is still being compiled. The construction process is known to utilize high-resolution satellite images and manual inspection of GIS experts. We used the 2014 map, which was released at the National Spatial Data Infrastructure Portal (`http://www.nsdi.go.kr/lxmap/index.do`). We calculated built-up ratio (i.e., *floor area*) as a ground truth over each grid.

1.4 Market Area Data

The regional socioeconomic status may be inferred from the number and size of authorized markets in North Korea (called *Jang-Madang*). Jang-Madang have been recognized as a significant development representing the growing private economy in North Korea. As of July 2002, the North Korean government implemented new reforms to save the economy from further recession, officially allowing state-sanctioned markets to operate (*21*).

The Korea Institute for National Unification (KINU) (*4*) in South Korea has collected the geolocations of North Korean markets since 2015 based on multiple sources, including interviews, satellite imagery, and other secondary resources. Later, the Center for Strategic and International Studies (CSIS) (*21*) and the North Korea Development Institute (NKDI) in South Korea also repeated the effort to detect Jang-Madang in 2018. The number of vendor stalls in each market is estimated by the size of the market seen in satellite images (i.e., these markets are state-built and have particular recognizable shapes). Both institutes considered each stall to occupy a space between 1.4 and 1.9 m^2.

We combined the above market maps to generate a unified version. We utilized the historical imagery function of Google Earth Pro to cross-check the 2015 and 2018 market locations and confirmed 404 markets based on satellite images from 2015 (with an average area per market of 4,552m^2 and an average number of stalls per market of 2,707) and 442 markets based on satellite images from 2018 (with an average market area of 4,413 m^2 and 2,691 stalls per market).

1.5 Industry Listing Dataset

The next data source we considered is the industry listings. The Korea Institute for Industrial and Economics and Trade (KIET) in South Korea has been monitoring news articles published by two major North Korean news outlets, *Rodong Sinmun* and *Minju Chosun* since 2000 to compile a list of company names that appear in the news. Three versions were published through 2013 (*5, 22, 23*). We combined these lists and further expanded them by following the same methodology listed by the original KIET report to include new company names mentioned after 2013. The report classifies North Korean industry into nine categories: (1) Food, Beverage, and Tobacco, (2) Textile, Garment, and Shoes, (3) Chemistry, (4) Building Materials, (5) Primary Metal Industry, (6) Electronics, (7) Transport Machinery, (8) Power, and (9) Furniture, Wood, and Miscellaneous Goods. Each of these types has subcategories.

We expanded the industry listing as of 2019, totaling 2,407 (2016), 2,546 (2017), 2,570 (2018), and 2,639 (2019) company names as shown in Fig. 6(D). Our compiled dataset contains the operating districts for each industry, which allows us to compare the district-level economic growth over the observed years. This industry listing is limited in its coverage and context. For example, the list misses the names of critical military and mining-related units. Furthermore, we could not check which companies are in real operation or infer their scale in terms of factors such as employment and investment. Noting these caveats, we make use of the industry listing dataset as one of the reference points.

1.6 Land Cover Classification Data

Land cover classification maps for the 30 m-resolution satellite imagery were released by the Ministry of Environment (MoE) in South Korea for 1989, 1999, 2009, and 2019. This classification task requires extensive human labor for validation and hence has been conducted only once every decade. We use the 2019 version, which is the closest to our observed period. Land cover types in this map (Fig. 5(A)) include (red color) used area, [6] (yellow) agricultural land, (dark green) forest, (green) grass, (purple) wetland, (aqua-blue) barren, and (blue) water. The MoE reports the classification accuracy to be approximately 70% (*24*). MOE controls access

[6] Note that ‘used area’ from the land cover classification data and ‘built-up area’ from the building footprint data are different. The former refers to how urbanized a given area is, whereas the latter refers to the total amount of construction represented by buildings in that area. The amount of ‘used area’ is utilized to generate a data-guided POG. The built-up ratio is used as a proxy of the ground truth economic development index to evaluate our models. See Fig. 5 for comparison.

to the land cover classification map of North Korea.

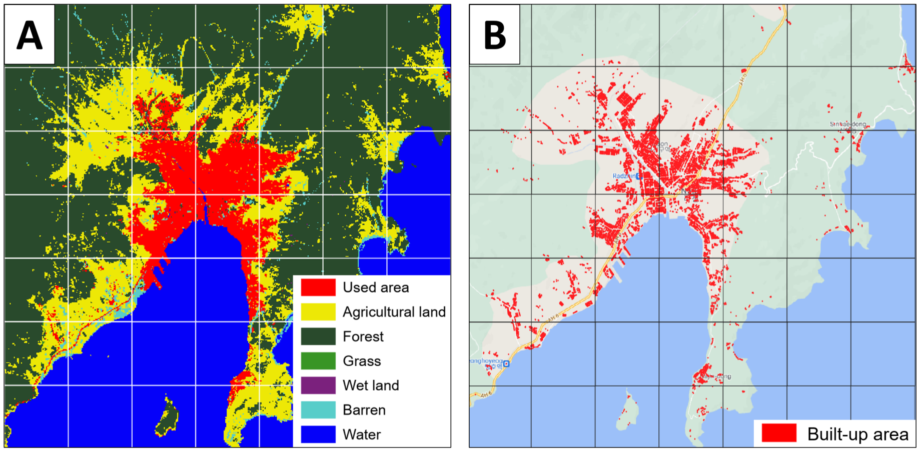


Figure 5: (A) 2019 land cover map and (B) 2014 building footprint data of Rason city, North Korea. The used area is calculated from the land cover map, while the built-up ratio is calculated from building footprint data.

1.7 North Korea Mine Data

To determine the location of the mining industry in North Korea, we use data from I-RENK (Information system for Resources in North Korea, `http://www.irenk.net/`), founded by SONOSA (South-North Korea Exchanges and Cooperation Support Association) in South Korea. I-RENK provides locations of significant mines for gold, copper, magnetite, molybdenum, coal, zinc, uranium, apatite, tungsten, iron, graphite, and rare earth elements (Fig. 6 E). The total number of mines listed in this dataset is 399.

1.8 Locations of Major Infrastructure

The locations of other major development-related infrastructure were collected through reports published by the South Korean government. Examples include the eight central trade locations (*25*): Chongjin port, Hungnam port in Hamhung, Rajin port, Sunbong port in Rason, Wonsan port, Nampo port, Haeju port, and Songrim port. Additionally, twelve border checkpoints facing China and Russia have been identified by the Korea Transport Institute (*26*), including Sinuiju, Manpo, Hyesan, Heoryeong, and Rason.

Information about special economic zones is relevant to this study, as it indicates the North

Korean government's strategic development initiative or the regime's message visible to outside observers. The Kim Jong-un regime has established 22 economic development zones (EDZs, *Gyeongje-gaebalgu*) to overcome the economic crisis caused by international isolation (*17*). EDZs each represent different development goals, from agricultural to industrial, export, tourism, technical, and green economy. North Korean companies within these EDZs were allowed to receive foreign investment prior to sanctions.[7] The average size of EDZs is 5.21 km^2, and the EDZ locations were identified based on reports published by North Korea (*17*).

We also gathered information about nuclear test sites. Studies on North Korea's nuclear programs date back to the 1950s. North Korea announced its withdrawal from the Nuclear Nonproliferation Treaty (NPT) in 2003 and declared in 2005 that it had achieved nuclear power status. Its six nuclear tests followed the nuclear weapon development programs in 2006, 2009, 2013, 2016 (January and September), and 2017. Reports have confirmed several nuclear facility complexes in Pyongsan, Yongbyon, Punggye-ri, and others. We found grids of satellite images where those nuclear facilities are located based on articles from CSIS (*27–29*) and geodatabase sources such as 38North (`http://38northdigitalatlas.org/`).

2 Satellite Imagery Data

2.1 Daytime Satellite Imagery

This research utilizes Sentinel-2 satellite images that provide the finest spatial resolution (10 m) of North Korea among public resources. A 10 m resolution is equivalent to a zoom level of 14 (9.547 meters/pixel). Zoom levels are a tile-based system that divides the entire satellite images into nonoverlapping 256×256 pixel square-shaped images (called tiles). The zoom level determines how many tiles are needed to show the entire world. For zoom level 0, the entire earth is displayed on a single tile. Zoom level 1 requires 2×2 tiles, whereas the world is divided into $2^{14} \times 2^{14}$ tiles for zoom level 14.

This research uses images taken in the summer season between June and September from 2016 to 2019. We used images with a cloud ratio less than 10%. Satellite images were downloaded via the United States Geological Survey (USGS) Earth Explorer at `https://earthexplorer.`

[7] In the previous Kim Jong-il regime, the predecessor of Kim Jong-un, five special economic zones (SEZs, *Gyeongje-teukgu*) had existed.

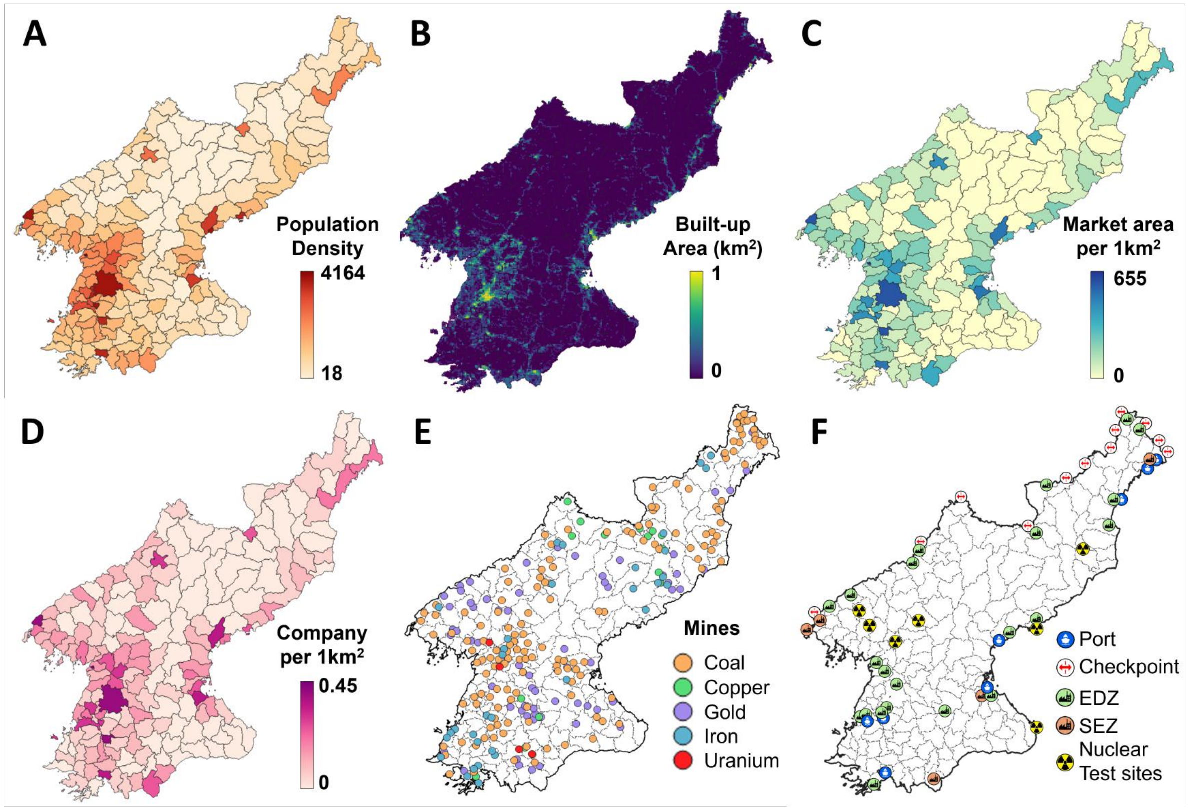


Figure 6: Comparisons of various socioeconomic data of North Korea. (A) The grid-level built-up ratio in 2014. A total of 9,959 of 32,578 grids (30.6%) had a null value for built-up areas in 2019. (B) Population density in 2008, (C) market areas per 1 km^2 in 2018, and (D) numbers of companies per 1 km^2 in 2019. (E) Locations of coal, copper, gold, iron, and uranium mines. (F) Locations of major ports, border checkpoints for trade, EDZs, SEZs, and nuclear test sites.

usgs.gov. The Sentinel-2 data comprise 13 multispectral bands, including visible, near-infrared (NIR), and shortwave infrared (SWIR) bands. Our research used three visible bands: red, green, and blue. In addition to the North Korean region, we also used images from Nepal and Cambodia to validate the model. We used images of the same resolution (i.e., zoom level 14) gathered from the World Imagery data resource via the ESRI ArcGIS REST API. We considered data from the same years, 2016 to 2019, to evaluate these countries' census data predictions. We used ESRI images instead of Sentinel-2 because ESRI has analysis-ready 256×256 pixel imagery datasets between 2016 and 2019 for these countries. ESRI images are not available during this period for many areas of North Korea.

Several procedures were followed after downloading the satellite images. The images were first combined in the QGIS program to construct a single large map of North Korea for each year. Then, the image map was split into 256×256 pixel square tiles for analysis, each covering an area of 5.97 km^2. The model used tiles whose three or more vertices belonged to district polygons (Fig. 7) and discarded others. A total of 32,578 image tiles were generated for each year's data, with an average of 183 tiles per county.

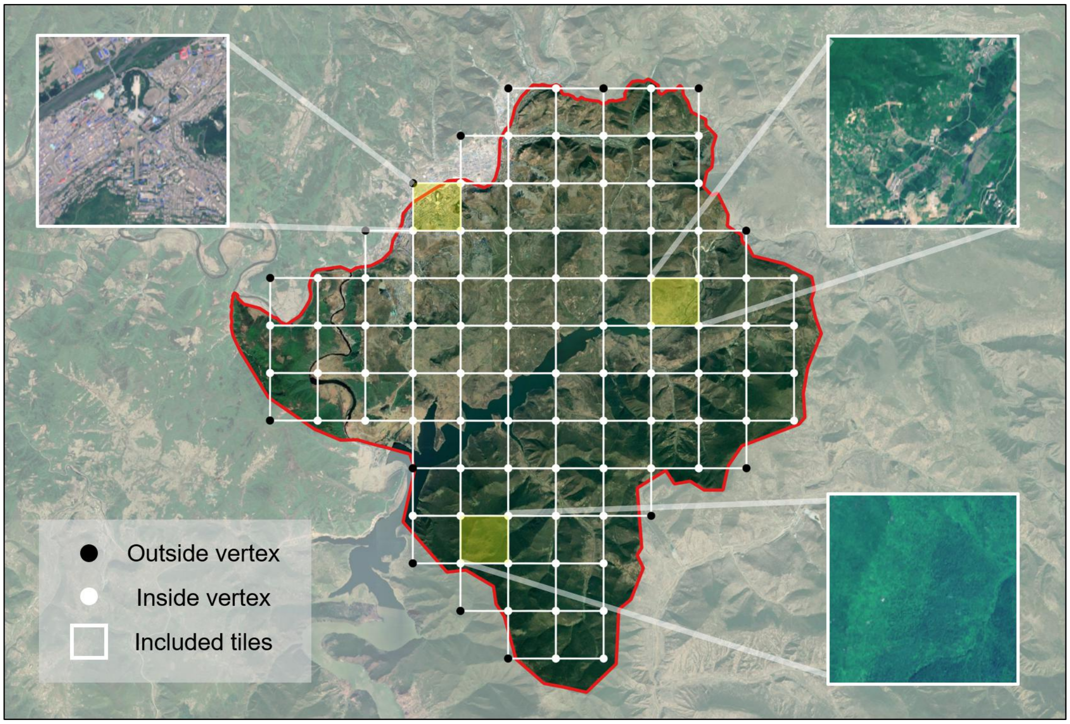


Figure 7: Methods for mapping districts and grid-level tiles for the zoom-level 14 satellite image of Hyesan City, North Korea. The model utilizes all tiles belonging to the target. At least three corners must belong to the district to be considered.

2.2 Night-light Satellite Imagery

One of the baseline models we employ in this research is based on the nighttime satellite imagery, released by the Earth Observations Group (EOG) via the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) at `https://eogdata.mines.edu/products/vnl/`. The VIIRS data cover 15 arc-second grids of the world since 2012, and annual aggregate versions exist until 2019. The aggregated dataset is used widely because it represents the average luminosity values of cloud-free and outlier-free images (e.g., fire).

The current research uses the annual VIIRS data, listed as the *Annual VNL V2* version, that filters out sunlit, moonlit, cloudy pixels, biomass burning, aurora, and background values. As outlined in our paper and several existing studies (*8*), nightlight satellite images have critical caveats. First, meaningful luminosity values are difficult to obtain for rural areas of low GDP countries like North Korea. Significant gaps in the luminosity values between rural and urban areas make comparisons challenging. Second, nightlight data are not suitable for granular inspection due to blooming and saturation effects (*30*).

3 Algorithm

Overview. We are given a total of N satellite images in a target area and these images are split into square-shaped grids, which are denoted as $\mathcal{I} = \{\mathbf{x}_i\}_{i=1}^{N}$. Our goal is to estimate the relative scores representing the economic development y_i of each grid image $\mathbf{x}_i$ by training a score model f_θ (i.e., $y_i = f_\theta(\mathbf{x}_i)$). This estimation of y_i is challenging because there is no ground truth available to train the network f_θ. We propose a *human-machine collaborative* algorithm that breaks the problem into multiple stages: (1) rely on the machine to group satellite images with similar visual features, (2) receive human guidance on the relative economic development of the image clusters, and (3) train the machine to produce a rankwise score model f_θ.

Stage 1: Clustering satellite images. Given the satellite image set $\mathcal{I}$, the first task is to create a cluster set $\mathcal{C} = \{c_i\}_{i=1}^{n_c}$ of visually similar images representing consistent economic activities. We start with an unsupervised clustering algorithm, DeepCluster (*15*). DeepCluster uses and updates an encoder e_ϕ by training with pseudo labels generated by k-means clustering over the embedded space of satellite images. For a given unlabeled satellite image set $\mathcal{U}$, the DeepCluster loss $\mathcal{L}_{dc}$ is defined as follows based on the cross-entropy loss function H:

$$L_{dc} = \frac{1}{|\mathcal{U}|} \sum_{\mathbf{x}} \mathbf{H}(W \cdot e_\phi(\mathbf{x}), \bar{\mathbf{y}}), \tag{1}$$

where W is the weight matrix for the clustering head and $\bar{\mathbf{y}}$ denotes the pseudo label of each image $\mathbf{x}$ from k-means clustering.

Rather than directly using DeepCluster (which has a limitation in its random initialization of pseudo labels), we add a novel modification by leveraging transfer learning to overcome the initial randomness. We set up an auxiliary dataset $\mathcal{X}$ in a weakly supervised fashion by randomly selecting 1,000 images and labeling them with coarse-grained category labels of urban, rural, and uninhabited (i.e., $\mathcal{X} = \{(\mathbf{x}_i, \hat{\mathbf{y}}_i)\}_{i=1}^{1000}$). Similar methods have been used in (*10*). Then, using the auxiliary dataset $\mathcal{X}$, we pretrain the encoder e_ϕ with the classification head W' such that it minimizes the cross-entropy loss $\mathcal{L}_{class}$:

$$L_{class} = \frac{1}{|\mathcal{X}|} \sum_{(\mathbf{x}, \hat{\mathbf{y}}) \in \mathcal{X}} \mathbf{H}(W' \cdot e_\phi(\mathbf{x}), \hat{\mathbf{y}}). \tag{2}$$

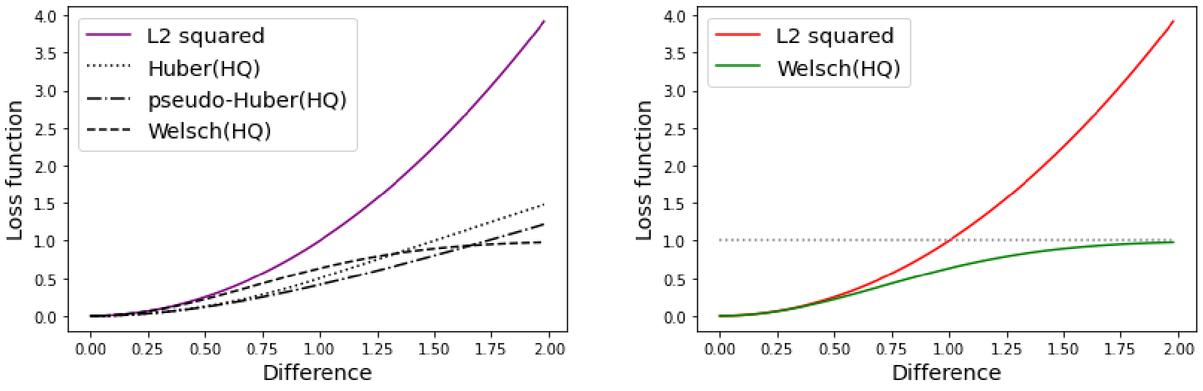
The pretrained classifier then divides the original image set $\mathcal{I}$ into three subsets: urban $\mathcal{I}_u$, rural $\mathcal{I}_r$, and uninhabited $\mathcal{I}_n$. The DeepCluster algorithm with the initialized encoder e_ϕ is applied separately to $\mathcal{I}_u$ and $\mathcal{I}_r$ to find two disjoint cluster sets $\mathcal{C}_u$ and $\mathcal{C}_r$. The uninhabited set $\mathcal{I}_n$ itself is regarded as a single cluster $\mathcal{C}_n$. The combined cluster $\mathcal{C}$ is the union of all three cluster sets (i.e., $\mathcal{C}$ = $\mathcal{C}_u \cup \mathcal{C}_r \cup \mathcal{C}_n$). To determine the number of clusters, we applied the silhouette analysis over the embedding space to measure the similarity of instances within a cluster.

Stage 2-1: Sorting clusters by human guides. We introduce a partial order graph (POG), which utilizes cluster set $\mathcal{C}$ from the previous stage and represents its rank orders. Each identified cluster contains satellite grid images of visually similar economic traits. Furthermore, the relatively small cluster size (i.e., 23 for North Korea) compared to the grid count (i.e., 32,578 for North Korea) allows humans to intervene and generate POGs based on their knowledge. Generating POGs at the cluster level is more efficient than grid-level pairwise comparisons, which would prohibit human participation. We recruited ten human participants who each generated a POG by inspecting the 23 clusters.

Stage 2-2: Combining POGs via an ensemble method. Given ten human-guided POGs, our next task is to aggregate diverse views into a single POG via an ensemble method. Formally, let us assume M human participants and denote the POG ranking from the m-th human as R^m (i.e., R^m, $m = 1, 2, 3, ..., M$). A typical ensemble approach takes the average of the collected POG rankings, which is identical to finding a ranking vector R^* that minimizes the L2 distance between every pair of POG ranks. This method, however, is sensitive to outliers (e.g., some POGs may have substantially different ordering than others). Instead, we adopt a half-quadratic (HQ) minimization algorithm to aggregate POGs, which is known to minimize the outlier effect by assigning lower weights to rarely appearing inputs. Fig. 8(A) illustrates the behaviors of

```
Algorithm 1: Algorithm for ensembling POGs
Input  :Rankings $R^m$, $m = 1, 2, ..., M$, HQ function $\delta(\cdot)$.
Output :Ensembled ranking $R^*$.
repeat
    $\alpha_m = 1 - \delta(||R^m - R^*||_2), m = 1, 2, ..., M$
    $w_m = \alpha_m / \sum_j \alpha_j$
    $R^* = \sum_m w_m R^m$
until Convergence of $R^*$;
```

three commonly used HQ algorithms over the distance between two ranks, demonstrating that the HQ function loss is less sensitive than the L2 loss under large ordering deviance. Among these three HQ functions, we utilized the Welsch function for the ensemble process due to its asymptotic upper bound (see Fig. 8(B)) and its robustness against outliers. The ensemble algorithm described in Algorithm 1 yields a single combined POG structure. The algorithm uses a one-dimensional k-means algorithm to consolidate clusters of similar economic development levels and sorts the clusters according to their average ranks.[8] This final POG reflects the rank ordering from all human guides while minimizing discordance among conflicting POGs.



(a) HQ functions vs L2 loss

(b) Welsch function

Figure 8: Three representative HQ functions demonstrate their superiority over the conventional L2 loss, especially for large ordering deviances. We utilize the Welsch function for the ensemble process because of its robustness against outliers.

Stage 3: Training a rankwise scoring model. The last stage is to train a CNN-based score model f_θ that produces a score between 0.0 and 1.0 for every satellite grid. This model learns a

[8] To reduce label noise, we prune the POG by removing clusters placed near the decision boundary.

```
Algorithm 2: Algorithm for training the score model.
Input  : Partial order graph $G$,
         the number of total epochs $Epochs$,
         the batch size per cluster $n_s$,
         the initial scoring model $f_\theta$
Output : Trained scoring model $f_\theta$
$\mathcal{P} \leftarrow$ Extract all available paths from $G$
for $j \leftarrow 1$ to $Epochs$ do
    $Scores \leftarrow []$
    foreach path $p_j \in \mathcal{P}$ do
        foreach cluster $c \in p_j$ do
            Select random batch $\mathcal{B}$ with size $n_s$ from $c$
            $Scores$.insert($f_\theta(\mathcal{B})$)
        end
        $\mathbf{s} \leftarrow Concat(Scores)$
        $\mathbf{r}_j \leftarrow [0, 1, ..., |p_j|]$
        $L_{score} = \frac{1}{n_s} \sum_{i=1}^{n_s} ||h(\mathbf{s}, i) - \mathbf{r}_j||_2^2$
        Update $f_\theta$ via back-propagation
    end
end
```

continuous-scale score for every satellite grid image such that these scores align with the POG ranking in the previous stage. The model training process is guided to preserve the POG ordering. Let us denote $\mathcal{P}$ as a collection of every ordered path from a POG. After randomly selecting one path p from $\mathcal{P}$ (i.e., $p = \{c_1, c_2, ..., c_m\} \in \mathcal{P}$, where c_i is the i-th cluster in the selected path p and m is the length of the path), we compute the score vector $\mathbf{s}$ as follows:

$$\mathbf{s} = [f_\theta(x_{c_1}), f_\theta(x_{c_2}), \cdots, f_\theta(x_{c_m})], \tag{3}$$

where x_{c_i} is a sampled image from each cluster c_i in the path p, f_θ is the scoring model, and $[\cdot]$ represents concatenation.

The Spearman rank correlation measures the degree of rank correlation between two sequences. By maximizing the rank correlation between the score vector $\mathbf{s}$ and the actual order $\mathbf{r}$ of a given path p, the scoring model f_θ can produce scores that mimic the POG ordering (Eq. 4):

$$\max \left(1 - \frac{6||h(\mathbf{s}) - \mathbf{r}||_2^2}{m(m^2 - 1)}\right), \tag{4}$$

where $h(\mathbf{s})$ returns the sorted order of s.

Assuming that the length of p is m and the complete set of score vectors is $\mathcal{S}$ (i.e., $\mathcal{S}$ = $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{n_s}\}$), we can maximize Eq. 4 by minimizing the following loss function:

$$L_s = \frac{1}{\mathcal{S}} \sum_{\mathbf{s} \in \mathcal{S}} ||h(\mathbf{s}) - \mathbf{r}||_2^2. \tag{5}$$

However, since mapping scores to ranks is a nondifferentiable process, typical deep learning models with gradient descent cannot be used to maximize this correlation. Instead, we apply the approximation in Eq. 6. This differentiable ranking function σ matches two variables, e_i and e_j, returns a value close to 1.0 if e_j is substantially larger than e_i, and returns a value close to 0.0, otherwise:

$$\sigma(e_i, e_j) = \frac{1}{1 + e^{-\lambda(e_j - e_i)}}, \text{ where } \lambda > 0. \tag{6}$$

Here λ is the hyperparameter that controls the precision of the rank estimation. Given the score vector s, the rank vector $h(\mathbf{s})$ is computed by summing the matching results among elements in the score vector. The rank of the i-th element in s can be calculated as follows:

$$h(\mathbf{s}, i) = \sum_{k=1;\ k \neq i}^{n} \sigma(\mathbf{s}^i, \mathbf{s}^k), \tag{7}$$

where s^i represents the i-th element of s. For every batch of clusters in the path, the Spearman rank correlation between the approximated rank of the corresponding image and that based on the POG order is maximized via the L_s loss (Eq. 5). The algorithm for training the score model is presented in Algorithm 2.

4 Performance Evaluation

4.1 Partial Order Graph (POG) evaluation

The POG structure was proposed to capture the level of human-perceived economic development from satellite images. Fig 9 illustrates the core concept of this process. Ten human participants with different backgrounds (i.e., economists, satellite imagery experts, and North Korean defectors) were recruited to build POGs. Here we quantitatively measure the reliability of the POGs contributed by these participants.

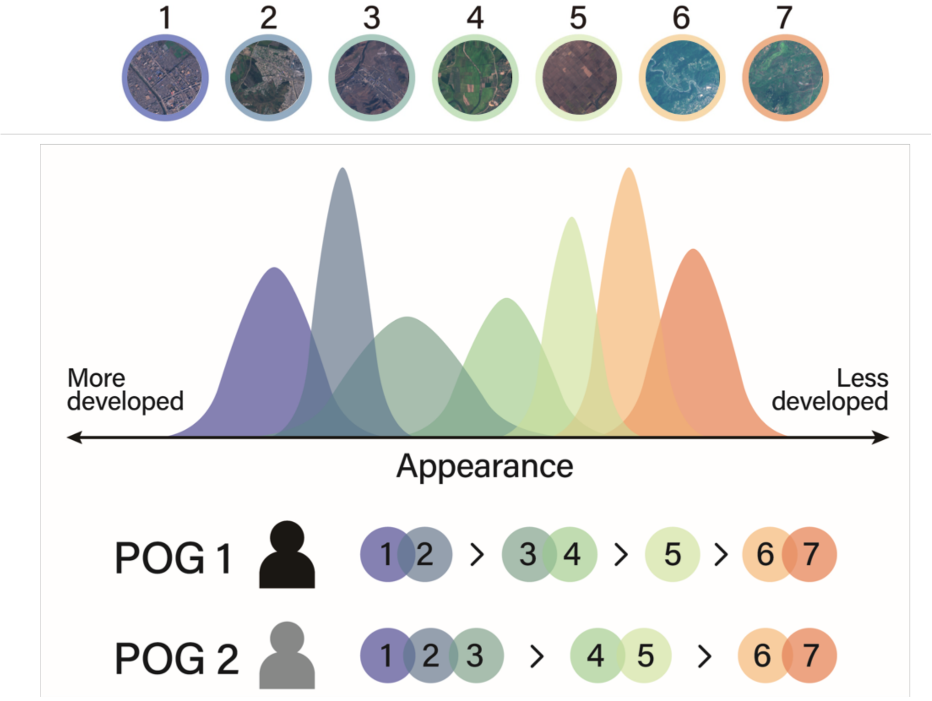


Figure 9: Illustration of the POG generation process. Clusters contain satellite images that show similar visual patterns of civil artifacts. Human participants order these clusters according to the perceived economic development level.

We utilized Kendall's rank correlation coefficient (τ_B) to evaluate the monotonicity between every pair of human-guided POGs. The coefficient ranges from -1 to +1, where -1 represents negative agreement, and +1 represents positive agreement. Table 2 displays the pairwise correlations across all participants. Most τ_B values show positive correlations close to 1, indicating high agreement among participants. As shown in Table 3, we confirm an overall high correlation at the group level. Despite the small sample, POGs generated by economists show the highest correlation with the utilized ground-truth data, followed by those generated by North Korean defectors and GIS experts. Moreover, all participants completed the ordering task within 1 to 2 hours, which shows that the work is feasible for humans. We nonetheless expect a higher degree of discordance when hiring participants via online crowdsourcing platforms.

Next, we evaluate the quality of the aggregated POG. Instead of training all POG ranks at once, our model produces a single representative POG employing an ensemble process via an HQ minimization algorithm. We investigated whether the combined POG accurately reflects the

Table 2: Cross-participant agreement measured by the Kendall-τ correlation based on POGs. (Sat: satellite expert, Loc: North Korean locals and defectors, Eco: economists)

Participant	Sat1	Sat2	Sat3	Loc1	Loc2	Loc3	Eco1	Eco2	Eco3	Eco4
Sat1	-	0.714	0.905	0.613	0.655	0.786	0.880	0.791	0.825	0.842
Sat2	0.714	-	0.795	0.768	0.681	0.804	0.763	0.725	0.696	0.774
Sat3	0.905	0.795	-	0.667	0.723	0.782	0.915	0.789	0.874	0.843
Loc1	0.613	0.769	0.667	-	0.589	0.752	0.729	0.643	0.561	0.703
Loc2	0.655	0.681	0.723	0.589	-	0.648	0.652	0.793	0.693	0.590
Loc3	0.786	0.804	0.782	0.752	0.648	-	0.837	0.851	0.672	0.809
Eco1	0.880	0.763	0.915	0.730	0.652	0.837	-	0.788	0.790	0.833
Eco2	0.791	0.725	0.856	0.643	0.792	0.724	0.788	-	0.800	0.759
Eco3	0.825	0.696	0.874	0.561	0.693	0.672	0.790	0.800	-	0.821
Eco4	0.842	0.774	0.843	0.703	0.590	0.809	0.833	0.760	0.821	-

Table 3: Cross-group agreement measured by the Kendall-τ correlation based on the ensemble rank of POGs.

Participant Type	Sat	Loc	Eco	Total
Satellite expert (Sat)	-	0.816	0.88	0.912
Local expert (Loc)	0.816	-	0.869	0.925
Economics expert (Eco)	0.662	0.869	-	0.966
Total	0.912	0.925	0.966	-

different viewpoints of participants. Table 3 summarizes the degree of agreement measured by Kendall's tau correlation between every pair of individual and combined POGs. We confirm a high correlation, which indicates that the ensemble algorithm produces appropriate meta-information.

4.2 Model Score Evaluation

We evaluate the model scores at two levels: grid and district. For the grid-level predictions, we use the log-scaled built-up ratio as the ground truth. Three evaluation metrics are used: R-squared value, Spearman correlation, and Pearson correlation. For the district-level predictions, which are the averages across all grids belonging to the same district, we use multiple economic scales as ground truth, including the built-up ratio, population density, markets per square kilometer, and industry count per square kilometer. We again use the log-scaled values for both the prediction scores and partial ground truth, as in related research (*8, 10*). We use the R-squared value for the evaluation metric.

Table 4: Evaluation of the model from 2016 to 2019

Model	Ground truth	Metric	2016	2017	2018	2019	Average
Grid-level	Built-up ratio	R-squared	0.4197	0.4132	0.5058	0.5007	0.5348
		Pearson	0.6478	0.6428	0.7112	0.7076	0.7313
		Spearman	0.6799	0.6684	0.7559	0.7528	0.7740
District-level	Built-up ratio	R-squared	**0.7689**	**0.7255**	**0.8070**	**0.8280**	**0.8286**
	Population density		0.7058	0.6733	0.7239	0.7417	0.7522
	Markets per 1km^2		0.5187	0.5326	0.5242	0.5397	0.5582
	Market areas per 1km^2		0.6595	0.6356	0.6663	0.6869	0.6988
	Market stalls per 1km^2		0.6334	0.6137	0.6286	0.6563	0.6677
	Industry per 1km^2		0.3816	0.3500	0.3496	0.3714	0.3853

Table 5 displays the evaluation results across the observed years. The averaged correlations are shown in the last column. We confirm reasonably high correlations at both the grid-level and district-level. The district-level results show the highest correlation with the built-up ratio dataset.

4.3 Baseline Comparisons

We next evaluate the model by comparing it with four baselines. We compute the Spearman correlation between the predicted scores on the built-up ratio dataset. The four baselines for comparison are as follows:

- **The nightlight data-guided POG model** is based on the observation that nightlight intensity positively correlates with the economic development index. We extracted the aggregated nightlight intensity value for each grid-level and used it to represent the cluster's economic indicator. Clusters can then be transformed into a POG based on their average luminosity scores without any human input.

- **The land cover-guided POG model** uses land cover classification data that contain rich contextual information on economic development from land usage. We compute the ratio of land labeled as the 'built-up' category for each grid, and then clusters were ordered into a POG based on this ratio.

- **The nightlight regression model** does not use any POG structure or deep learning model but directly utilizes the light intensity value from nighttime satellite imagery.

- **The land cover regression model** similarly applies regression on the ratio of land labeled ‘built-up’ from the land cover classification data.

Table 5: Comparison with possible baselines

Method	Data Source	Spearman Corr.
Human-guided POG	None	**0.7740**
Data-guided POG	Nightlight	0.5682
	land cover	0.7695
Regression	Nightlight	0.5096
	land cover	0.7464

The nightlight intensity, which is currently utilized as a proxy of economic development in many studies (*6, 7*), shows sufficient performance for both methods (i.e., Spearman correlation above 0.5). The land cover-based models perform even better because of the rich contextual information from manual labels. Table 5 displays the results, demonstrating that the human-machine collaborative method achieves outstanding performance compared to all baselines. Our model produces comparable results to the land cover data-guided model, which requires high resolution data and manual inspection. In addition, we emphasize another benefit of the human-machine collaborative algorithm in that it enables the score model to detect artifacts based on expert knowledge, where land cover data cannot be detected. For example, Fig. 10 shows that our model can distinguish between densely populated areas and simple built-up areas.

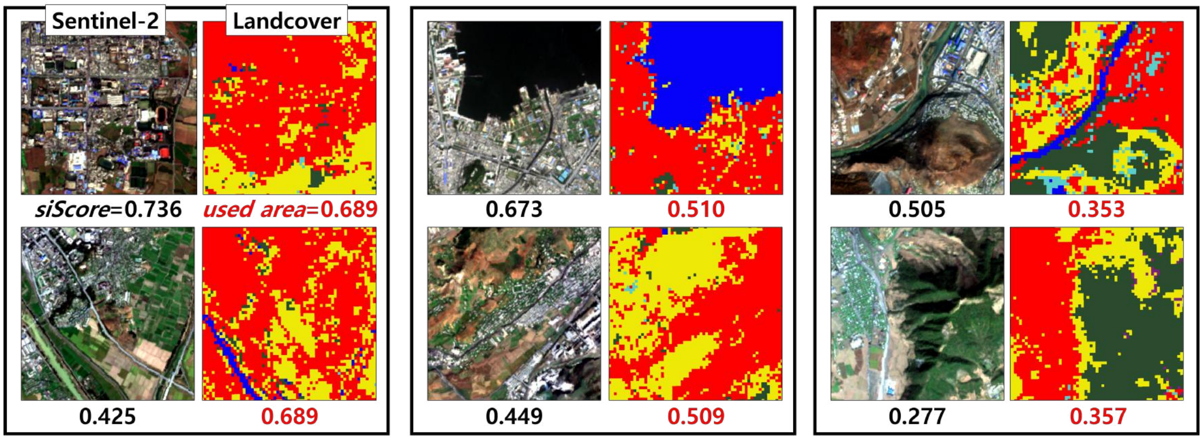


Figure 10: Comparison of our model (black) with the land cover use-area ratio (red). The pair of pictures grouped in a box has a similar economic development degree based on the land cover data. Pairs of Sentinel-2 images with similar land cover ratios (e.g., 0.689) can have different scores (e.g., 0.736 and 0.425) due to the appearance of artifacts in the images.

4.4 Change Detection over Time

Table 6: Model score statistics

Descriptive statistics of model score (2016, 2019)

Variable	Mean	Median	S.D.	P25	P75
Model score					
2016	0.1135	0.0981	0.1004	0.0256	0.1805
2019	0.1176	0.0969	0.1130	0.0077	0.1941
Difference					
2019-2016	0.0041	0.0	0.06167	-0.0278	0.0336

The main result of the model was shown with the averaged scores over four years between 2016 and 2019 (Fig. 11). Here, we present the score differences from 2016 to 2019 to understand which regions underwent the highest degree of change (Fig. 12). The descriptive statistics of the model score over 2016 and 2019 are presented in Table 6. Our prediction map shows that North Korea's economic development focused on areas near Pyongyang and the western plain, and economic development in these areas has increased gradually over the four years. The 2016 and 2019 score differences shown in both figures indicate that most of the eastern area had a score of zero.

4.5 Robustness Testing

The robustness test checks the model's ability to adapt to datasets for other countries. We computed the R-squared value and Spearman correlation of predictions for two other countries, Cambodia and Nepal, because past studies on North Korea also compared results to developing countries (e.g., Nepal) and transitioning economies (e.g., Cambodia, China, and Vietnam) in Asia. Cambodia has a similar standard of living in terms of household assets as North Korea (*31*). Nepal has long been classified as a representative low-income Asian country by the World Bank (*32*) and is known to have a GNI per capita close to North Korea (*33*). Nepal also shares trade-isolated geological conditions due to its land-locked location.

A total of six human participants helped build the POGs for these countries. The cluster count was identified in the same way as for North Korea, resulting in 16 for Cambodia and 21 for Nepal. Fig. 13 visualizes the model's prediction scores. The grid-level map distinguishes the most developed capitals of the two nations: Phnom Penh and Kathmandu. The two countries exhibit strikingly different patterns. In Cambodia, development is spread over the expansive agricultural plains along the Tonlé Sap Lake and the Mekong River. In Nepal, development is limited to a few urban areas, whereas the vast mountainous area remains less developed.

We use grid-level Facebook population data as ground truth instead of building footprint data, because they are not available for these countries. The Facebook Connectivity Lab and the Center for International Earth Science Information Network (CIESIN) at Columbia University have jointly released a high-resolution population density map of world regions (c.f., excluding North Korea) based on satellite imagery at `https://data.humdata.org/organization/facebook`. We use these statistics as ground truth for validation.

For the district-level evaluation, we compare our predictions against the population, the numbers of establishments, and the numbers of people engaged in establishments (i.e., employment). The 2011 Nepal Census[9] and National Economic Census 2018[10] provide data for 75 districts of Nepal. Similarly, the Economic Census 2011 and Population 2016 from the Commune Database (CDB) of the Ministry of Planning in Cambodia contain data for its 186 districts.

Fig. 14 shows that our model can successfully be applied to countries outside North Korea. The district-level evaluation with population and economy census data confirms that our method can explain up to 70% of the variation seen in the economic development of these countries. This result reaffirms the potential use of our model in many other developing countries.

[9] https://unstats.un.org/unsd/demographic/sources/census/wphc/Nepal/Nepal-Census-2011-Vol1.pdf

[10] https://nada.cbs.gov.np/index.php/catalog/92

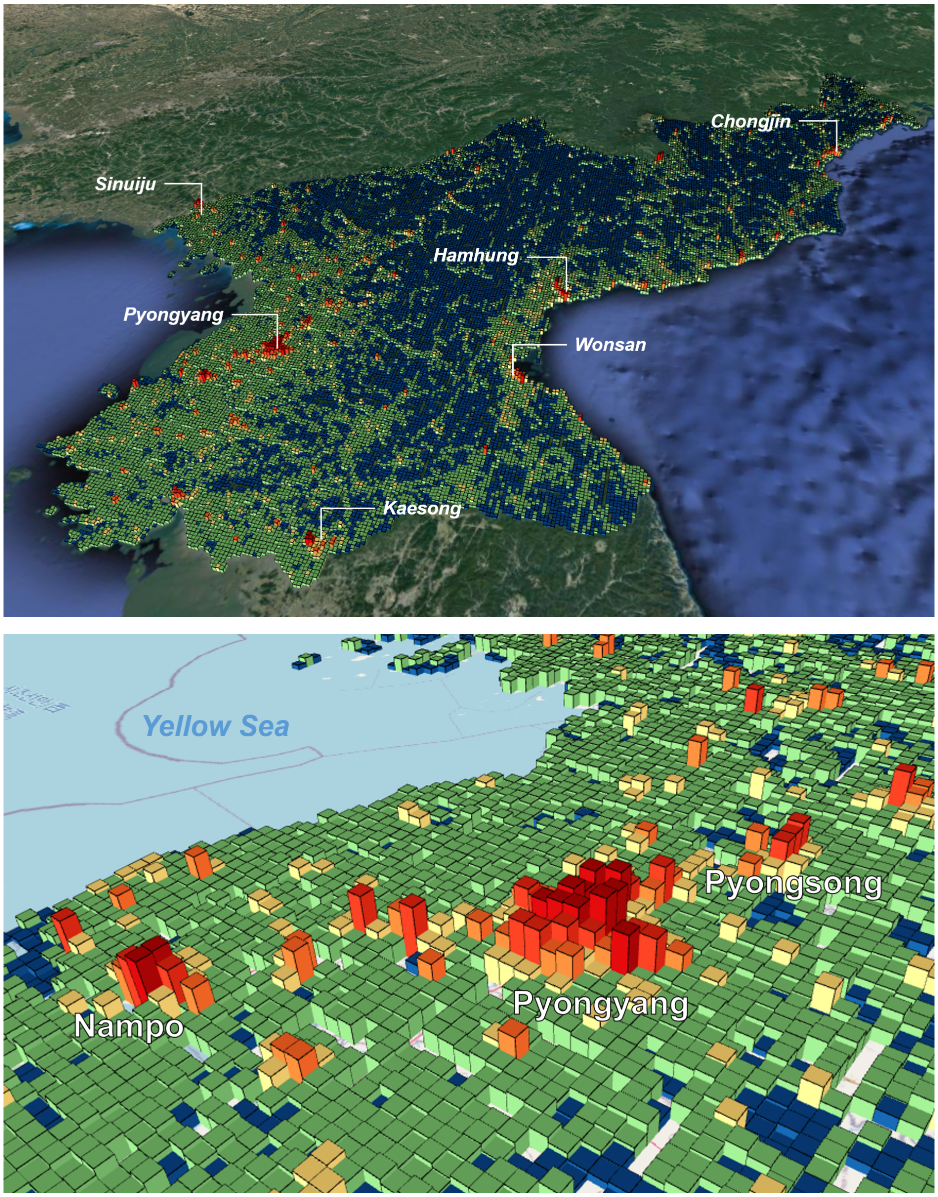


Figure 11: 3D visualization of *siScore* averaged over four years from 2016 to 2019 for North Korea (top) and Pyongyang region (bottom)

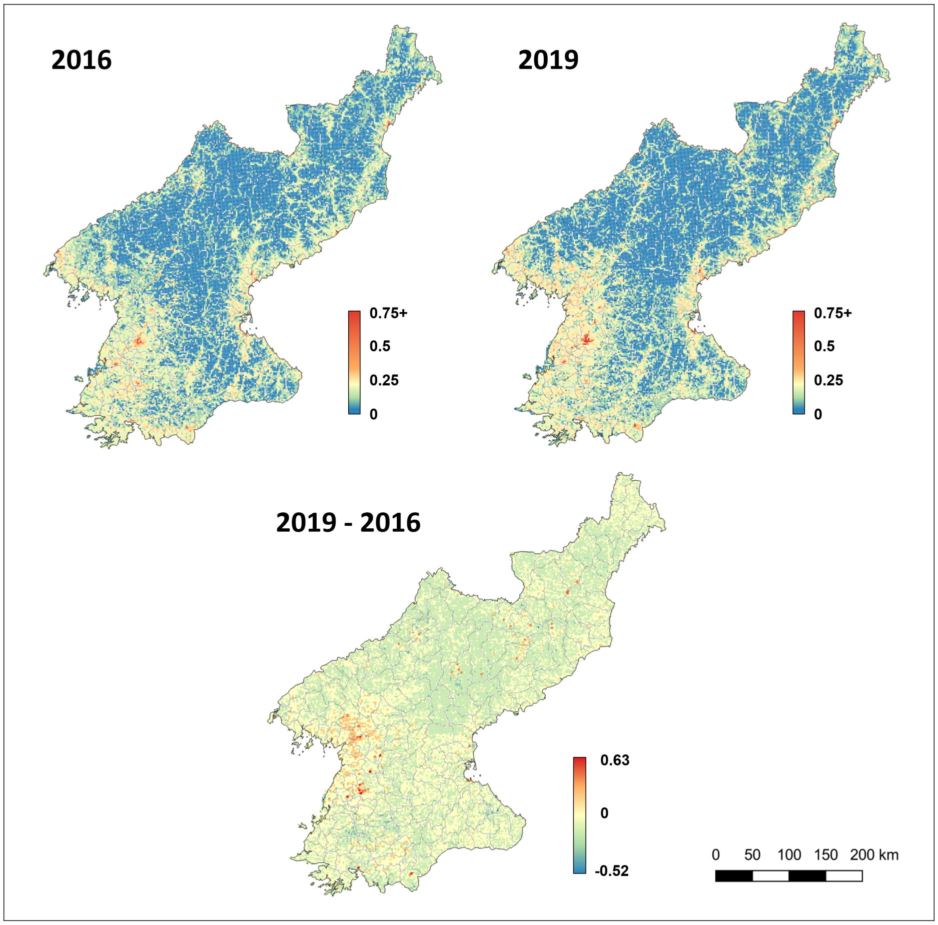


Figure 12: The changes in *siScore* from 2016 to 2019. The bottom image represents the *siScore* difference between 2019 and 2016.

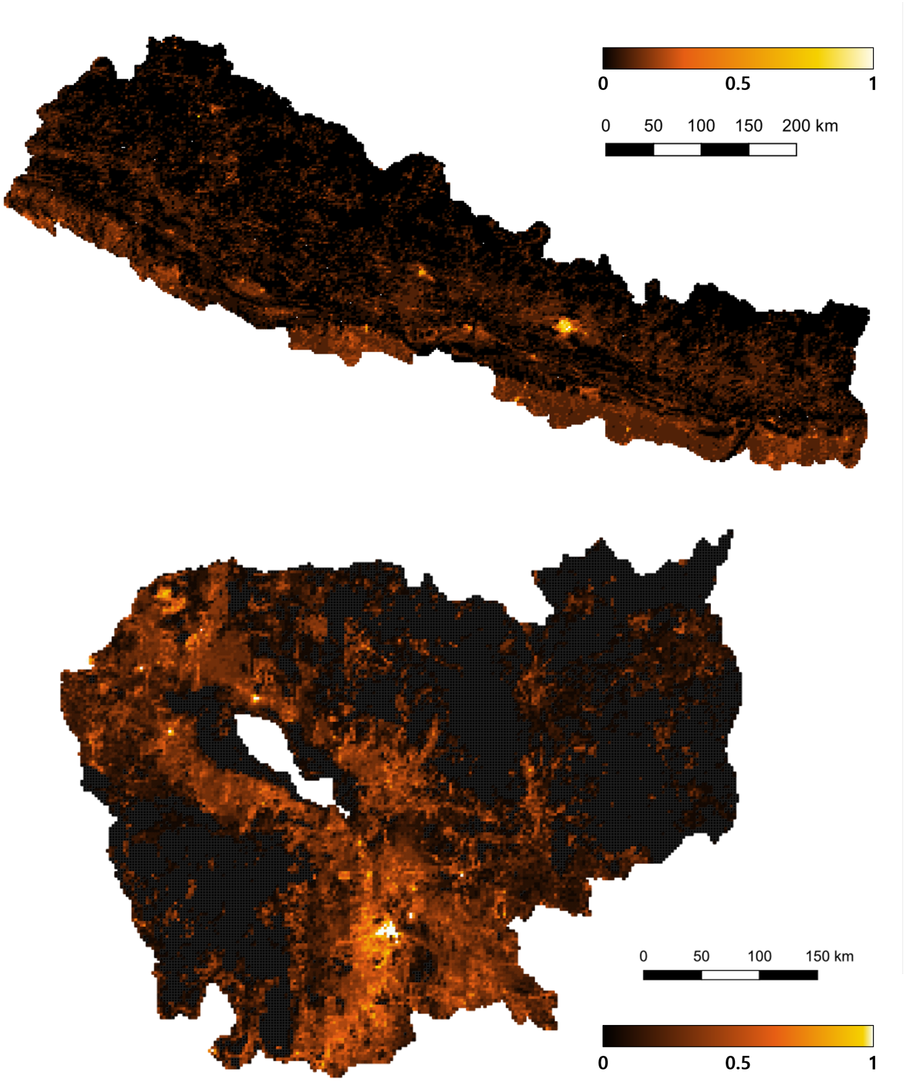


Figure 13: Visualization of the economic development predicted by the human-machine collaboration model for Cambodia (bottom) and Nepal (above). Predictions are based on images from 2016 to 2019.

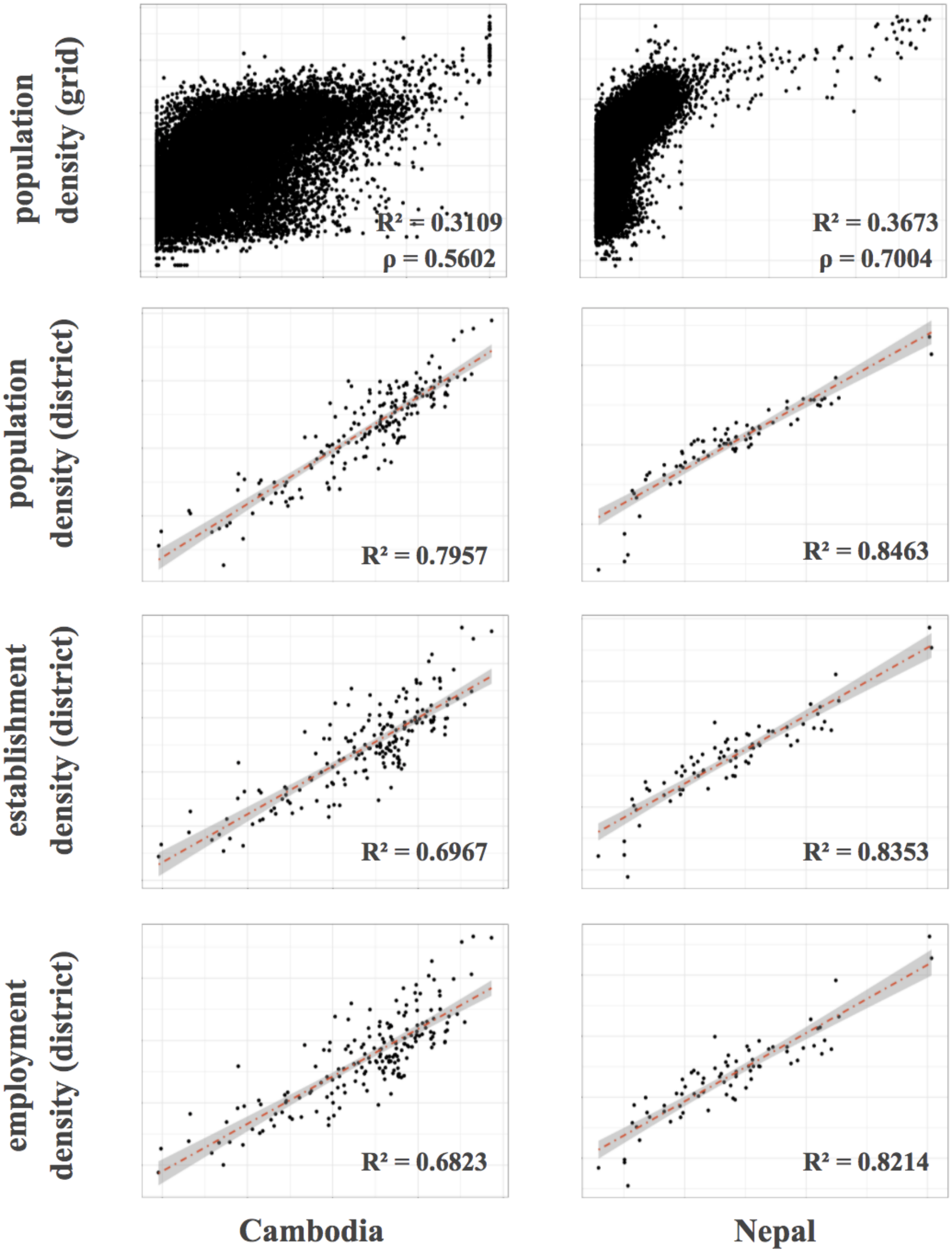


Figure 14: Robustness testing for Cambodia and Nepal. The first row shows grid-level evaluation with Facebook population data. The remaining rows examine district-level evaluation with population density and the numbers of establishments and employment from censuses.