Gridded population projection of South Korea using ensemble machine learning (2022–2050)

1. Background

South Korea, facing the challenges of the world's lowest birthrate and a rapidly aging population, requires comprehensive demographic data for effective policy responses. Gridded population projections offer advantages such as temporal consistency and flexibility for integrated analysis with other fields, including public health, regional development, and climate change adaptation. This research presents the first grid cell-level population projection data for the entire country.

2. Data and Methods

We developed a machine learning (ML)-based dasymetric mapping approach to disaggregate municipal-level population projections to a 500m grid. The projections were sourced from Statistics Korea, Korea University (Kim and Kim, 2020; Kim et al., 2021), and Shared Socioeconomic Pathways 2 (SSP2; Wang et al., 2022), hereafter referred to as the Statistics Korea, Korea University, and SSP2 projections.

Recognizing significant spatial heterogeneity, we separated municipal regions into urban and rural areas based on legal definitions before training the ML models. We trained XGBoost and CatBoost models, with Ordinary Least Squares (OLS) as a baseline, resulting in six models. Training data included population density (dependent variable) and features such as eight land cover classes, distance to primary roads, nighttime lights, elevation, and slope, all preprocessed at the census tract level. These features were also prepared at the 500m grid cell level and input into the trained ML models to estimate population density at the grid cell level, which was then used as a weight to downscale municipal-level projections using the following equation:

$$P_i = \frac{D_i}{\sum_{i \in s}^{n} D_i} \bullet P_s$$

where P_i represents the projected population of grid cell i, D_i indicates the estimated population density of grid cell i, P_s refers to the population at the municipal level region s from the population projections, n denotes the number of grid cells belonging to each municipal-level region.

Model accuracy was evaluated using R-squared and Root Mean Squared Error (RMSE). SHapley Additive exPlanations (SHAP) analysis was conducted to assess variable importance and the

relationship between features and the target variable. Without SHAP, ML methods would remain a black box.

Finally, we validated the gridded population data by comparing it with Seoul's hourly population data at 25 districts derived from real-time mobile phone signals. We compared daytime, daily average, and nighttime populations to evaluate how well the urban model captured different temporal patterns.

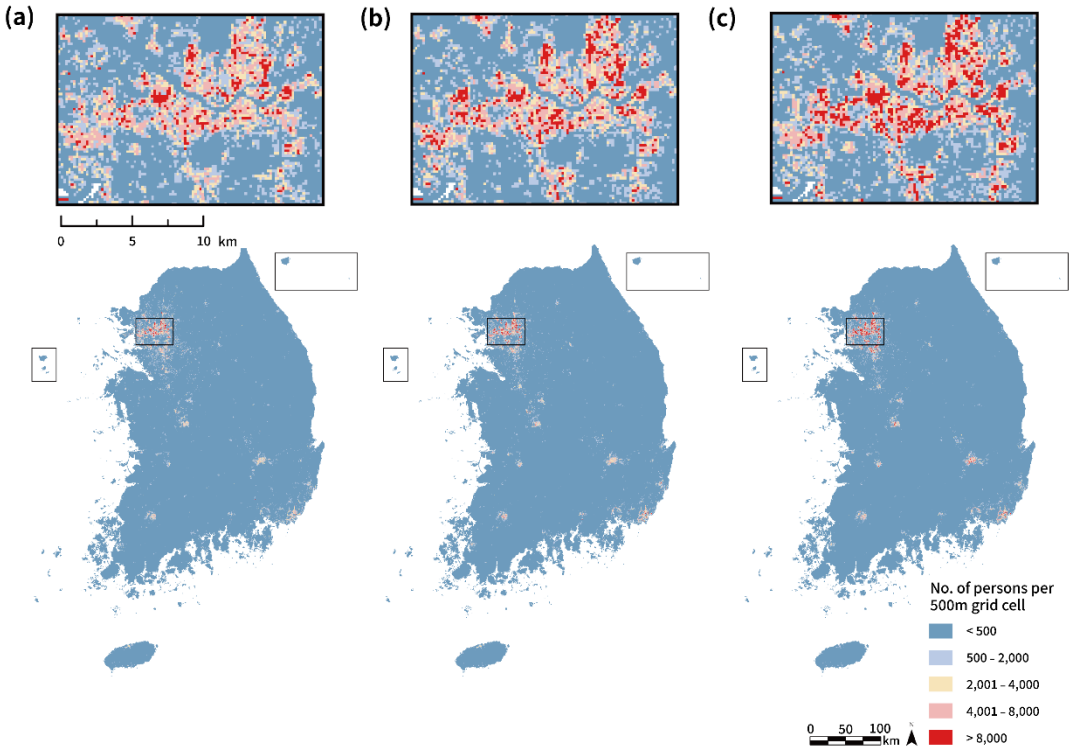
3. Results

Model evaluation was based on R-squared and RMSE values for test data. CatBoost showed the best performance for both the urban and non-urban models. All three projections indicated that most of the population is concentrated in major metropolitan regions, especially Seoul (Figure 1). Compared to the Statistics Korea and Korea University projections, the SSP2 projection showed the highest population.

SHAP analysis revealed that, for the urban model, Residential area was the most important feature, while Wetland had the lowest importance (Figure 2). Residential area and Nighttime lights had positive correlations with population density, while Non-residential built-up area, Water body, and Forest land showed negative relationships. For the non-urban model, Residential area was again most important, while Slope had minimal impact. Notably, the contribution of features differed between urban and non-urban models.

To validate the urban model, we calculated MAPEs for the three gridded population projections using both ML and OLS models (Figure 3). The projections most closely matched the nighttime population. For nighttime, OLS performed slightly better, but ML models had lower MAPEs for daily average and daytime populations, likely because they captured daytime activity patterns in non-residential areas (e.g., industrial and commercial zones) as well as nighttime patterns from census data.

ML-based dasymetric mapping significantly improves population downscaling in South Korea, capturing both spatial heterogeneity and temporal patterns. Traditional planning and vulnerability assessments have relied on coarse administrative-level data, risking misallocation of resources and underestimation of local risks. As demographic and environmental conditions change, high-resolution population projections that preserve demographic attributes (from the Statistics Korea and Korea University projections) or align with climate scenarios (from the SSP2 projection) will be essential for effective adaptation and preparedness.



(a) (b)

Figure 1. Gridded population using CatBoost in (a) 2022 and (b) 2050

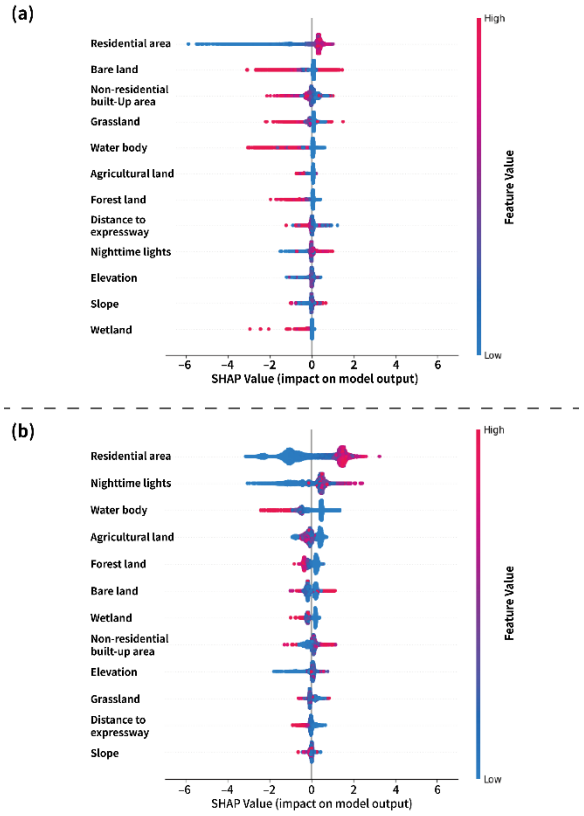


Figure 2. SHAP summary plots of (a) urban model and (b) non-urban model by CatBoost.

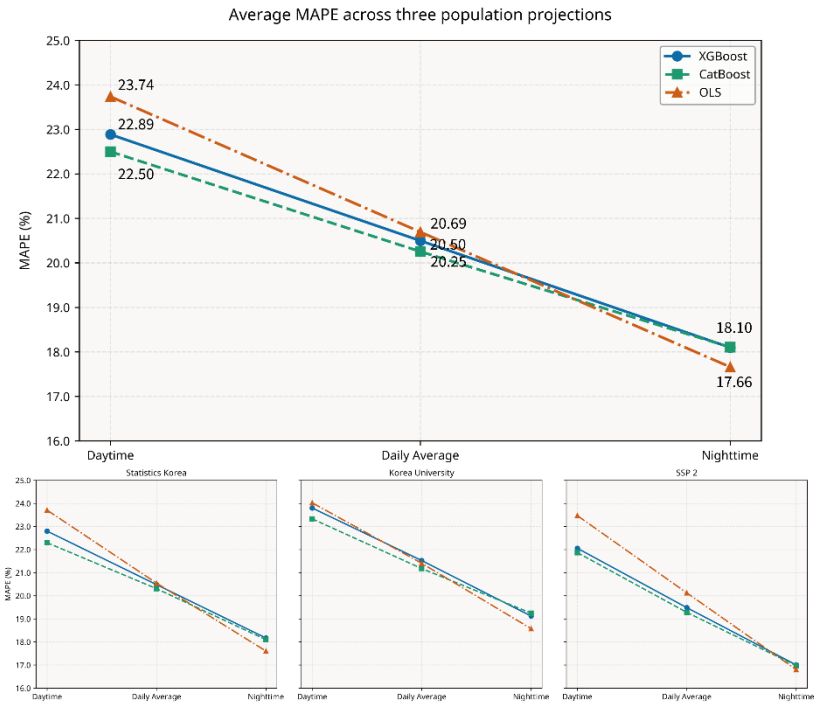


Figure 3. MAPEs of the gridded population projections of Statistics Korea, Korea University, and SSP2.