Supplementary Materials

Based on biophysical components and CA binarization, identify the scope of built-up areas in Shanghai

Yiwen Ji [1,†], Lang Zhang [2,3,4,†], Haoran Yu [2,3], Xinchen Gu [5,6,*] and Lei Zhang [7,*]

[1] Art School, Anhui Jianzhu University, Hefei 230022, China; yiwenji@njfu.edu.cn (Y.J.);
[2] School of Landscape Architecture, Nanjing Forestry University, Nanjing 210037, China; zl@shsyky.com (Lang Zhang); yuhaoran2021@ahjzu.edu.cn (H.Y.);
[3] Key Laboratory of National Forestry and Grassland Administration on Ecological Landscaping of Challenging Urban Sites, Shanghai Academy of Landscape Architecture Science and Planning Shanghai 200232,China;
[4] College of Architecture and Urban Planning, Tongji University, Shanghai 200092, China;
[5] State Key Laboratory of Hydraulic Engineering Simulation and Safety, School of Civil Engineering, Tianjin University, Tianjin 300072, China; gxc@tju.edu.cn (X.G.);
[6] State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, Beijing 100044, China;
[7] School of Architecture and Planning, Anhui Jianzhu University, Hefei 230022, China; 188371280@ahjzu.edu.cn (Lei Zhang).
† These authors contributed equally to this work.
* Correspondence: gxc@tju.edu.cn (X.G.); 188371280@ahjzu.edu.cn (Lei Zhang).

1. Extracting Impermeable Surface Based on Biophysical Component Index (BCI)

The remote sensing data used in this article is Landsat TOA (on satellite reflectance) data provided by the GEE platform, which is processed using radiometric calibration methods provided by Chander et al. The Landsat system is the main system for medium resolution remote sensing within the United States Earth Observation System, mainly used for land resource investigation and management, water resource investigation and management, surveying and mapping, etc. Landsat has developed for four generations, with the first generation being Landsat 1-3, the second generation being Landsat 4 and 5, the third generation being Landsat 6 and 7, and the fourth generation being Landsat 8. After generations of development, the technological level of Landsat has steadily improved and has initially achieved commercial operation. The Landsat-8 satellite will continue the duration of Earth observation data recording, and the global land observation mission will be extended to over 40 years. It will play an important role in many fields such as energy and water resource management, forest resource monitoring, human and environmental health, urban planning, post disaster reconstruction, and agriculture. The calculation process is shown in Figure S1:
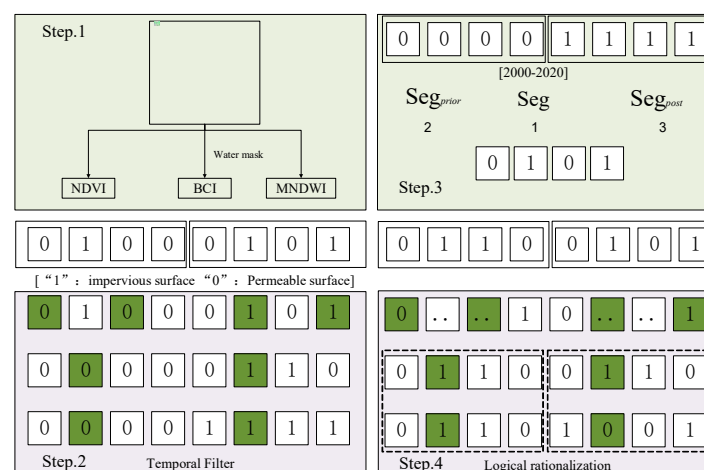
Figure S1. Calculation process.

GEE is a cloud platform provided by Google for analyzing and visualizing remote sensing images. It collects a large number of datasets and images, and users can freely use the data provided by the platform or upload their own data. With the powerful computing power of the cloud platform, they can quickly process and analyze large batches of data [2004]. During the years 2000, 2005, 2010, 2015, and 2020 in Shanghai, most remote sensing images were covered with clouds between June and September, so cloud removal processing is necessary. TOA images contain a band that measures cloud cover scores. The higher the value, the higher the likelihood that this pixel is a cloud. In this paper, pixels with cloud cover scores greater than 40 are masked, and then the median of all images within each year is calculated pixel by pixel to obtain a composite image for cloud removal. This article selects Landsat 5 images from 2000, 2005, and 2010, as well as Landsat 8 images from 2015 and 2020 as research data. Because the Landsat 7 satellite has been experiencing stripe issues since 2003, this article selected Landsat 5 and Landsat 8 images after 2000. However, due to the fact that the Landsat 8 satellite in 2010 has not yet been launched, the Landsat 5 image is still selected for the 2010 data in this article, and the impact of stripes is reduced to some extent through the synthesis of images. The synthetic data of image related bands is introduced in Table S1:

Table S1. Summary of Landsat TM band synthesis.

RGB	Formula	Describe
3, 2, 1	True color images.	Used for various types of land identification. The image is flat, the color tone is gray, the color is not saturated, and the information content is relatively small.
4, 3, 2	Standard false color image.	Its ground feature images are rich, distinct, and well layered, used for vegetation classification, water body recognition, and vegetation display in red.
7, 4, 3	Simulate true color images.	Used for identification of residential areas and water bodies.
7, 5, 4	Non-standard false color image.	The screen is slightly blue, used for special geological structure investigation.
5, 4, 1	Non-standard false color image.	Rich vegetation types used for studying plant classification.
4, 5, 3	Non-standard false color image.	Emphasize the display of water bodies, making it easy to distinguish between rivers and roads.
3, 4, 5	Non standard pseudo. color images close to true color.	It is more advantageous for distinguishing water systems, residential areas, urban streets, and forest land.
3, 2, 1	True color images.	Used for various types of land identification. The image is flat, the color tone is gray, the color is not saturated, and the information content is relatively small.

1.1. Cloud Removal

Firstly, using cloud cover scores on the GEE platform, all Landsat images from 2000 to 2020 in the study area were de cloud processed, and the median values of all images within each year were calculated pixel by pixel; Then calculate the annual NDVI and MNDWI, mask the water body in the image based on MNDWI, and use the masked image to calculate the annual BCI; Next, use the adaptive iteration method to obtain the thresholds of BCI and NDVI, and use the thresholds to obtain the initial extraction results of the impermeable surface; Finally, perform a time consistency check on the extraction results to obtain the final extraction result, and evaluate the accuracy. The specific steps will be described in detail below.

1.2. Index Calculation

This article uses a combination of BCI and NDVI to obtain the initial impermeable surface map. Compared with commonly used indices such as NDBI, NDVI, and NDISI,

BCI can better distinguish bare soil and vegetation from impermeable surfaces, and is therefore widely used for extracting impermeable surfaces. Before calculating BCI, it is necessary to mask the water body. In this paper, MNDWI is used to extract the water body, and the formula is as follows:

$$MNDWI = \frac{GREEN - MIR}{GREEN + MIR} \tag{1}$$

In the formula: GREEN is the green light band; MIR is in the mid infrared band.

Use GEE to perform TC transformation on the masked image. The TC transform, also known as the K-T transform or the tassel hat transform, has clear physical meanings for the bands obtained, such as brightness, greenness, and humidity, and is therefore widely used in land cover classification.

Normalize the first three bands obtained from TC transformation on GEE, limit the numerical range of the bands to 0-1, and then perform BCI calculation. The BCI calculation formula is as follows:

$$BCI = \frac{\frac{TC1 + TC2}{2} - TC2}{\frac{TC1 + TC3}{2} + TC2} \tag{2}$$

In the formula, TC1, TC2, and TC3 correspond to the first three bands obtained from the normalized TC transformation, namely brightness, greenness, and humidity.

In BCI images, the areas with higher values are the impermeable surface, while the areas with lower values are the permeable surface. However, Deng et al. pointed out that BCI can only remove some bare soil, and using only BCI for impermeable surface extraction will still mix up a lot of bare soil. NDVI can effectively distinguish vegetation from other land features and is widely used for vegetation extraction. For bare soil with vegetation alternation (such as arable land, wasteland, etc.), there will be significant fluctuations in NDVI relative to impermeable surfaces during the year, and the median NDVI throughout the year is relatively high. Therefore, this article combines BCI and NDVI to extract impermeable surfaces and attempts to eliminate this type of bare soil to a certain extent. Using the GEE platform to calculate NDVI, the NDVI calculation formula is as follows:

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{3}$$

In the formula: RED is the red light band; NIR is in the near-infrared band. In NDVI images, vegetation has higher values.

Export BCI images and NDVI images from the GEE platform for further processing.

1.3. Threshold Selection

How to find a suitable threshold to distinguish the target object from the background area has always been a problem that troubles researchers. Compared to other commonly used threshold selection methods, the adaptive iteration method has the advantages of objectivity, speed, and accuracy, and has certain applications in remote sensing image threshold segmentation. Therefore, this article applies the adaptive iteration method to obtain the thresholds of BCI and NDVI images. The main principle is the mathematical approximation idea, and the steps are as follows.

Find the maximum and minimum values of the exponential image, denoted as Z_{Max} and Z_{Min} and set the initial threshold to:

$$T_0 = Z_{max} + Z_{min} \tag{4}$$

Divide the image into foreground and background based on the threshold T, and then calculate the average value Z of the two separately Z_0 and Z_B. Set a new threshold:

$$T_0 = \frac{Z_0 + Z_B}{2} \tag{5}$$

If the new threshold does not change, it is the optimal threshold. Otherwise, repeat the above steps.

It should be noted that the numerical range of BCI and NDVI images should be stretched to an integer of 0-255, otherwise the iteration will fall into a dead cycle. After obtaining the annual thresholds of BCI and NDVI images, pixels that meet both BCI greater than threshold and NDVI less than threshold conditions are extracted to obtain the initial impermeable surface extraction results.

1.4. Time Consistency Analysis

In classification mapping of long time series, when the category of a pixel changes, it is necessary to determine whether the pixel has been misclassified or indeed changed. It is generally believed that once the object becomes impermeable, it will not return to the permeable surface in a short period of time. Based on this assumption, this article conducts a Temporal Consistency Check (TCC) on the extracted results to make the time series of impermeable surfaces more reasonable. Time consistency testing is divided into two steps: time filtering and logical rationalization.

Taking a pixel as an example, first, place the classification results of all years in a sequence, and mark the impermeable surface as 1 and the permeable surface as 0. Then, starting from 2001, add the labels of the previous year, itself, and the following year. Assuming that the label for 2001 is 0 and the result of the addition is 2, it indicates that this pixel was classified as impermeable in both 2000 and 2002. At this point, it can be concluded that the classification results for 2001 were incorrect and the label should be changed to 1. Similarly, assuming that the label for 2001 is 1 and the sum result is 1, it indicates that the classification results for this pixel in 2000 and 2002 are both transparent. Therefore, the label for 2001 will be changed to 0; Otherwise, do not change its markings. By analogy, filter the labels of all years except 2000 and 2020 to reduce misclassification in the classification results.

After time filtering the data, the phenomenon of misclassification will be reduced, but the logic of the time series may not be reasonable. For example, there may be alternating occurrences of continuous impermeable surfaces and continuous permeable surfaces in the time series, so it is necessary to logically rationalize them. Firstly, divide all years into three segments, with the years 2000-2002 and 2015-2017 lacking time context information recorded as Seg, respectively_ Prior and Seg_ Post, and from 2003 to 2014, it was recorded as Seg. Then, based on the frequency of the occurrence of impermeable and permeable surfaces in Seg for each pixel, it is divided into two types: impermeable and permeable dominated.

For pixels dominated by impermeable surfaces, in Seg, the first year classified as impermeable surfaces is found starting from 2003. Based on the aforementioned ideas, it can be considered that pixels classified as impermeable surfaces after this year are all misclassified. Therefore, the labels for subsequent years are changed to impermeable surfaces, while Seg_ All posts have been modified to be impermeable. For pixels dominated by permeable surfaces, search for the last continuous impermeable surface in Seg and change all subsequent years (excluding Seg_post) to impermeable surfaces. If this operation does not change the dominant type of Seg, it can be considered that this section of impermeable surface is correctly classified, but the previous impermeable surface and the subsequent permeable surface are both misclassified, so they will be changed. If this would change the dominant type of Seg, then it is believed that this section of impermeable surface is a misidentification, so Seg and Seg are separated_ Change all priors to permeable surfaces.

2. Identifying Land Use in Shanghai Based on GEE Platform and Multiple Machine Learning Methods

2.1. Data Preparation

The dynamic change span of land use change monitoring data in Shanghai is 20 years, divided into 5 time periods: 2000, 2005, 2010, 2015, and 2020. Due to different data sources in different periods, different spatial resolutions, and different processing methods.

The information source for remote sensing monitoring data in 2000, 2005, and 2010 is LADSAT-TM images from the United States, with a resolution of 30 × 30m; In order to better highlight the characteristics of surface vegetation, TM4, 3, and 2 bands and MSS4, 5, and 7 bands were selected for false color synthesis.

When determining the seasonal phase of remote sensing images, attention should be paid not only to the quality of the instantaneous coverage of remote sensing information within the survey area (such as technical indicators such as cloud cover<10%), but also to the seasonal differences in timeliness in different regions. Based on the principle of maximizing the richness of surface information reflected in the image under instantaneous conditions, images from late May to mid June or from late August to mid September are mainly selected. For some remote areas and areas where data acquisition is difficult, the time limit can be appropriately relaxed.

2.2. Classification Strategy

Collecting accurate training and validation samples is a necessary condition for accurate land cover classification. Non representative and/or insufficient samples will lead to significant uncertainty in the results of land cover classification. Considering the actual situation of Shanghai and referring to the existing land cover classification system and products, the classification system for this study is shown in Table S2. The classification system used in this study is a land cover classification system based on ecosystem types, which can be divided into agricultural land, forest land, grass land, wetlands, water bodies, and construction land.

Table S2. Land cover classification system and products.

Classification	Type	Abbreviation	Features
1	Agriculture	Agr	A type of land cover that is greatly affected by intensive human activities. During the year, there are significant changes from bare land to sowing, to crop growth and harvest. It includes paddy fields, greenhouse agriculture, and other cultivated and cultivated land.
2	Forests	For	Areas with a tree coverage rate of 15% and a tree height of>3m, including natural forests, artificial forests, and fruit trees.
3	Grass	GL	Areas with a herbaceous coverage of 15%, including natural grasslands and pastures.
4	Water	WB	All inland water bodies; Mainly natural and artificial water bodies.
5	Constructed	Con	Including urban, rural, and industrial and mining land that is greatly affected by human activities.
6	Wetlands	Wl	It usually has a significant high reflectivity in the near-infrared band; A swamp covered with aquatic herbaceous plants; Mud flats are also included.

Using area estimation adjustment method to achieve quality and quantity control of sample point data. This can adjust the bias caused by stratified sampling. Calculate the confusion matrix n of classification results using sample data n_{ij}. A more informative representation of the error matrix is the unbiased estimation of the area ratio of the units i and j of the error matrix p_{ij}:

$$p_{ij} = W_i \frac{n_{ij}}{n_i} \qquad (6)$$

In the equation, W_i is the proportion of land cover type area in the total area of the study area in the classification results, and the proportion of area classified as *i* is:

$$W_i = \frac{A_i}{A_{tot}} \tag{7}$$

In the equation, A_{tot} is the total area of the map, A_i is the map area of land cover type *i*. Then it is the unbiased estimator of the total area of type *j*:

$$A_j = A_{tot} \times p_j \tag{8}$$

Where, the calculation formula for p_j is as follows:

$$p_j = \sum_i W_i \frac{n_{ij}}{n_i} \tag{9}$$

The estimated standard error of the estimated area ratio is:

$$S(p_j) = \sqrt{\sum_{i=1}^{q} W_i^2 \frac{\frac{n_{ij}}{n_i}\left(1 - \frac{n_{ij}}{n_i}\right)}{n_i - 1}} \tag{10}$$

The standard error of the estimated area after error adjustment is:

$$S(A_j) = A_{tot} \times S(p_j) \tag{11}$$

The approximate 95% confidence interval for A_j is:

$$A_j \pm 2 \times S(A_j) \tag{12}$$

Among them, the error range is defined as z-score multiplied by standard error, and the value of z-score is related to the confidence level (for 95% confidence, Z=1.96). When the area of each land cover type in the classification results is within the estimated area range, it can be considered that the sample data on which the classification results are based is reasonable. This method accurately quantifies the classification error caused by inaccurate sampling, and the final number of sample points is shown in Table S3. By using the "Random Column" function in the GEE cloud platform, 70% of the sample data can be randomly selected for classifier training and image classification, and the classification results can be verified and evaluated using 30% of the sample data.

Table S3. Sample quantity points of various types of land use.

Classification	Type	Amount
1	Agriculture	600
2	Forests	400
3	Grass	400
4	Water	100
5	Constructed	400
6	Wetlands	100
Total	/	2000

2.3. Structural Features

2.3.1. Spectral Index

Previous studies have shown that the application of remote sensing spectral indices can effectively improve the accuracy of identifying different land cover types. This study used the GEE cloud platform to calculate the normalized differential vegetation index (NDVI), vegetation enhancement index (EVI), land surface water index (LSWI), normalized differential water index (NDWI), soil adjusted vegetation index (SA VI), and bare

land index (BSI) based on the calculation formulas for each spectral index. Add each indicator to the original remote sensing image in sequence, and the calculation formula for each indicator is as follows:

$$NDVI = \frac{\rho_{NIR} - \rho_{RED}}{\rho_{NIR} + \rho_{RED}} \tag{13}$$

$$EVI = 2.5 \times \frac{\rho_{NIR} - \rho_{RED}}{\rho_{NIR} + 6 \times \rho_{RED} - 7.5 \times \rho_{BLUE} + 1} \tag{14}$$

$$LSWI = \frac{\rho_{NIR} - \rho_{SWIR}}{\rho_{NIR} + \rho_{SWIR}} \tag{15}$$

$$NDWI = \frac{\rho_{GREEN} - \rho_{NIR}}{\rho_{GREEN} + \rho_{NIR}} \tag{16}$$

$$SAVI = \frac{(\rho_{NIR} - \rho_{RED})(1 + L)}{(\rho_{NIR} + \rho_{RED} + L)} \tag{17}$$

$$BSI = \frac{(\rho_{RED} + \rho_{SWIR}) - (\rho_{NIR} + \rho_{BLUE})}{(\rho_{RED} + \rho_{SWIR}) + (\rho_{NIR} + \rho_{BLUE})} \tag{18}$$

Among them, $\rho SWIR$, ρNIR, ρRED, $\rho GREEN$, $\rho BLUE$ represents the surface reflectance values in the short wave infrared, near-infrared, red, green, and blue bands. L is the soil conditioning factor, with a value range of 0-1, usually taken as 0.5, to better reduce background differences in the soil and eliminate the impact of soil noise.

2.3.2. Texture Features

Texture features are important attributes of remote sensing images, and different types of land cover have different texture features. Based on texture features, the accuracy of recognition and classification can be improved. The Gray Level Co occurrence Matrix (GLCM) is a classic method for extracting texture features. It extracts texture through conditional probability density between grayscale levels of remote sensing images and is widely used in land cover classification research. The calculation of GLCM can be obtained through the "Glcm Texture" function in the GEE cloud platform. This study selected several commonly used texture features for classification, including NDVI based feature parameters such as angular second moment (asm), contrast (Con), correlation (Corr), variance (Var), inverse moment of difference (Idm), mean (Savg), and entropy (Ent). The calculation of each feature variable is based on GLCM, which is dimensionless and their value range is not completely uniform.

2.3.3. Radar Features

Research has shown that SAR data is sensitive to land cover types such as water bodies, construction land, and farmland. The radar variables involved in constructing feature variables in this study include the backscatter coefficients of Landsat data in VV polarization and VH polarization modes [2008].

2.3.4. Topographical Features

In the recognition and classification of remote sensing images, the participation of terrain features can improve the accuracy of classification. Therefore, based on SRTMGL1_ The "ee. Terrain." of the 003 digital elevation data product. Use the "product" function provided by GEE to calculate slope aspect, slope shade, slope, and elevation, and add them as independent features to remote sensing images.

Based on the above feature variables, the image is classified into five combinations of feature variables. In the first input feature variable combination, only the spectral bands of Landsat are involved in image classification, and then the feature variable combinations of spectral indicators, terrain features, radar features, and texture features are sequentially added

2.4. Classification Method

At present, remote sensing classification methods such as Support Vector Machine (SVM), Classification Regression Tree (CART), and Random Forest (RF) algorithms have been widely applied in land cover mapping and crop type recognition. In order to analyze the differences in extraction and classification accuracy of different land cover types using different classification algorithms, as well as the differences in extraction and classification accuracy of different land cover types when using the same feature variable combination, this study used three classification algorithms. The following will introduce these three classification algorithms.

2.4.1. Support Vector Machine (SVM)

Support Vector Machine (SVM) was proposed by Vapnik in 1995. Support vector machines have significant advantages in nonlinear problems, small samples, high-dimensional, and other aspects. Due to its small training sample size and support for high-dimensional feature spaces, it is widely used. The parameters that need to be adjusted when using SVM are kernel function type, kernel gamma value, and cost parameters.

2.4.2. Classification and Regression Tree (CART)

CART (classification and compression tree) was proposed by Breiman et al. in 1984. Due to its simple structure, fast computation speed, and ease of understanding, it is widely used in land cover extraction and remote sensing image classification research. When CART classifies remote sensing images, the parameters that need to be optimized are the maximum and minimum values of each leaf node.

2.4.3. Random Forest (RF)

The Random Forest (RF) algorithm was proposed by Leo Beriman in 2001. Research has shown that RF processing of remote sensing data has advantages such as stability, speed, and high accuracy; Therefore, it has become the most widely used classifier in remote sensing classification algorithms. It has important applications in crop extraction, image classification, agricultural regression models, and other fields. When classifying remote sensing images, the two main parameters that RF needs to adjust and optimize are the number of decision trees to be created and the minimum number of leaf nodes. Research has shown that the value of RF parameters has little impact on accuracy, so we set the number of decision trees to 300. In addition, when using RF for classification in GEE, the importance score of the parameters participating in the classifier can be calculated using the "Explain" function. The value of importance score is not absolute and there is no unified fixed range of values, but varies with the number of sampling data and feature parameters involved in classification. The importance and contribution level of feature parameters participating in classification can be determined by the relative values of importance scores in specific situations.

2.5. Accuracy Verification

Confusion matrix is a commonly used method for evaluating the accuracy of image classification. This study calculates the confusion matrix through online programming on the GEE cloud platform, and then calculates the overall accuracy (OA), Kappa coefficient, producer accuracy (PA), and user accuracy (UA). Among them, OA and Kappa coefficients can fully reflect the comprehensive accuracy of the results, while PA and UA can be used to evaluate the classification accuracy of a certain land cover type.

Overall Accuracy refers to the accuracy index obtained by dividing the total number of grid units extracted from any random sample of land use classification results that are consistent with the actual land use type by the total number of grid units extracted. It is the probability that any land use classification result to be evaluated and the actual land use in the same location in the validation sample map are the same type of land use.

The Kappa coefficient refers to the sum of the total number of grid units in a land use type multiplied by the diagonal of the confusion matrix, minus the product of the total

number of grid units in the actual land use type in a certain category and the total number of grid units classified into that category in the evaluated land use type, and then dividing by the square of the total number of pixels minus the product of the real surface pixels in this category and the total number of pixels classified in this category. It is a precision evaluation index established on the basis of discrete multivariate evaluation used in the accuracy evaluation stage of land use type maps (Table S4). The calculation formula is as follows:

$$OA = \frac{\sum_{i=1}^{n} X_{ii}}{N} \times 100\% \tag{19}$$

$$PA = \frac{X_{ii}}{X_{+i}} \times 100\% \tag{20}$$

$$UA = \frac{X_{ii}}{X_{i+}} \times 100\% \tag{21}$$

In the formula, N is the total number of samples used for accuracy evaluation; N is the total number of columns in the confusion matrix; X_{ii} is the number of samples in the ith row and ith column of the confusion matrix; The uncertainty analysis of classification result accuracy is achieved by calculating the error range, which is defined as z-score multiplied by standard error. In this study, all data processing and calculations, including data collection, data processing, image synthesis, feature space construction, parameter calculation, classifier implementation, confusion matrix calculation, etc., were implemented using the GEE JavaScript API.

Table S4. Kappa coefficient classification quality table.

Kappa Coefficient	0.00-0.20	0.20-0.40	0.40-0.60	0.60-0.80	0.80-1.00
Classification Quality	Not good	Commonly	Good	Very good	Excellent