Title

Gridded datasets for Japanese total, male, and female population over 2001-2020

Authors

Chao Li[1], Shunsuke Managi*[1]

Affiliations

[1] Urban Institute & School of Engineering, Kyushu University, Japan
* Correspondent to: Shunsuke Managi, managi@doc.kyushu-u.ac.jp, Kyushu University 744 Motooka, Nishi-ku, Fukuoka 819-0395 Japan

Abstract

Japan is one of the highly urbanized and severely aging societies. In an aging society, chronic disease and disability are prevalent, and the populaiton is sensitive to environmental issues and climate change. To detect the impacts of population changes, formulate the population and public health policies, and assist environmental applications, the high-resolution and accurate gridded population dataset is strongly desired. To provide basic data for these studies, we create an open access annual dataset of the total, male, and female population counts in each grid at a 500m resolution from 2001 to 2020. The yearly population dataset is based on the 4th mesh data from the Statistics Bureau of Japan to make it easy to use. The dataset is demonstrated here, along with the descriptions of the data and methods used in the fitting, cross-validation, and prediction processes.

Background & Summary

An increasing number of open access gridded datasets are available, which provides more possibility for complex spatial analyses and, in turn, leads to a development in spatial analysis technologies [1-3]. Since more high-resolution remote sensing data [4] and efficient machine learning packages [5] are publicly available, the spatial and temporal resolutions and accuracy of gridded data continue increasing. Additionally, with the development of computer technologies in both software and hardware, big data analyses have become accessible to most researchers based on high-performance computers, which also causes the gridded data to become available in most fields.

Japan has a population of approximately 125 million and is one of the highly urbanized and severely aging societies [6-8]. Aging society is a fatal issue for all developed countries and also threatens some developing countries. In an aging society, chronic disease and disability are prevalent [9], and the population is sensitive to environmental issues and climate change. Furthermore, in the next decades, the population will keep decreasing in Japan [6], and numerous facilities will be discarded. To formulate effective population, public health, and land use policy, high-resolution and accurate population data are desired. In fact, accurate gridded human population data are vital in environmental, public health, economic, urban planning, and policy analyses [3,10,11]. For example, the estimation of the negative impacts of various pollution [12-14], diseases prevalence and mortality distribution, inclusive wealth estimation [15], land use policy [16], among others, desire and rely on the high-resolution and accurate gridded population distribution.

Since 2000, the Japanese government has offered the gridded population data per five years based on nationwide surveys. These publicly available official datasets make high-resolution and accurate predictions possible. Currently, the famous WorldPop Project (www.worldpop.org) also provides high-resolution population data, including population density at a 1km resolution and population count at 1km and 100m resolution. However, their estimation used the aggregated data and the random forest-based dasymetric mapping approach [3]. Redistributions of aggregated data to gridded data cause some residuals, which undoubtedly reduce the accuracy of the estimation. Using the gridded data in several years of the research period from the Japanese government could avoid these residuals. Furthermore, our dataset is based on the grids divided by the Statistics Bureau of Japan, which makes the data easy to connect with other datasets from the Japanese government without further raster resampling or reprojecting. Therefore, in terms of Japan's population distribution, our dataset exceeds other datasets in the accuracy and ease to use.

This paper presents a gridded dataset including total, male, and female population distribution in Japan from 2001 to 2020, with cross-validation accuracy scores of 92.00%, 91.90%, and 92.00%, respectively. The dataset is stored in a polygon shape file with a resolution of 500m in the standard WGS84 coordinate system.

Remote sensing data from the National Aeronautics and Space Administration (NASA) and the Japan Aerospace Exploration Agency (JAXA), and statistical spatial data from the Japanese government are employed to estimate our dataset. The spatial resolutions of some remote sensing data are mainly 500m, while others are 30m, 1km, or 0.1-arc-degree. All remote sensing data are raster data, which are resampled to a roughly 500m resolution, reprojected to the standard WGS84 coordinate system, and extracted to a spatial point data frame for further analyses. The statistical spatial data from the Japanese government are vector data, which are spatially joined to the spatial point data frame by returning the distance. In total, 55 features are used to estimate the gridded population. The schematic overview of the workflow is shown in **Figure 1**.

Previous studies in various fields, including human well-being[16], environmental impacts [12,13], diseases [17,18], among others, mainly use aggregated population data in their analysis, such as city-level or prefecture-level. However, with accurately high-resolution gridded data, the spatial heterogeneity in those tropics can be more deeply detected. Furthermore, the data-gap-free annual dataset over 2001-2020 provides more possibility for other potential research to detect the time-fixed effects on population distribution within each mesh. Our dataset could also be used as the base data to predict other population-related gridded datasets, such as disease distributions, income distributions, transportation densities, among others.

Methods

In this section, we demonstrate the dataset-producing process. The data sources, data gathering, and further data processing are reported. Three variables, the logarithms of the total, male, and female population counts in each mesh, are taken as the output variables in the machine learning models. Additionally, the random forest models employ 55 features from the various data sources.

Materials

Japan Regional Mesh and Population Data

Regional mesh data are a series of grids divided by the Statistics Bureau of Japan (https://www.stat.go.jp/english/data/mesh/05.html). There are six levels of mesh data at different spatial resolutions. The resolutions of 1st to 6th mesh data are 80km, 10km, 1km, 500m, 250m, and 125m, respectively. Because the resolution of remote sensing data is mainly 500m, the 4th mesh data at a 500m resolution are the best choice (https://www.e-stat.go.jp/gis/statmap-search?page=1&type=2&aggregateUnitForBoundary=H&coordsys=1&format=shape). To further processing, we reproject the polygon shape file to the standard WGS84 coordinate system. To extract data from remote sensing raster and build distance to features of interest datasets, we convert the polygon shape file to the point shape file by using centroids of each grid.

The Japanese government conducts a nationwide survey every five years to obtain population distribution. From 2001 to 2020, there were four surveys in 2005, 2010, 2015, and 2020. The 4th mesh population data of 2005, 2010, 2015, and 2020 are publicly accessible. The population data include three variables: the total population counts in each mesh, the female population counts in each mesh, and the male population counts in each mesh, respectively. Seemingly, there are 5th mesh population data of 2010 and 2015 at a 250m resolution, but the data are only available in metropolitan areas and missing in the low-population-density areas. Therefore, the 4th mesh population data in 2005, 2010, 2015, and 2020 are the best choice to be regarded as the output in the models.

The population counts of each mesh range from 0 to over 10,000. If the population counts are directly used as the output variables, the large standard deviation of the output variables might reduce the model's accuracy. Hence, a link function to squeeze the range of the output variables is needed. Logarithmization is an effective method. The following equation is used:

$$LPC_i = \ln(PC_i + 1) \quad (1)$$

where LPC_i is the logarithm of population count in mesh i, and PC_i is the population count in mesh i.

Land Cover Types and Distances to Certain Land Types

Land cover data are provided by NASA. The MCD12Q1 is a moderate resolution imaging spectroradiometer (MODIS) dataset, which includes yearly global land cover data at a 500m resolution from 2001 to 2020 based on the observations of the MODIS satellites [19]. The MCD12Q1 has five different land cover classification schemes, including InternationalGeosphere-BiosphereProgramme (IGBP), UniversityofMaryland(UMD), Leaf Area Index (LAI), BIOME-Biogeochemical Cycles (BGC), and PlantFunctionalTypes(PFT). The IGBP has the most classification, followed by the UMD. Compared with the UMD, the IGBP has one more land type, permanent snow and ice, but this land type does not exist in Japan. Therefore, we use the UMD classification, which contains 16 land types: water bodies, evergreen needle leaf forests, evergreen broad leaf forests, deciduous needle leaf forests, deciduous broad leaf forests, mixed forests, closed shrublands, open shrublands, woody savannas, savannas, grasslands, permanent wetlands, croplands, urban and built-up lands, cropland/natural vegetation mosaics, and non-vegetated lands. The raw resolution of the MCD12Q1 is 463.312m, and the projection is the MODIS Sinusoidal coordinate system [19]. We reproject the data to the WGS84 coordinate system and resample the data to a 0.004-arc-degree resolution (roughly 500m) by the mode method. The point shape file and yearly land cover data from MCD12Q1 are employed to extract land type data. The extracted land type data is a categorical variable, ranging from 0 to 15. We use the one-hot vector method to convert them into a data

frame with 16 dummy variables. Furthermore, we calculate the nearest distances of each point to all land types, which are the other 16 variables.

Nighttime Light Data

The nighttime light (NTL) satellite data that report the light intensity have been widely applied to indicate human activity and development intensity [20,21]. The connection between gross domestic product (GDP) and NTL is significant, and NTL is usually used to represent the GDP in developing countries [22,23]. Previous studies indicate that NTL is associated with population density [24,25]. To accurately estimate the population in each mesh, we put the NTL variable into the models. Currently, two NTL datasets are publicly available and widely used: the Defense Meteorological Satellite Program Operational Linescan System (DMSP-OLS) and Suomi National Polar-orbiting Partnership Visible Infrared Imaging Radiometer Suite (NPP-VIIRS), respectively [26,27]. The DMSP-OLS covers 2000-2012, and NPP-VIIRS data are available from 2012. However, because they are from different sensors, their calibrations are not consistent. Chen et al. created an extended yearly NPP-VIIRS-like NTL data set from 2000-2018, and the data in 2019 and 2020 are also stored in their data archive [28]. The NTL dataset is in the WGS84 coordinate system at a spatial resolution of 15 arcsec (roughly 0.004 arc degree). The point shape file could directly extract the NTL data from the raster in the data archive.

Net Primary Production

Yearly net primary production (NPP) data are also from NASA MODIS satellites. NPP demonstrates the solar energy captured and stored by the planes through photosynthesis [29]. The NPP is an essential confounder of population distribution because human populations depend on NPP "imports" [30]. MOD17A3HGF and MYD17A3HGF are yearly global NPP products at a spatial resolution of 500m based on MODIS Terra's and Aqua's observation, respectively. The projection of the raw MOD17A3HGF and MYD17A3HGF data is the MODIS Sinusoidal coordinate system. To make the data consistent with our project, we reproject them into the WGS84 coordinate system and use the average method to resample them into a 0.004-arc-degree resolution.

Temperature and Precipitation

The meteorological variables, average temperature and precipitation, are employed in previous population distribution studies [3,31]. The temperature data are from NASA MOD11A2 and MYD11A2 at a 1km spatial resolution. MOD11A2 and MYD11A2 data include 8-day average daytime and nighttime temperature. To make the temporal resolution concordant with the output variable, we average the 8-day data annually. The temperature difference is also an important indicator of livability. Hence, we put the annual average temperature and standard deviation of temperature into the models. Since the MODIS data all have the MODIS Sinusoidal coordinate system, we must reproject them into the WGS84 coordinate system. We directly use the point shape file to extract the temperature data. Although the resolutions of the temperature data and the point shape file are inconsistent, the points located in the same grid will obtain the same values. One grid covers at most four points, so the data are valid at a large scale. NASA global precipitation measurement provides monthly precipitation data at a 0.1-arc-degree resolution, included in the product GPM_3IMERGM. We use the same method to extract monthly precipitation data as the temperature extraction. Although the spatial resolution of the precipitation dataset is insufficient to some degree, they are still better than aggregated data, such as city-level or prefecture-level data in Japan.

Elevation and Slope

The JAXA published the global elevation data at a 30m resolution in 2015. We assume that the elevations in each mesh in Japan have remained constant in the past 20 years. First, we resample the elevation data to the 0.04-arc-degree resolution by the averaging method.

Second, we use the 0.04-arc-degree raster dataset to generate the slope raster at a 0.04-arc-degree resolution. Then, we extract the elevation and slope data using the point shape file.

Distance to Features of Interest

The Japanese government provided the shape files of several features of interest, including rivers, coastal lines, high population zones, railways, railway stations, entertainment facilities, government branches, police stations, fireman stations, schools, hospitals, post offices, and disabled or senior support facilities. Although these data are not updated yearly, we assume that they are the same as the nearest data-available year. We use the point shape file to calculate the distance to the nearest features of interest. It must be mentioned that the road density data are not a line shape file but 3rd mesh data at a 1km resolution. Hence, we can only know the road density of each mesh but cannot obtain the distances between roads and each mesh point.

Location Information

The latitudes and longitudes of the mesh centroids are put into the model. Different from the traditional regression methods, directly using location features in the analyses is valid. Random forests divide a specific feature range binarily several times. In other words, this feature range is separated into several intervals, and within an interval, the output variable values of each observation are similar to some degree. If the features are location information, the dataset is divided into numerous clusters based on the spatial contexts. These spatial clusters would improve the accuracy of estimation by grasping the spatial variability.

Data Summary

Table 1 summarizes variable name, processing approach, data source, time stamps, and other necessary information.

Machine Learning Model

Random forest is taken as the algorithm to predict the gridded population dataset in our study because it is good at grasping the non-parametric relationship between the output variable and features [5,32] and is widely used in previous population prediction studies [3,33].

Decision Tree

A decision tree is the basic element of the random forest method. The decision tree predicts the output variable based on a series of binary judgments [5,32]. The binary splitting makes the decision trees extremely efficient in grasping the nonlinear relationship. When a decision tree analyses the continuous variable, it will judge a feature several times to break the feature range into several ranges. For example, the first judgment in a decision tree might be whether the average temperature is higher than 25℃; if true, the second one might be whether the average temperature is higher than 27℃, while if false, the second one might be whether the average temperature is higher than 23℃. Based on these judgments, the temperature range is divided into $(-\infty, 23℃]$, $(23℃, 25℃]$, $(25℃, 27℃]$, and $(27℃, +\infty)$. The rules of each judgment and feature range splits are critical to obtaining high accuracy. The residual sum of squares (RSS) is the widely used accuracy indicator, and the machines "learn" the optimal rules of judgment and split strategies to minimize the RSS. The greedy split approach is applied for training the individual regression trees to minimize the RSS [34]:

$$RSS = \sum_{l \in leaves} \sum_{i \in C_l} (y_i - \bar{y}_{C_l})^2 \quad (2)$$

where l is a leaf, C_l is the cases in leaf l, y_i is the observed value and $\bar{y}_{C_l}$ is the average observed value in leaf l. In this approach, the splits will continue as long as RSS can decrease.

However, the price of the minimized RSS is high variance, i.e., the unlimited greedy approach leads to over-fitting. Two sophisticated rules are designed to avoid over-fitting. We set the thresholds of RSS and the remaining case numbers in the end leaves [34]. If the RSS or the remaining case numbers in end leaves is smaller than the thresholds, the further split in a certain feature stops.

Random Forest

Decision trees always come with over-fitting or low accuracy. As we mentioned above, we set the thresholds in the greedy split approach to avoid over-fitting, but these limitations increase the RSS. Decision trees' ability is far from the big data prediction requirements, based on the balance between accuracy and over-fitting that is the bias-variance tradeoff technologically. To improve the prediction ability and capture complex relationships like decision trees, the random forest built on a bundle of decision trees is created [5,35].

The random forest first resamples a large number of subdatasets; second builds hundreds of decision trees based on the subdatasets and lets them individually predict their results; and third averages all the individual tree's results. Bootstrapping is the sampling technology to randomly sample subdatasets with replacements, which is the vital part of completing the first two steps of the random forest. The number of bootstrapped subdatasets is the same as the tree number in the random forest. In our analysis, the tree number is set to be 1,000, which is enough to obtain a reliable result [36]. Therefore, there will be 1,000 bootstrapped subdatasets to train the random forest. The sizes of each subdataset are 2/3 of the total sample size. To improve the heterogeneity among the trees, the subdatasets just contain partial features rather than all features in the total dataset. The default number of selected features in the subdatasets is one-third of the total number of features in the total dataset [5]. Each subdataset is used to train a single decision tree. In the third step, the random forest aggregates the predicted values from each tree by using the averaging method to predict the output variable. Since the random forest used bootstrapping and aggregating technologies, the full process of the model training is terminologically called "bagging". Because each tree only uses approximately 2/3 data in the bagging, the remained data are called out-of-bagging (OOB) data. OOB error is the residuals of OOB data from the model trained by the bagging data [5]. The OOB error is the cross-validation result in a way, but the estimation process has been assembled in the random forest algorithm. Generally, the OOB error rate represents the over-fitting degree of a random forest. OOB score is the ratio of the variance grasped by the model, so the sum of the OOB error rate and OOB score always equals 100%.

Cross-Validation

Although the effects of the OOB scores in random forests are similar to cross-validation, there are still some differences. For a single decision tree, the OOB score is estimated using "new" data, but for the entire model, all data are used to train the model. In the cross-validation, the total dataset is randomly divided into training and testing datasets according to the ratio of 8 to 2. The training dataset is only used to train the model, while the testing dataset is employed to individually examine the model's reliability. In fact, this process is closer to real-world situations. Furthermore, our model is to predict annual population distribution based on several-year data, so the model must be reliable temporally. To check the method's temporal reliability, we execute temporal cross-validation. Three-year data are used to train the model, while the remaining one-year data are employed to test the reliability. Since we have three-year data, the temporal cross-validation would be performed three times.

Statistical Indicator

Several statistical indicators, including R^2, root mean square error (RMSE), mean absolute error (MAE), and regression coefficients between observed and predicted values are widely used to

indicate the accuracy of models. R^2 is a critical statistical indicator describing the goodness of fit, and in this study, it is taken as the accuracy score. The R^2 calculation is as follows:

$$R^2 = 1 - \frac{\sum_{k=1}^{n}(OV_k - PV_k)^2}{\sum_{k=1}^{n}(OV_k - \overline{OV})^2} \quad (3)$$

where n represents the record number of the dataset, OV_k represents the kth record of the observed population data in a certain mesh, PV_k represents the kth record of the predicted population data in a certain mesh, and $\overline{OV}$ represents the mean of the observed population data in a certain mesh. It must be noted that the record numbers n vary because the dataset are different in the fitting process, the 7:3 cross-validation, and the temporal cross-validation. The RMSE is imputed as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(OV_k - PV_k)^2} \quad (4)$$

The MAE is calculated as follows:

$$MAE = \frac{1}{n}\sum_{k=1}^{n}|OV_k - PV_k| \quad (5)$$

In the analysis, the RMSE and MAE should be low. Furthermore, the regression coefficients are estimated as follows:

$$PV_k = \alpha + \beta OV_k + \varepsilon_k \quad (6)$$

where α is the intercept in the regression and the ideal value of α is 0, β is the slope and its ideal value is 1, and δ_k is a random error term.

Data Records

The datasets are based on the 4th Japanese regional mesh data with a resolution of 500m in the standard WGS84 coordinate system. The data are provided in shapefile format. Population data are stored as attributes of each polygon element. To make usage convenient, we keep the mesh id number in the dataset following the rules from the Statistics Bureau of Japan (http://data.e-stat.go.jp/lodw/en/provdata/lodRegion). The attribute name of the mesh id number is "meshID". In total, 1,553,024 mesh grid data are predicted. **Figures 2 – 4** demonstrate the total, male, and female population distributions in Japan from 2001 to 2020.

The total population data are stored in 20 attributes, which are named in "X0000" style. The "0000" of "X0000" represents the year number. For example, the attribute "X2001" reports the total population in the mesh. The dataset also keeps the direct output from the random forest, the logarithms of the total population. The attributes of the logarithms of the total population in each year are named in "X0000_log" style. The "0000" of "X0000_log" also stands for the year number. The attributes of the female population, the logarithm of the female population, the male population, and the logarithm of the male population are written in "X0000_fema", "X0000_fe_l", "X0000_male", and "X0000_ma_l", respectively.

Technical Validation

The goodness of fit of the fitting model for the logarithm of the total population is 98.65% (**Table 2**). The MAE and RMSE of this fitting model are 0.13 and 0.24. The regression intercept and slope are 0.06 and 0.95, respectively. Because the output variable is the logarithm of population count, the MAE and RMSE from the model are difficult to be understood. Therefore, we convert the logarithms into population counts and recalculate the indicators again. In our study, the R^2s of the fitting model and cross-validations all increase. According to **Figure 5.a**, the model tends to underestimate the value since the blue line (linear fit line) is always under the red one (1:1 line). After the data transformation, the residuals become

relatively larger, as shown in **Figure 5.b**, but the accuracy is still around 97.99%. **Figure 5.c** and **d** illustrate the result of the 8:2 cross-validation for the total population. The shapes of scatter plots are similar to the figures of the fitting model (**Figure 5.a** and **b**), but residuals are larger. The accuracy of the cross-validation for the total population after data transformation is 92.00%. The MAE and RMSE after data transformation are 9.32 capita/mesh and 44.97 capita/mesh, respectively, while the mean of observed total population count data is 81.95 capita/mesh. The accuracy scores of the fitting model for the logarithm of the male and female populations are 98.57% and 98.70% (**Table 2**), similar to the fitting model for the logarithm of the total population. **Figures 6** and **7** show that the model situations for male and female populations are the same as the model of the total population. The accuracy scores of the cross-validation for the male and female population after data transformation are 91.90% and 92.00%, respectively. Furthermore, the OOB scores of the three fitting models are 90.02%, 90.45% and 90.41%, respectively, which are close to the results of the cross-validations. To sum up, although it is relatively over-fitting, the models are reliable because the differences among accuracy scores of the fitting models and cross-validations and OOC scores are tiny.

The accuracy scores of cross-validation are regarded as actual accuracy scores since the fitting models are overfitting. The accuracy scores of the cross-validations of the models for the logarithms of total, male, and female populations using the randomly divided dataset according to the ratio of 8 to 2 are 88.43%, 88.95%, and 88.90%, respectively (**Table 2**). After data transformation from the logarithm to the count, the accuracy scores are 92.00% for the total population, 91.90% for the male population, and 92.00% for the female population. Although the accuracy scores are relatively lower than the fitting models', the values are still excellent. We compare our results with the widely used dataset from WorldPop. We use population density adjusted by the corresponding official United Nations population estimates at a 1km resolution. The accuracy score, MAE, RMSE, intercept, and slope of their prediction of the total population in 2005, 2010, 2015, and 2020 are 74.52%, 57.77 capita/mesh, 194.81 capita/mesh, 26.08, and 0.73, respectively, while the mean of observed data is 84.60 capita/mesh. Our model's are 92.00%, 23.73 capita/mesh, 107.36 capita/mesh, -2.17, and 0.84, respectively, while the mean of observed data is 81.95 capita/mesh. Obviously, our model for the total population is better.

The temporal reliabilities of the model for the logarithms of the total, female, and male populations are 88.32%, 88.80%, and 88.82%, respectively, which are the mean values of the three temporal cross-validation accuracy scores (**Table 2**). **Figures 8 – 10** demonstrate the results of three temporal cross-validations that take the total, male, and female as the output variables. After data transformation from logarithms to counts, the temporal reliabilities of the three models increase to 87.65%, 87.60%, and 88.00%, respectively. Based on the high temporal reliabilities, the model predictions of the population in different years are also reliable.

Code Availability

The fully reproducible codes are publicly available at https://github.com/MichaelChaoLi-cpu/JapanPop.git. Data are all from publicly available data sources.

Acknowledgements

This research was supported by the following funding agencies: JSPS KAKENHI (Grant No. JP20H00648), the Environment Research and Technology Development Fund of the

Environmental Restoration and Conservation Agency of Japan (Grant No. JPMEERF20201001), and also JST SPRING (Grant No. JPMJSP2136).

Author contributions

Chao Li: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Original Draft, Visualization **Shunsuke Managi**: Review & Editing, Supervision, Project administration, Funding acquisition

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Figures

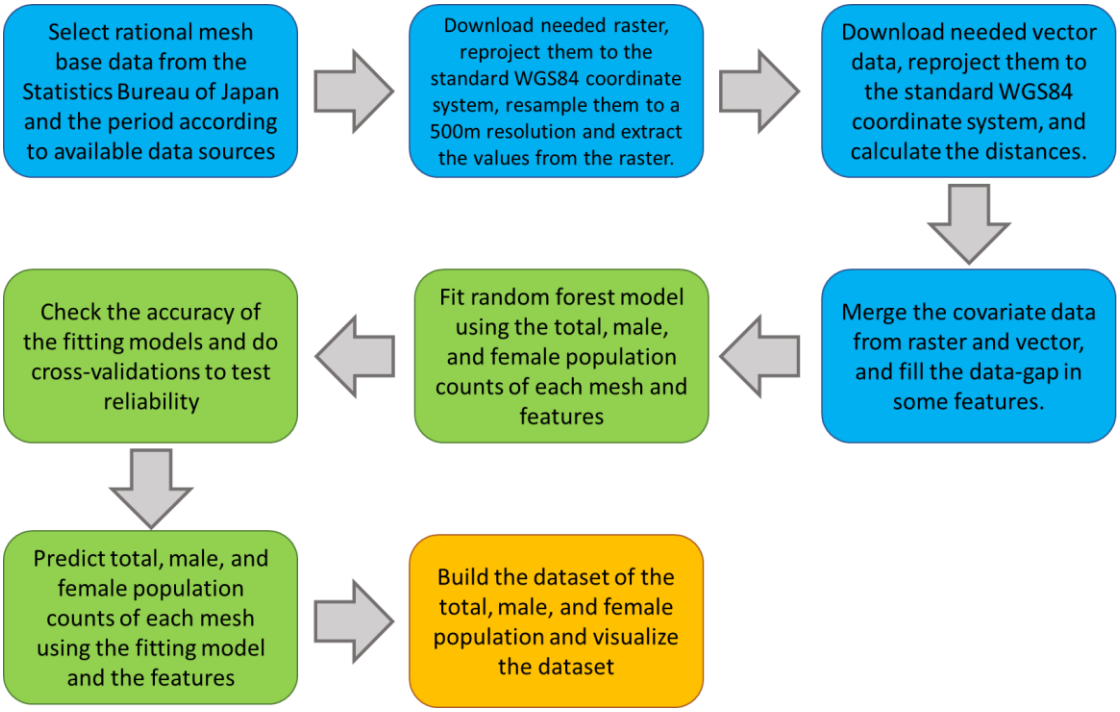
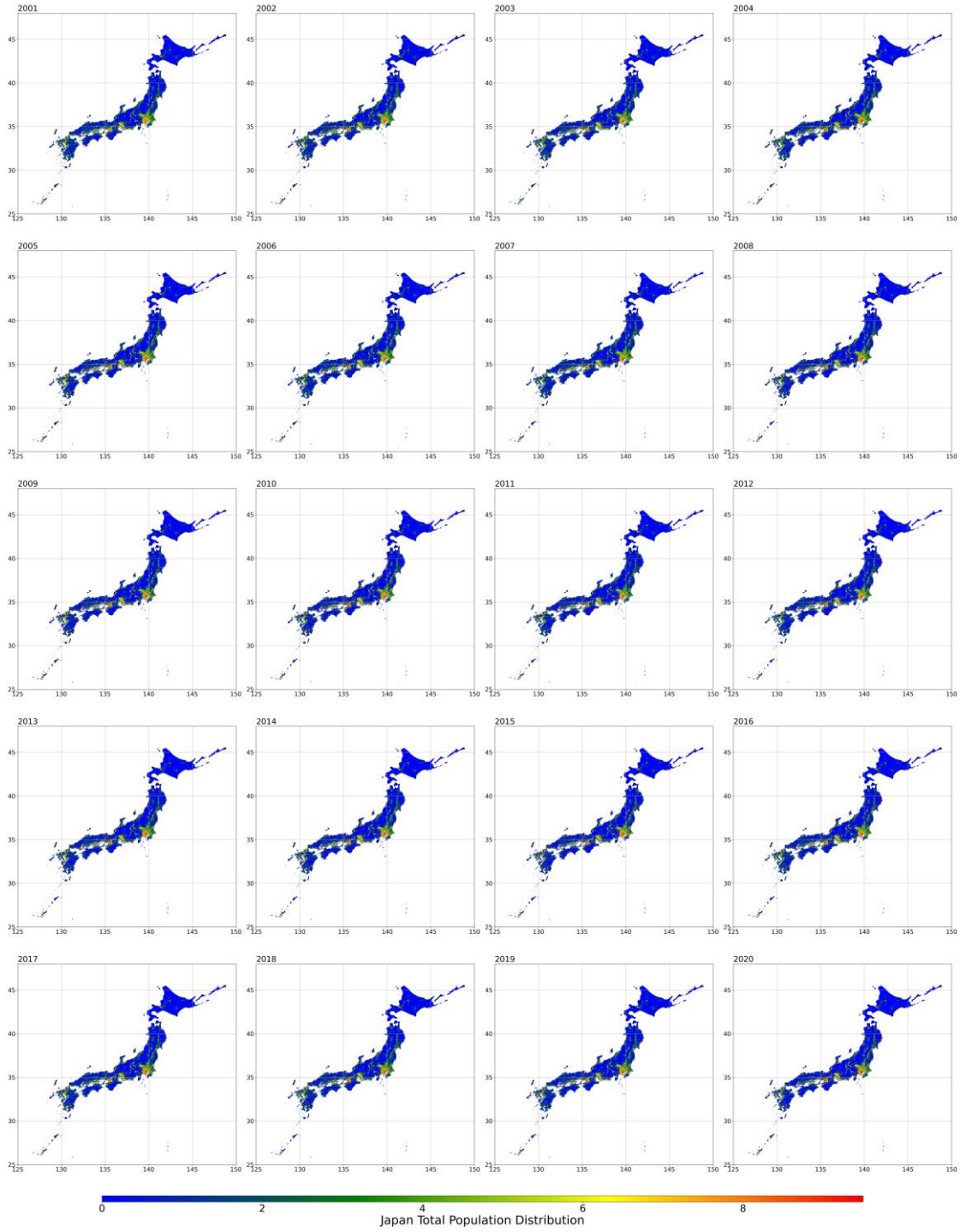
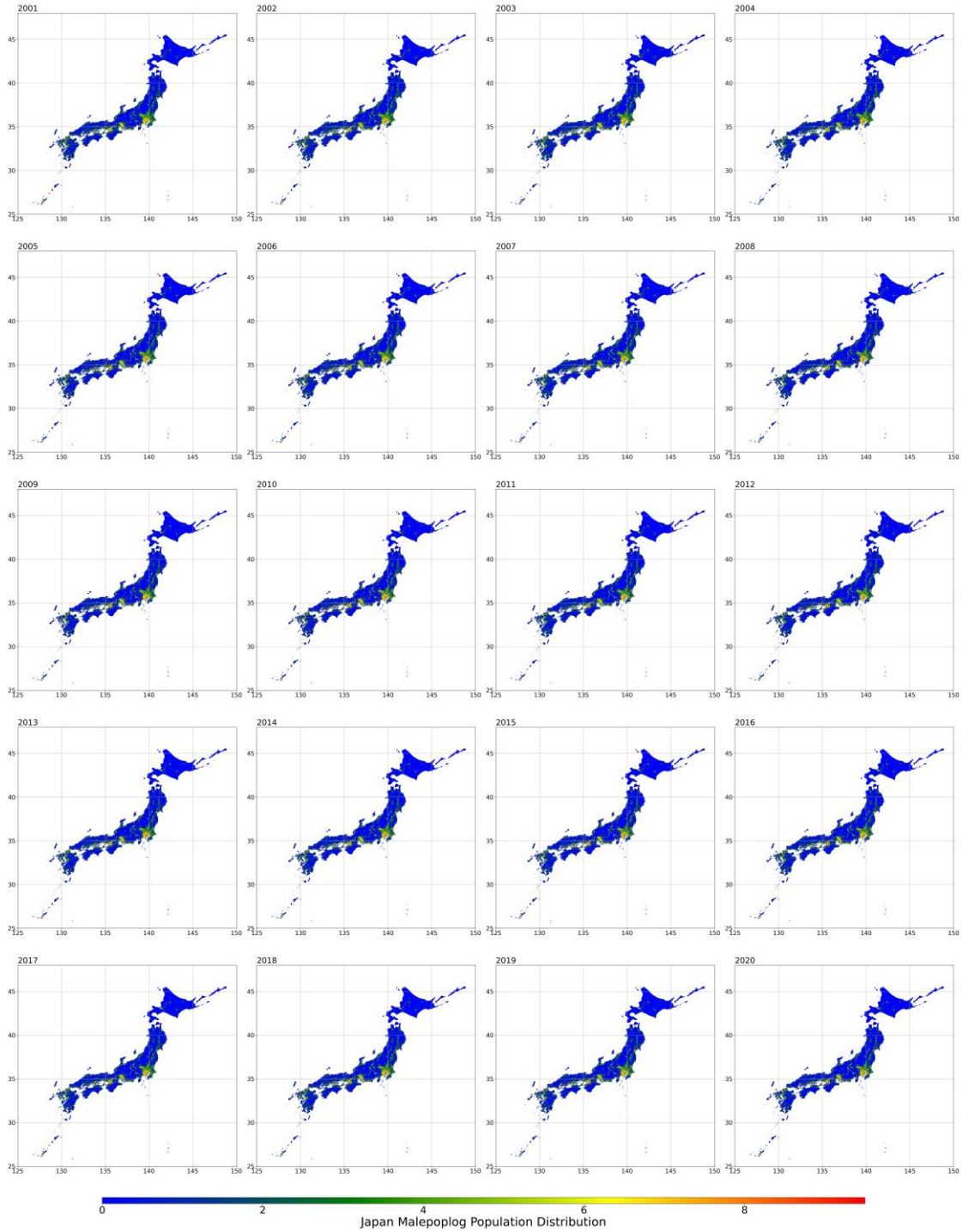


Figure 1. The schematic overview of the workflow.



403
404 **Figure 2. Total population distribution from 2001 to 2020**



405
406 **Figure 3. Male population distribution from 2001 to 2020**

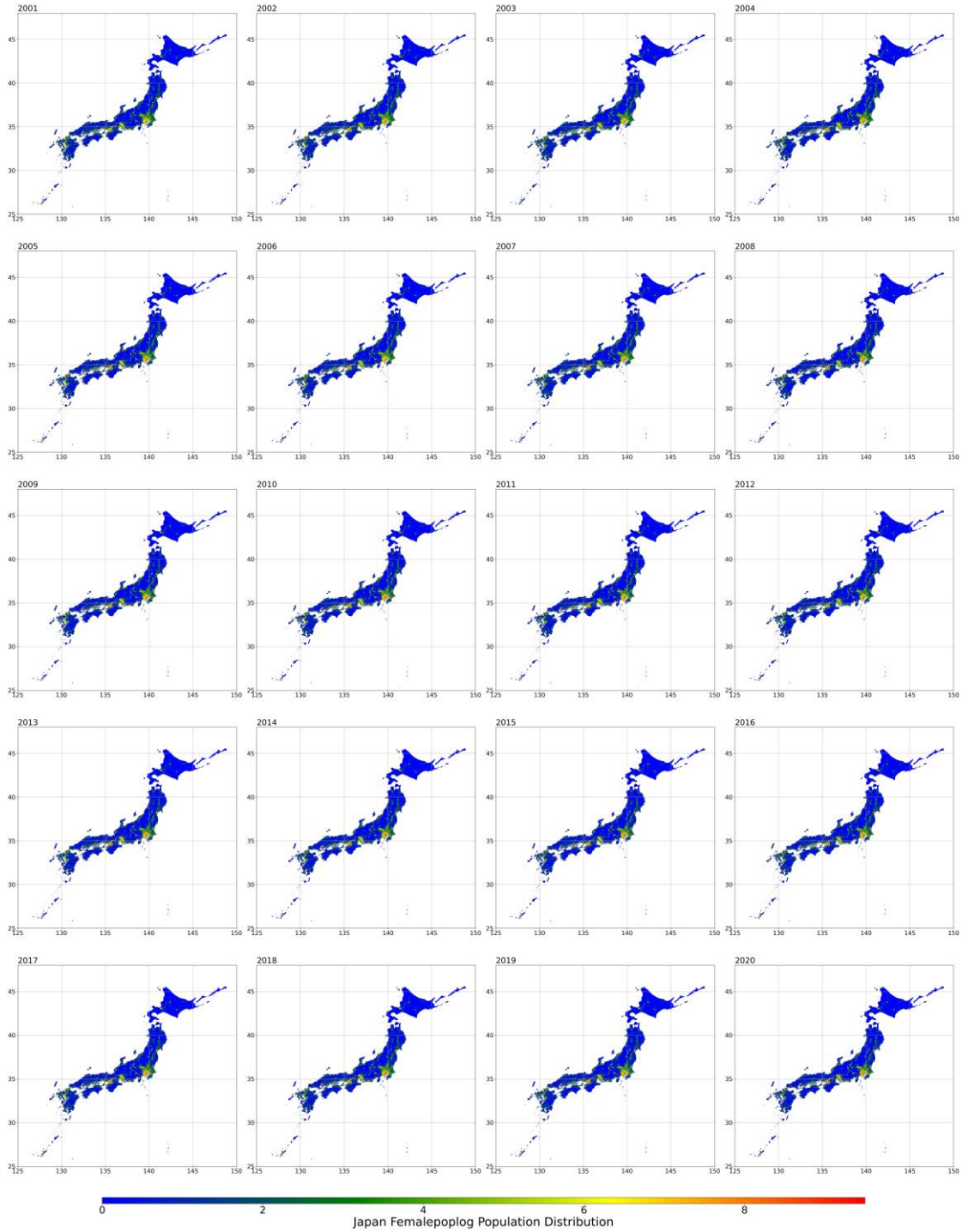
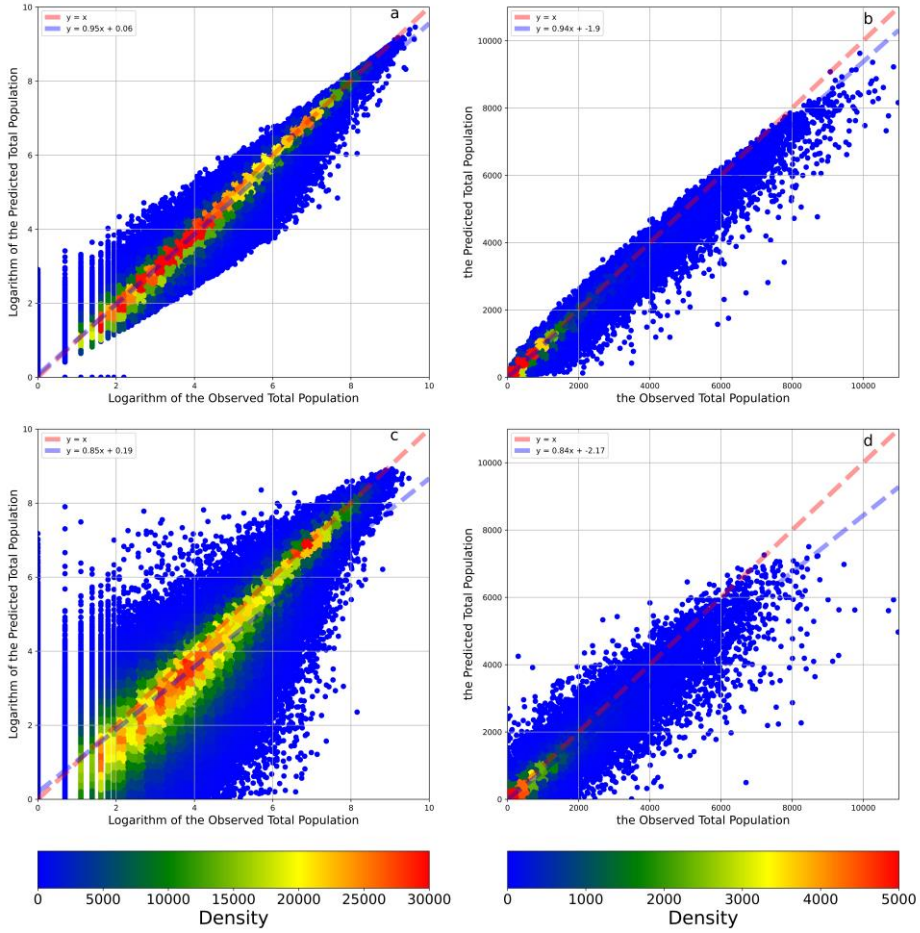


Figure 4. Female population distribution from 2001 to 2020



411
412 **Figure 5. The density plots between the observed and predicted total population and their**
413 **logarithms.** Panel **a** illustrates the density plots between the observed and predicted
414 logarithms of the total population. Panel **b** illustrates the density plots between the observed
415 and predicted total population. Panel **c** illustrates the density plots between the observed and
416 predicted logarithms of the total population in the 8:2 cross-validation. Panel **d** illustrates the
417 density plots between the observed and predicted total population in the 8:2 cross-validation.
418 The red dashed line is a 1:1 auxiliary line. The blue dashed line is the fit line between observed
419 and predicted data based on the linear regression.
420

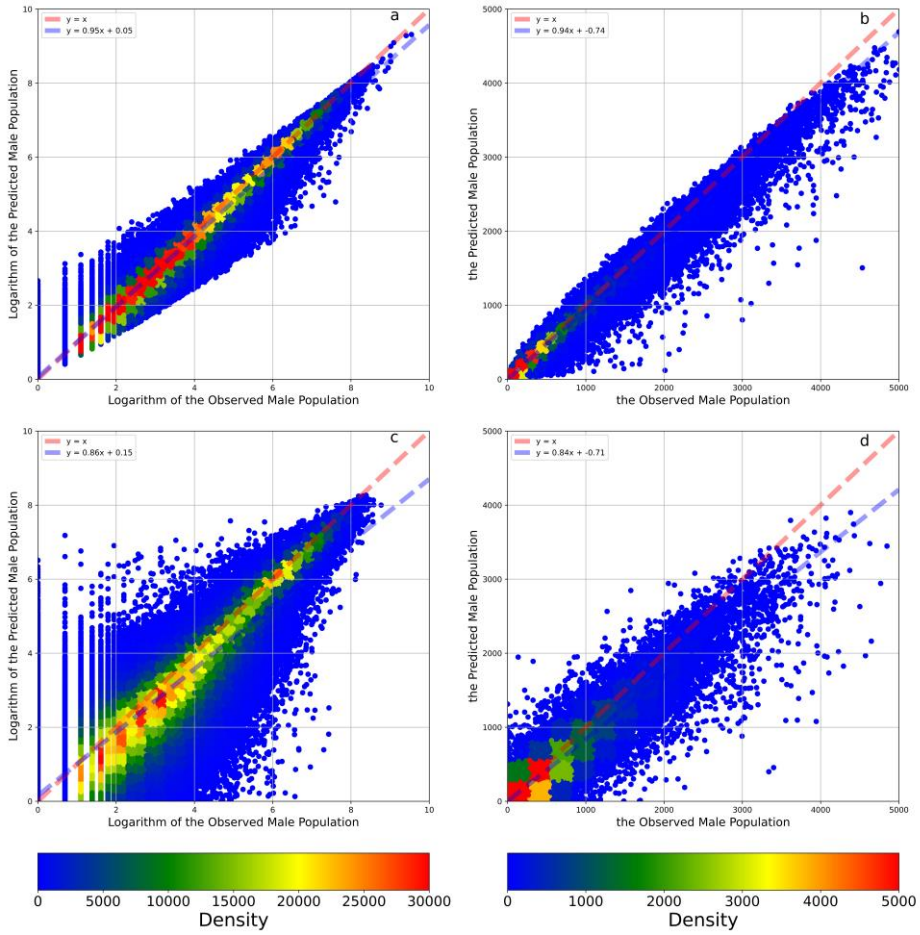


Figure 6. The density plots between the observed and predicted male population and their logarithms. Panel **a** illustrates the density plots between the observed and predicted logarithms of the male population. Panel **b** illustrates the density plots between the observed and predicted male population. Panel **c** illustrates the density plots between the observed and predicted logarithms of the male population in the 8:2 cross-validation. Panel **d** illustrates the density plots between the observed and predicted male population in the 8:2 cross-validation. The red dashed line is a 1:1 auxiliary line. The blue dashed line is the fit line between observed and predicted data based on the linear regression.

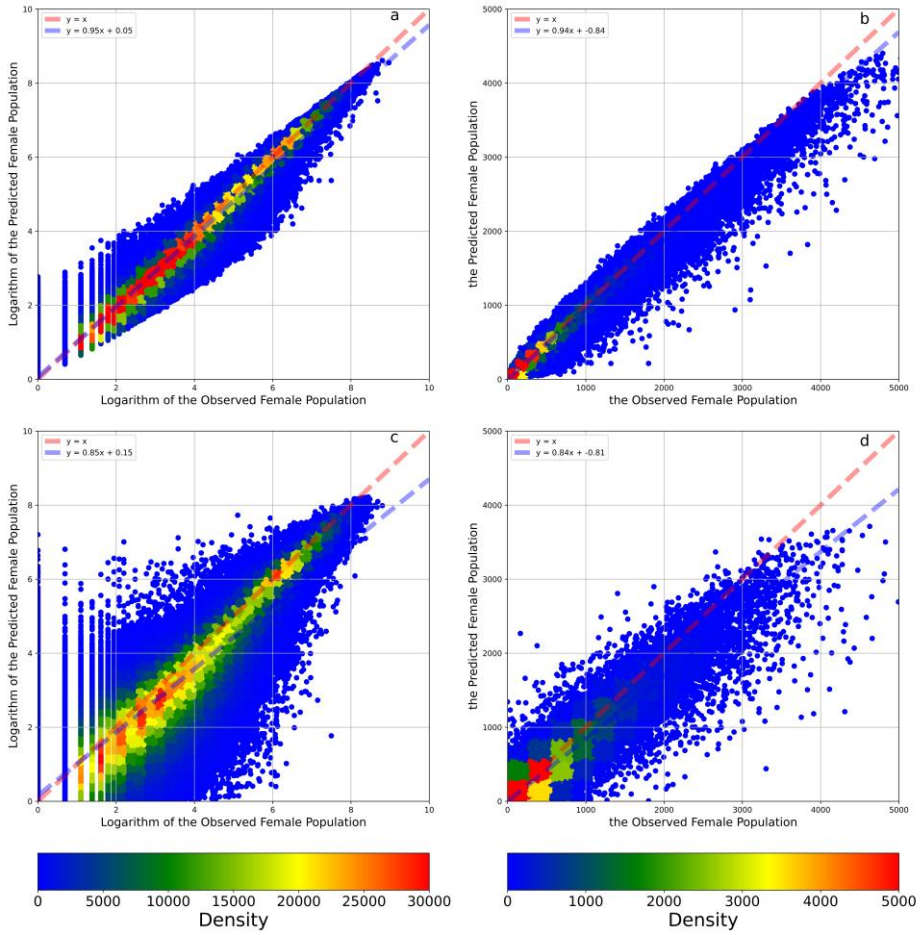


Figure 7. The density plots between the observed and predicted female population and their logarithms. Panel **a** illustrates the density plots between the observed and predicted logarithms of the female population. Panel **b** illustrates the density plots between the observed and predicted female population. Panel **c** illustrates the density plots between the observed and predicted logarithms of the female population in the 8:2 cross-validation. Panel **d** illustrates the density plots between the observed and predicted female population in the 8:2 cross-validation. The red dashed line is a 1:1 auxiliary line. The blue dashed line is the fit line between observed and predicted data based on the linear regression.

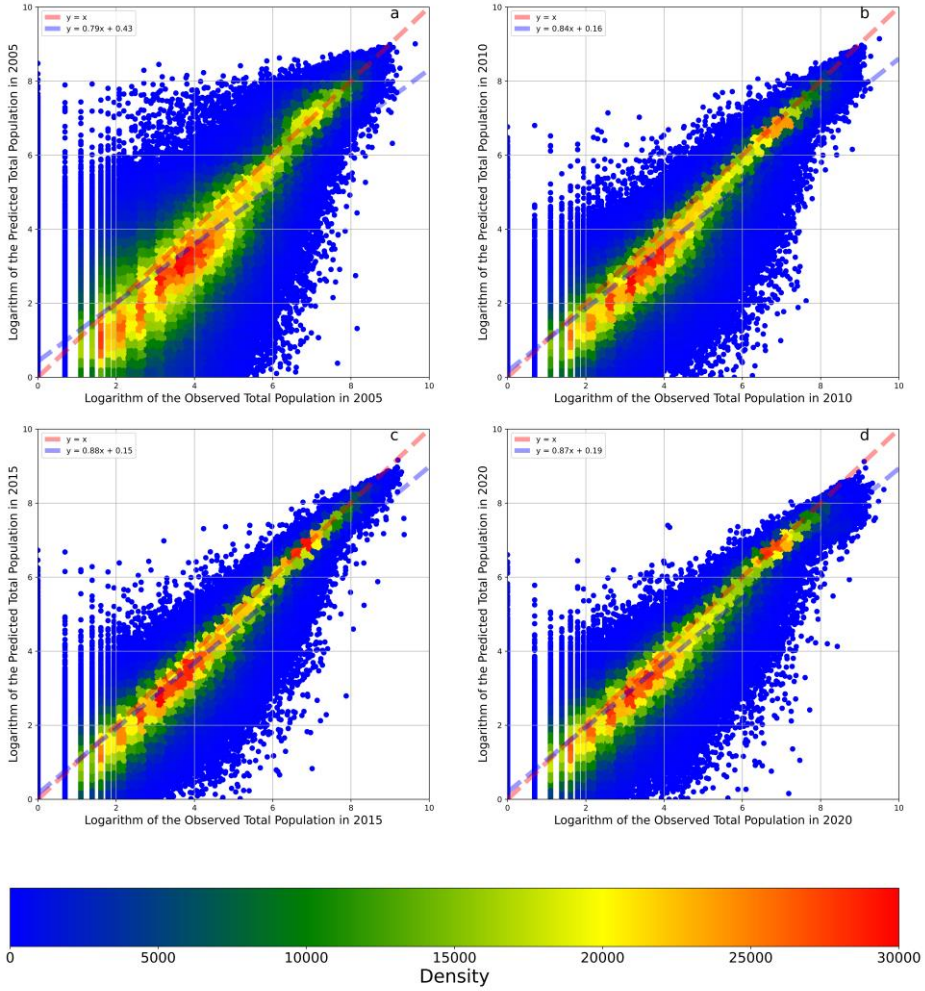


Figure 8. Temporal cross-validations of the model taking the logarithm of the total population as the output variable. Panel **a** illustrates the cross-validation result of the model trained by the data in 2010, 2015, and 2020 and tested by the data in 2005. Panel **b** illustrates the cross-validation result of the model trained by the data in 2005, 2015, and 2020 and tested by the data in 2010. Panel **c** illustrates the cross-validation result of the model trained by the data in 2005, 2010, and 2020 and tested by the data in 2015. Panel **d** illustrates the cross-validation result of the model trained by the data in 2005, 2010, and 2015 and tested by the data in 2020. The red dashed line is a 1:1 auxiliary line. The blue dashed line is the fit line between observed and predicted data based on the linear regression.

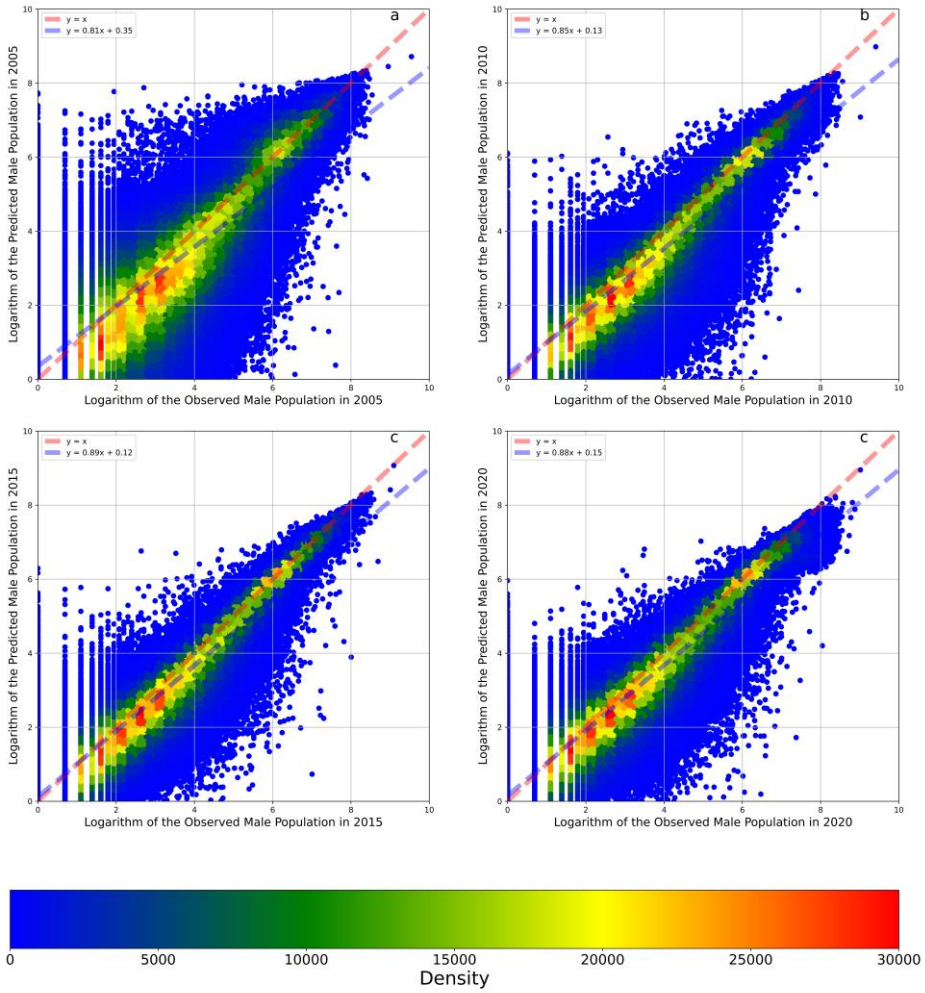


Figure 9. Temporal cross-validations of the model taking the logarithm of the male population as the output variable. Panel **a** illustrates the cross-validation result of the model trained by the data in 2010, 2015, and 2020 and tested by the data in 2005. Panel **b** illustrates the cross-validation result of the model trained by the data in 2005, 2015, and 2020 and tested by the data in 2010. Panel **c** illustrates the cross-validation result of the model trained by the data in 2005, 2010, and 2020 and tested by the data in 2015. Panel **d** illustrates the cross-validation result of the model trained by the data in 2005, 2010, and 2015 and tested by the data in 2020. The red dashed line is a 1:1 auxiliary line. The blue dashed line is the fit line between observed and predicted data based on the linear regression.

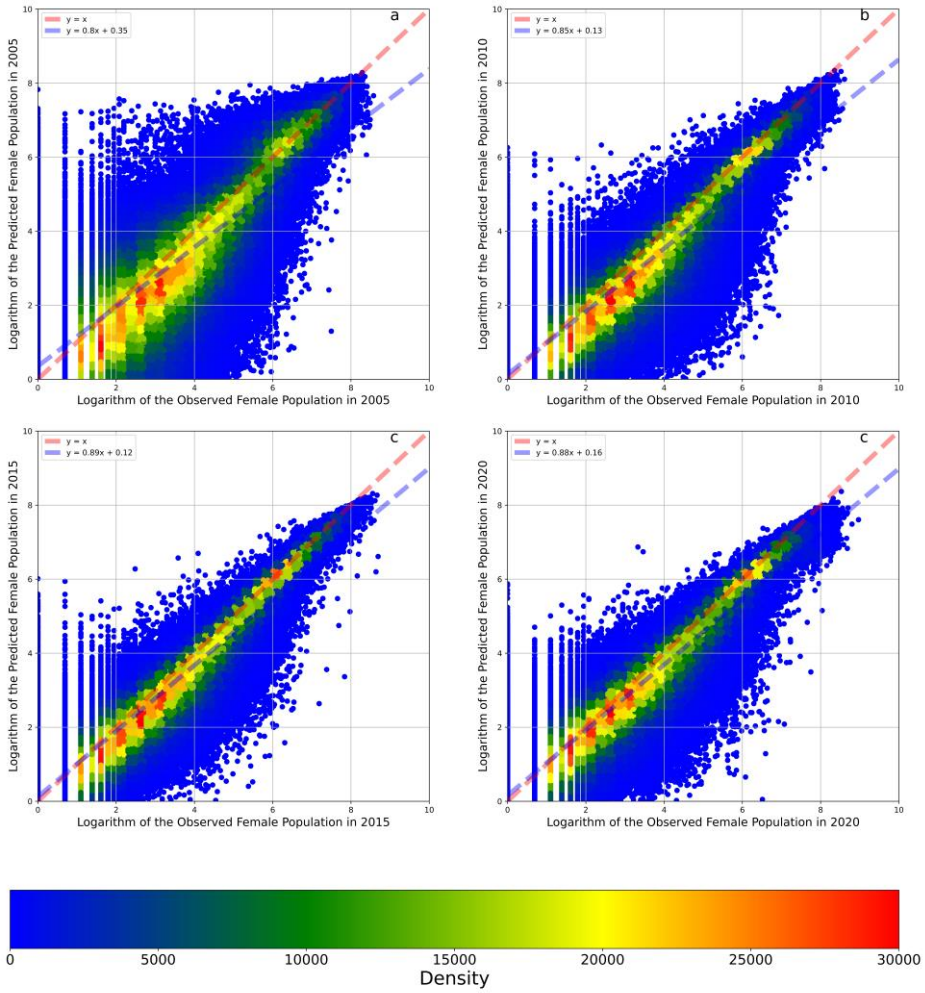


Figure 10. Temporal cross-validations of the model taking the logarithm of the female population as the output variable. Panel **a** illustrates the cross-validation result of the model trained by the data in 2010, 2015, and 2020 and tested by the data in 2005. Panel **b** illustrates the cross-validation result of the model trained by the data in 2005, 2015, and 2020 and tested by the data in 2010. Panel **c** illustrates the cross-validation result of the model trained by the data in 2005, 2010, and 2020 and tested by the data in 2015. Panel **d** illustrates the cross-validation result of the model trained by the data in 2005, 2010, and 2015 and tested by the data in 2020. The red dashed line is a 1:1 auxiliary line. The blue dashed line is the fit line between observed and predicted data based on the linear regression.

476 # *Tables*

Table 1: Variable information summary

Category	Variable Name	Processing Approach	Data Source	Time Stamps	Other Information
Output variable	Total population	Logarithmization	Japanese Government https://www.e-stat.go.jp/gis/statmap-search?page=1&type=1&toukeiCode=00200521	2005, 2010, 2015, 2020	Populaiton count per mesh at a 500m resolution
	Male population	Logarithmization			
	Female population	Logarithmization			
Land cover variable	Water bodies dummy	Spatial extraction	NASA MCD12Q1 https://lpdaac.usgs.gov/products/mcd12q1v006/	from 2001 to 2020	At a 500m resolution
	Distance to water bodies	Spatial join nearest			
	Evergreen needle leaf forests dummy	Spatial extraction			
	Distance of evergreen needle leaf forests	Spatial join nearest			
	Deciduous needle leaf forests dummy	Spatial extraction			
	Distance of deciduous needle leaf forests	Spatial join nearest			
	Deciduous broad leaf forests dummy	Spatial extraction			
	Distance of deciduous broad leaf forests	Spatial join nearest			
	Mixed forests dummy	Spatial extraction			
	Distance of mixed forests	Spatial join nearest			
	Closed shrublands dummy	Spatial extraction			
	Distance of closed shrublands	Spatial join nearest			
	Open shrublands dummy	Spatial extraction			
	Distance of open shrublands	Spatial join nearest			
	Woody savannas dummy	Spatial extraction			
	Distance of woody savannas	Spatial join nearest			
	Savannas dummy	Spatial extraction			
	Distance to savannas	Spatial join nearest			
	Grasslands dummy	Spatial extraction			
	Distance of grasslands	Spatial join nearest			
	Permanent wetlands dummy	Spatial extraction			
	Distance of permanent wetlands	Spatial join nearest			
	Croplands dummy	Spatial extraction			
	Distance of croplands	Spatial join nearest			
	Urban and built-up lands dummy	Spatial extraction			
	Distance of urban and built-up lands	Spatial join nearest			
	Cropland/natural vegetation mosaics dummy	Spatial extraction			
	Distance of cropland/natural vegetation mosaics	Spatial join nearest			
	Non-vegetated lands dummy	Spatial extraction			
	Distance of non-vegetated lands	Spatial join nearest			

NTL	NTL	Spatial extraction	Previous Research	from 2001 to 2020	At a 1km resolution
NPP	NPP	Averaging the data in the same year and spatial extraction	NASA MOD17A3HGF & MYD17A3HGF https://lpdaac.usgs.gov/products/mod17a3hgfv006/	from 2001 to 2020	At a 500m resolution
Temperature and precipitation	Annual average daytime temperature	Averaging the data in the same year and spatial extraction	NASA MOD11A2 & MYD11A2 https://lpdaac.usgs.gov/products/mod11a2v006/	from 2001 to 2020	At a 1km resolution
	Annual standard deviation of daytime temperature	Calculating the standard deviation of the data in the same year and spatial extraction			
	Annual average nighttime temperature	Averaging the data in the same year and spatial extraction			
	Annual standard deviation of nighttime temperature	Calculating the standard deviation of the data in the same year and spatial extraction			
	Annual precipitation	Spatial extraction	NASA GPM_3IMERGM https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGM_06/summary	from 2001 to 2020	At a 0.1-arc-degree resolution
Elevation and slope	Elevation	Resampling by the averaging mehtod and spatial extraction	JAXA https://global.jaxa.jp/press/2015/05/20150518_daichi.html	2015	At a 30m resolution
	Slope	Resampling by the averaging mehtod, generating slope raster by Gdal and spatial extraction			
Distance to Features of Interest	Distance to rivers	Spatial join nearest	Japanese Government https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-W05.html	2007	Line shape file
	Distance to coastal lines	Spatial join nearest	Japan Government https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-C23.html	2006	Line shape file
	Distance high population zones	Spatial join nearest	Japan Government https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-A16-v2_3.html	2005, 2010, 2015	Polygon shape file
	Distance to railways	Spatial join nearest	Japan Government https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-N02-v3_0.html	from 2005 to 2008 and from 2011 to 2020	Line shape file
	Distance to railway stations	Spatial join nearest	Japan Government https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-N02-v3_0.html	from 2005 to 2008 and from 2011 to 2020	Point shape file
	Distance to entertainment facilities	Spatial join nearest	Japan Government https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-P02-v4_0.html	2006	Point shape file
	Distance to government branches	Spatial join nearest			
	Distance to police stations	Spatial join nearest			
	Distance to fireman stations	Spatial join nearest			
	Distance to schools	Spatial join nearest			
	Distance to hospitals	Spatial join nearest			
	Distance to post offices	Spatial join nearest			
	Distance to disabled or senior support facilities	Spatial join nearest			
	Road density	Spatial extraction	Japan Government https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-N04.html	2002, 2003, 2004, 2010	Polygon shape file
Location Information	Longitude Latitude	Centroid attribute	From the mesh data	from 2001 to 2020	Point shape file

477

Table 2: Accuracy Indicators

	Indicator	Logarithm of Total Population	Total Population	Logarithm of Male Population	Male Population	Logarithm of Female Population	Female Population
Fitting Model	OOB Score	90.02%	--	90.45%	--	90.41%	--
	R2	98.65%	97.99%	98.70%	98.57%	98.70%	98.63%
	MAE	0.13	9.32	0.11	4.48	0.11	4.70
	RMSE	0.24	44.97	0.20	22.25	0.21	22.75
	Intercept	0.06	-1.90	0.05	-0.74	0.05	-0.84
	Coefficient	0.95	0.94	0.95	0.94	0.95	0.94
Cross-Validation	R2	88.43%	92.00%	88.95%	91.90%	88.90%	92.00%
	MAE	0.38	23.73	0.32	11.57	0.33	12.13
	RMSE	0.71	107.36	0.60	52.87	0.61	55.01
	Intercept	0.19	-2.17	0.15	-0.71	0.15	-0.81
	Coefficient	0.85	0.84	0.85	0.84	0.85	0.84
Temporal Cross-Validation using Data in 2005	R2	76.51%	81.90%	77.57%	81.59%	77.34%	81.62%
	MAE	0.62	36.19	0.52	17.76	0.52	18.75
	RMSE	1.02	157.50	0.86	78.28	0.88	80.86
	Intercept	0.43	2.00	0.35	1.80	0.35	1.65
	Coefficient	0.79	0.89	0.81	0.89	0.80	0.89
Temporal Cross-Validation using Data in 2010	R2	90.73%	89.82%	91.18%	89.53%	91.14%	90.11%
	MAE	0.36	24.79	0.30	12.04	0.30	12.55
	RMSE	0.63	120.32	0.53	59.86	0.54	60.68
	Intercept	0.16	-1.61	0.13	-0.19	0.13	-0.20
	Coefficient	0.84	0.78	0.85	0.78	0.85	0.79
Temporal Cross-Validation using Data in 2015	R2	93.38%	95.13%	93.59%	95.01%	93.76%	95.10%
	MAE	0.30	18.05	0.25	8.86	0.25	9.22
	RMSE	0.53	84.45	0.43	41.81	0.45	43.41
	Intercept	0.15	-2.36	0.12	-0.88	0.12	-0.90
	Coefficient	0.88	0.88	0.89	0.88	0.89	0.87
Temporal Cross-Validation using Data in 2020	R2	92.65%	83.75%	92.86%	84.28%	93.03%	85.18%
	MAE	0.32	24.02	0.27	11.54	0.27	11.99
	RMSE	0.55	157.18	0.47	75.44	0.47	77.17
	Intercept	0.19	6.33	0.15	3.39	0.16	3.34
	Coefficient	0.87	0.69	0.88	0.70	0.88	0.71
	Temporal Reliability	88.32%	87.65%	88.80%	87.60%	88.82%	88.00%

478

479

References

1 Li, C. & Managi, S. Estimating monthly global ground-level NO2 concentrations using geographically weighted panel regression. *Remote Sens. Environ.* **280**, 113152, doi:https://doi.org/10.1016/j.rse.2022.113152 (2022).

2 Li, L. & Wu, J. Spatiotemporal estimation of satellite-borne and ground-level NO2 using full residual deep networks. *Remote Sens. Environ.* **254**, 112257, doi:10.1016/j.rse.2020.112257 (2021).

3 Stevens, F. R., Gaughan, A. E., Linard, C. & Tatem, A. J. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLoS One* **10**, e0107042, doi:10.1371/journal.pone.0107042 (2015).

4 Savtchenko, A. *et al.* Terra and Aqua MODIS products available from NASA GES DAAC. *Advances in Space Research* **34**, 710-714, doi:https://doi.org/10.1016/j.asr.2004.03.012 (2004).

5 Breiman, L. Random Forests. *Machine Learning* **45**, 5-32, doi:10.1023/a:1010933404324 (2001).

6 UN. *World Urbanization Prospects: The 2018 Revision*. (United Nations, 2019).

7 UN. *World Population Prospects 2019: Highlights*. (United Nations, 2019).

8 Muramatsu, N. & Akiyama, H. Japan: Super-Aging Society Preparing for the Future. *The Gerontologist* **51**, 425-432, doi:10.1093/geront/gnr067 (2011).

9 Chen, B. K. *et al.* Forecasting trends in disability in a super-aging society: Adapting the Future Elderly Model to Japan. *The Journal of the Economics of Ageing* **8**, 42-51, doi:10.1016/j.jeoa.2016.06.001 (2016).

10 Lloyd, C. T. *et al.* Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. *Big Earth Data* **3**, 108-139, doi:10.1080/20964471.2019.1625151 (2019).

11 Sorichetta, A. *et al.* High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Scientific Data* **2**, 150045, doi:10.1038/sdata.2015.45 (2015).

12 Iwai, K., Mizuno, S., Miyasaka, Y. & Mori, T. Correlation between suspended particles in the environmental air and causes of disease among inhabitants: Cross-sectional studies using the vital statistics and air pollution data in Japan. *Environ. Res.* **99**, 106-117, doi:https://doi.org/10.1016/j.envres.2004.11.004 (2005).

13 Azuma, K., Kagi, N., Kim, H. & Hayashi, M. Impact of climate and ambient air pollution on the epidemic growth during COVID-19 outbreak in Japan. *Environ. Res.* **190**, 110042, doi:https://doi.org/10.1016/j.envres.2020.110042 (2020).

14 Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D. & Pozzer, A. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature* **525**, 367-371, doi:10.1038/nature15371 (2015).

15 Zhang, B., Nozawa, W. & Managi, S. Sustainability measurements in China and Japan: an application of the inclusive wealth concept from a geographical perspective. *Regional Environmental Change* **20**, 65, doi:10.1007/s10113-020-01658-x (2020).

16 Li, C. & Managi, S. Land cover matters to human well-being. *Sci Rep* **11**, doi:10.1038/s41598-021-95351-6 (2021).

17 Martinez, G. S., Imai, C. & Masumo, K. Local Heat Stroke Prevention Plans in Japan: Characteristics and Elements for Public Health Adaptation to Climate Change. *Int. J. Environ. Res. Public Health* **8**, 4563-4581, doi:10.3390/ijerph8124563 (2011).

18 Ng, C. F. S., Ueda, K., Ono, M., Nitta, H. & Takami, A. Characterizing the effect of summer temperature on heatstroke-related emergency ambulance dispatches in the Kanto area of Japan. *International Journal of Biometeorology* **58**, 941-948, doi:10.1007/s00484-013-0677-4 (2014).

19 Sulla-Menashe, D. & Friedl, M. A. User guide to collection 6 MODIS land cover (MCD12Q1 and MCD12C1) product. *USGS: Reston, VA, USA* **1**, 18 (2018).
20 Chen, X. & Nordhaus, W. D. VIIRS Nighttime Lights in the Estimation of Cross-Sectional and Time-Series GDP. *Remote Sens.* **11**, 1057, doi:10.3390/rs11091057 (2019).
21 Chen, X. & Nordhaus, W. D. Using luminosity data as a proxy for economic statistics. **108**, 8589-8594, doi:doi:10.1073/pnas.1017031108 (2011).
22 Henderson, J. V., Storeygard, A. & Weil, D. N. Measuring Economic Growth from Outer Space. *American Economic Review* **102**, 994-1028, doi:10.1257/aer.102.2.994 (2012).
23 Jean, N. *et al.* Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790-794, doi:10.1126/science.aaf7894 (2016).
24 Tan, M. *et al.* Modeling population density based on nighttime light images and land use data in China. *Applied Geography* **90**, 239-247, doi:https://doi.org/10.1016/j.apgeog.2017.12.012 (2018).
25 Zeng, C., Zhou, Y., Wang, S., Yan, F. & Zhao, Q. Population spatialization in China based on night-time imagery and land use data. *International Journal of Remote Sensing* **32**, 9599-9620, doi:10.1080/01431161.2011.569581 (2011).
26 Zhang, Q. & Seto, K. C. Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data. *Remote Sens. Environ.* **115**, 2320-2329, doi:10.1016/j.rse.2011.04.032 (2011).
27 Zhao, N. Z., Liu, Y., Cao, G. F., Samson, E. L. & Zhang, J. Q. Forecasting China's GDP at the pixel level using nighttime lights time series and population images. *GISci. Remote Sens.* **54**, 407-425, doi:10.1080/15481603.2016.1276705 (2017).
28 Chen, Z. *et al.* An extended time series (2000–2018) of global NPP-VIIRS-like nighttime light data from a cross-sensor calibration. *Earth System Science Data* **13**, 889-906, doi:10.5194/essd-13-889-2021 (2021).
29 Field, C. B., Randerson, J. T. & Malmström, C. M. Global net primary production: combining ecology and remote sensing. *Remote Sens. Environ.* **51**, 74-88 (1995).
30 Imhoff, M. L. *et al.* Global patterns in human consumption of net primary production. *Nature* **429**, 870-873, doi:10.1038/nature02619 (2004).
31 Linard, C., Gilbert, M., Snow, R. W., Noor, A. M. & Tatem, A. J. Population Distribution, Settlement Patterns and Accessibility across Africa in 2010. *PLoS One* **7**, e31743, doi:10.1371/journal.pone.0031743 (2012).
32 Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18-22 (2002).
33 Kummu, M., Taka, M. & Guillaume, J. H. A. Gridded global datasets for Gross Domestic Product and Human Development Index over 1990–2015. *Scientific Data* **5**, 180004, doi:10.1038/sdata.2018.4 (2018).
34 Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. Classification And Regression Trees. doi:10.1201/9781315139470 (2017).
35 Schapire, R. E. 149-171 (Springer New York, 2003).
36 Probst, P. & Boulesteix, A.-L. To tune or not to tune the number of trees in random forest. *The Journal of Machine Learning Research* **18**, 6673-6690 (2017).