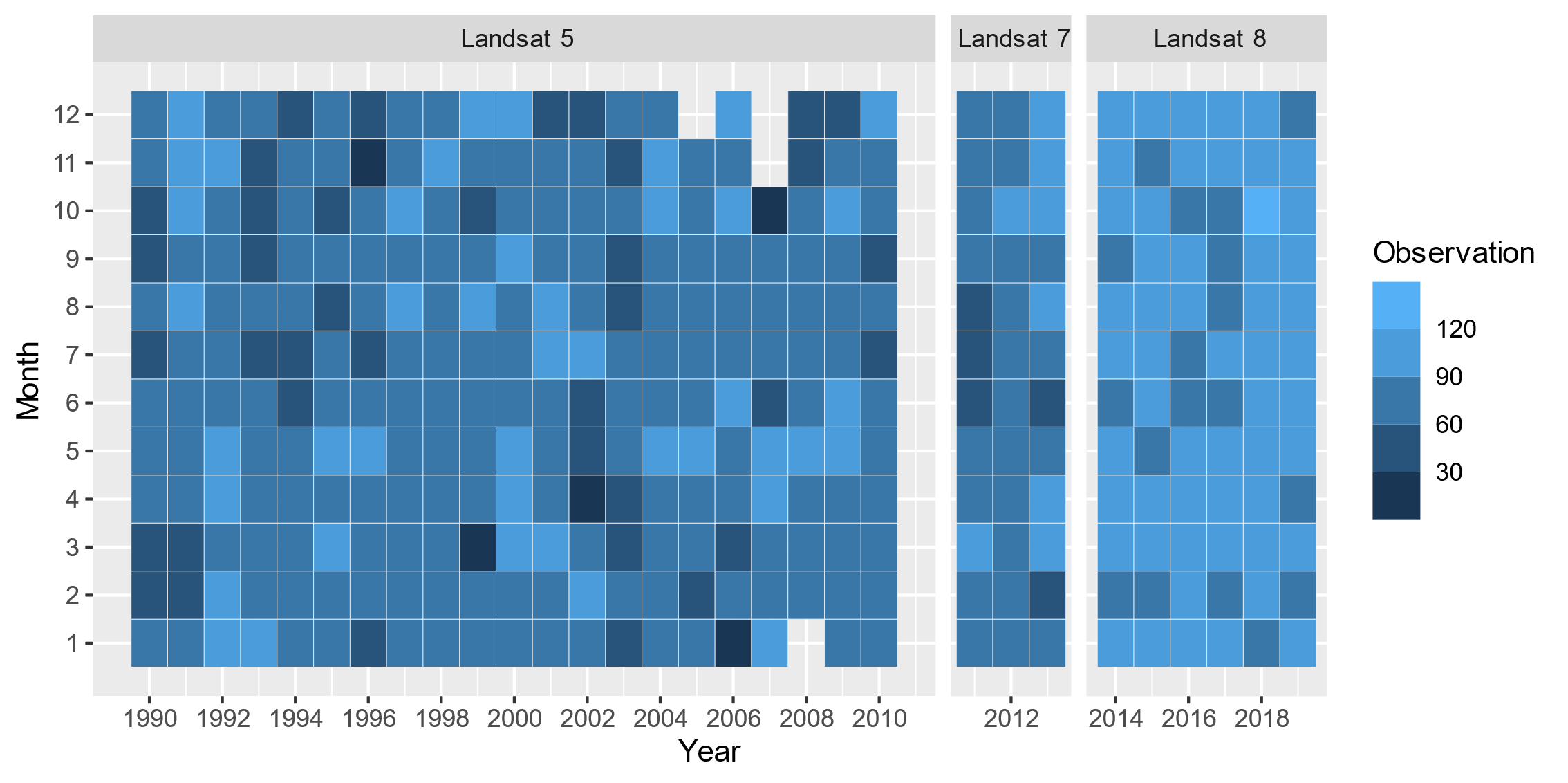
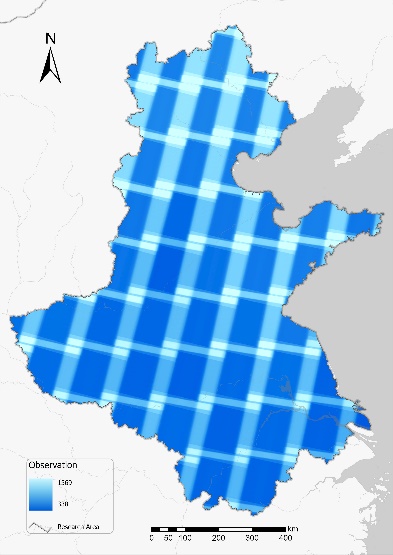
Supplementary information for the manuscript

This supplementary information concludes the additional information for the built-up land mapping. Some of the parameter settings are explained here because they are not essential for the main manuscript. Readers can use this supplementary information to understand the detailed settings for the method. Also, the code for this study is publicly accessible on GitHub (<https://github.com/wangjinzhulala/North_China_Plain_GEE_Organized>). The data can be interactively inspected at [GEE-APP](https://wangjinzhulala.users.earthengine.app/view/built-up-evolution-in-5-middle-eastern-china-provinces) and downloaded at [Google-Drive](https://drive.google.com/file/d/1_aMbMaylWWjAJClCJ0hxbSDy9K3GXbbz/view?usp=sharing) or [Baidu-Netdisk](http://pan.baidu.com/s/1xJel0-iRD3_suhJku1EEuw) (download code: mg85).

# 1. Preparation of the spectral and temporal predictors

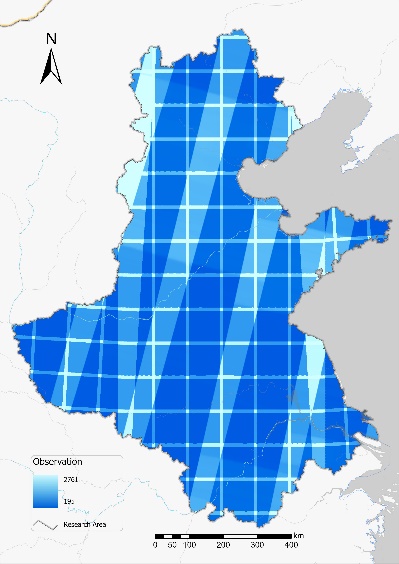
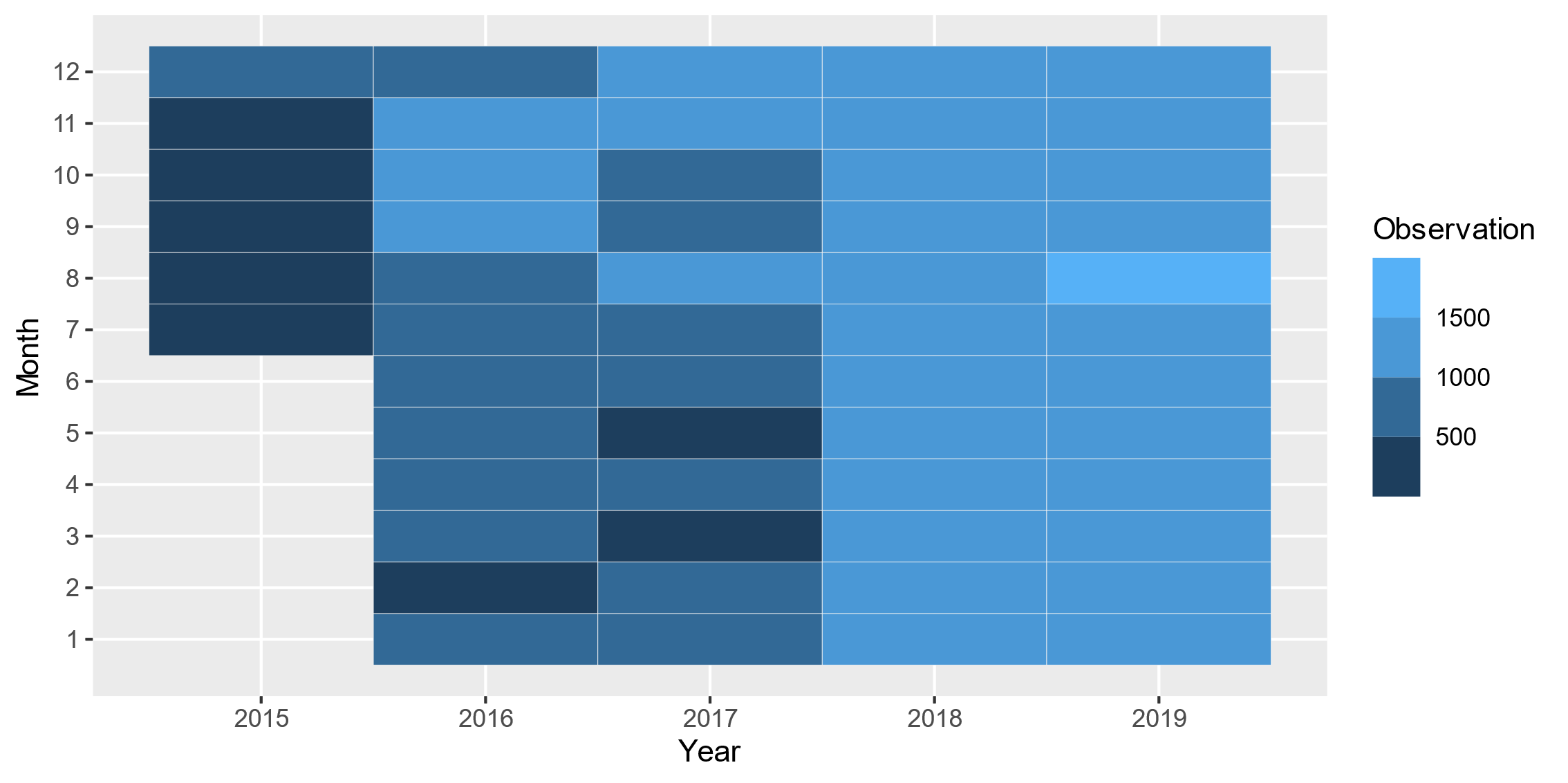
## 1.1 Spectral data

The spectral predictors were cloud-free images computed from Landsat top of atmosphere (TOA) TM/ETM+/OLI and Sentinel 2A multispectral instrument (MSI) data. All the Landsat TOA data passed geometric and radiometric corrections and can be accessed on the Google Earth Engine (GEE) platform. A total of 28,361 scenes were used in this study. Among them, 18,625 scenes were acquired from Landsat 5 TM, 2,796 from Landsat 7 ETM+, and 6,940 from Landsat 8 OLI. Because of the Scan Line Corrector (SCL) failure of Landsat 7 in 2003, the ETM+ data were used when there were no other available Landsat data. As a result, the TM, ETM+, and OLI data used in this study were set as 1990–2010, 2011–2013, and 2014–2019, respectively (Figure 1).



**Figure 1. The spatial and temporal distribution of Landsat top of atmosphere.**

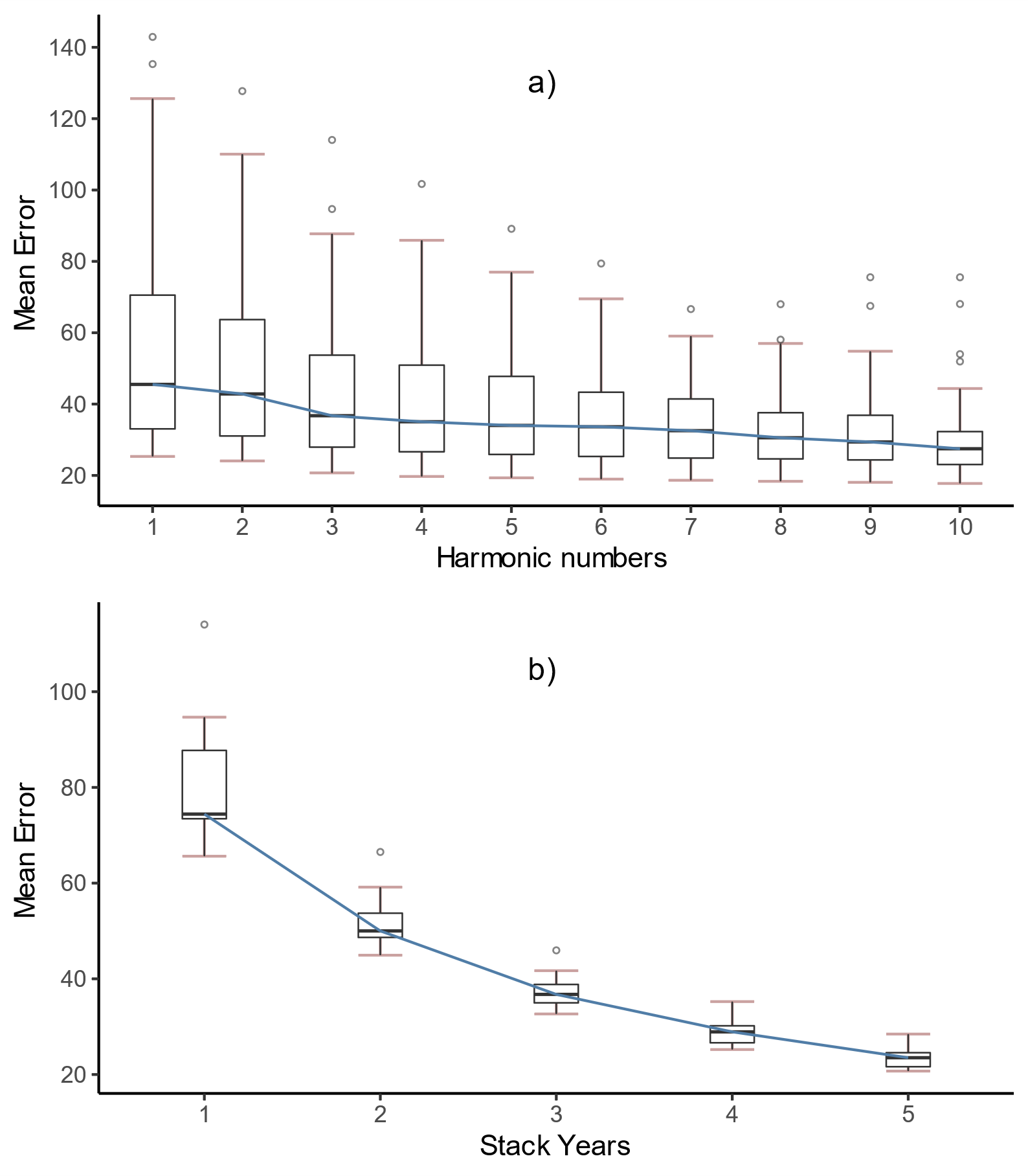
The level-1C product from Sentinel 2 MSI, which passed geometric and radiometric corrections, was selected because data from level-2A, which underwent one more radiometric correction than level-1C, were not fully available. All level-1C data can be accessed on GEE. Because the earliest accessible data were from June 2015, Sentinel 2 MSI was used with Landsat 8 OLI data for classification after 2015 (Figure 2).



**Figure 2. Spatial and temporal distribution of the Sentinel 2 MSI data.**

## 1.2 Discrete Fourier Transformation

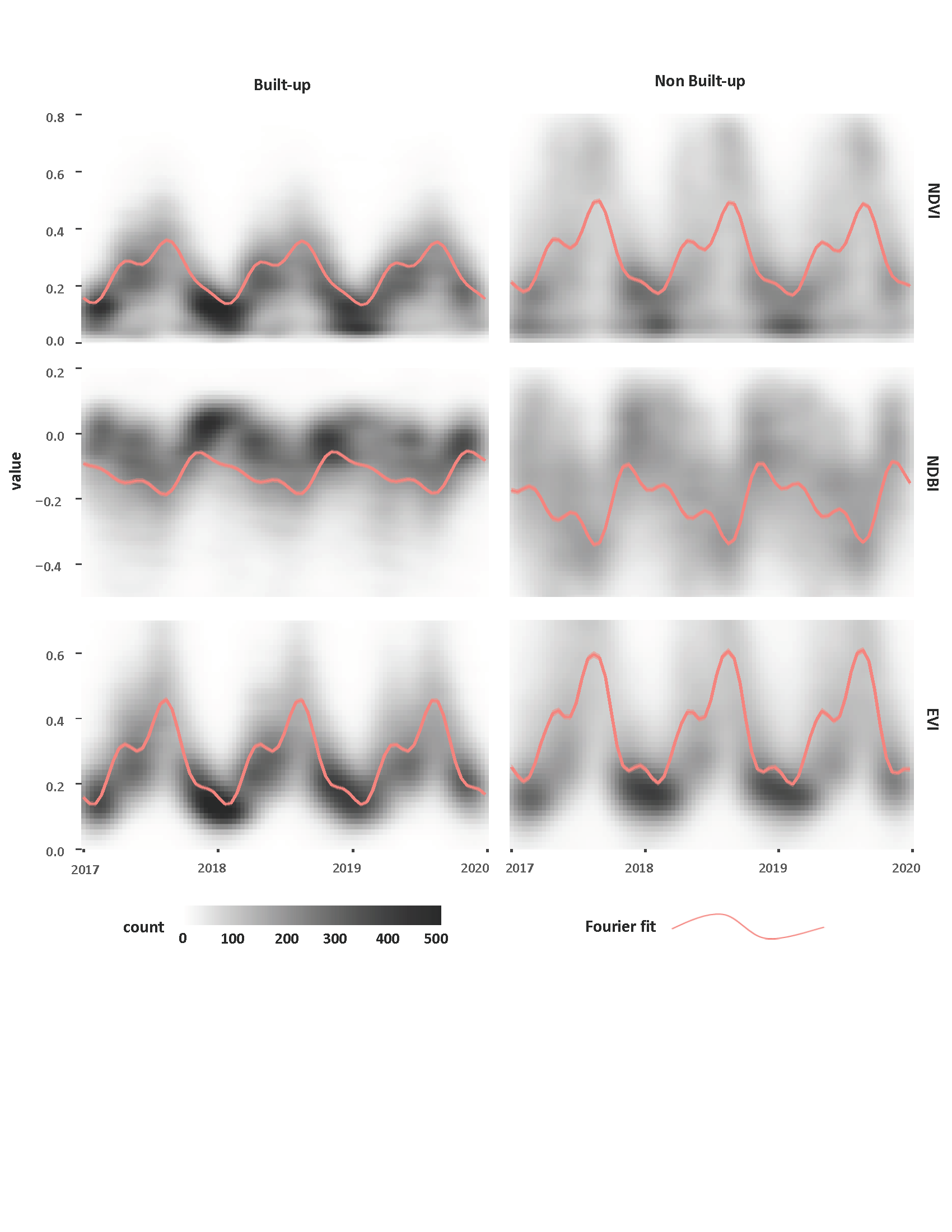
Before using the discrete Fourier Transform, the number of sinuates functions (harmonic numbers) and the volume of normalized indices (stack years) should be determined. A sensitivity test was conducted to select these parameters (Figure 3). Specifically, 100 random points were distributed through the research area, and the mean error between the original value and the fitted value for each point was computed with different harmonic numbers and stack years. The harmonic numbers were set to 1–10, and the stack years were set to 1–5 (where 1 means using only the normalized data from 2015, while 5 means using all the normalized indices from 2015–2019). The harmonic number was determined to be 3, i.e., where the most significant drop in mean error occurred. Fewer harmonic numbers are preferred as they produce fewer coefficients for later classification. The stack year was also determined to be 3 by balancing the data used for the discrete Fourier transform and the mean error decrease. Fewer stack years are preferred because built-up land can be mapped at a higher frequency if fewer data are used for the discrete Fourier transform.



**Figure 3. Sensitivity analysis for harmonic numbers and stack years in the Discrete Fourier Transformation.**

## 1.3 Discrete Fourier transform for validation samples

The indices data (NDVI, NDVI, and EVI) of hold-out samples in 2017–2019 are shown in Figure 4. The values of built-up samples are more concentrated than those for non-built-up samples, possibly because they comprise less heterogeneous land covers. The Fourier fitting lines of the built-up samples are lower than those of non-built-up samples, providing a robust way for the RF classifier to distinguish built from non-built-up pixels.



**Figure 4. The Discrete Fourier Transform of hold-out samples in 2017–2019. The density plot in the background shows the temporal distribution of values for the normalized indices. The red fitted lines are the result of the Discrete Fourier Transform using background points.**

# 2. Sample collection and inspection

## 2.1 Framework for sample collection

Because the conversion from non-built-up land to built-up land is unlikely to occur (Gong et al. 2020), built-up samples collected using base-maps from 1990–1992 were used for classification in 1990–2013. The non-built-up samples collected using Google Earth High Definition (HD) maps were used for the classification from 1990 to 2019. The sample collection framework is summarized in Figure *5*. Given the 30-year research period, a few sample points may be incorrect. The built-up samples were re-inspected using the Google Earth HD map of 2014 and then used for classification in 2014–2019. We also re-inspected samples to ensure high accuracy in the last two classifications, which will be used as masks to remove inconsistencies in the former classifications.

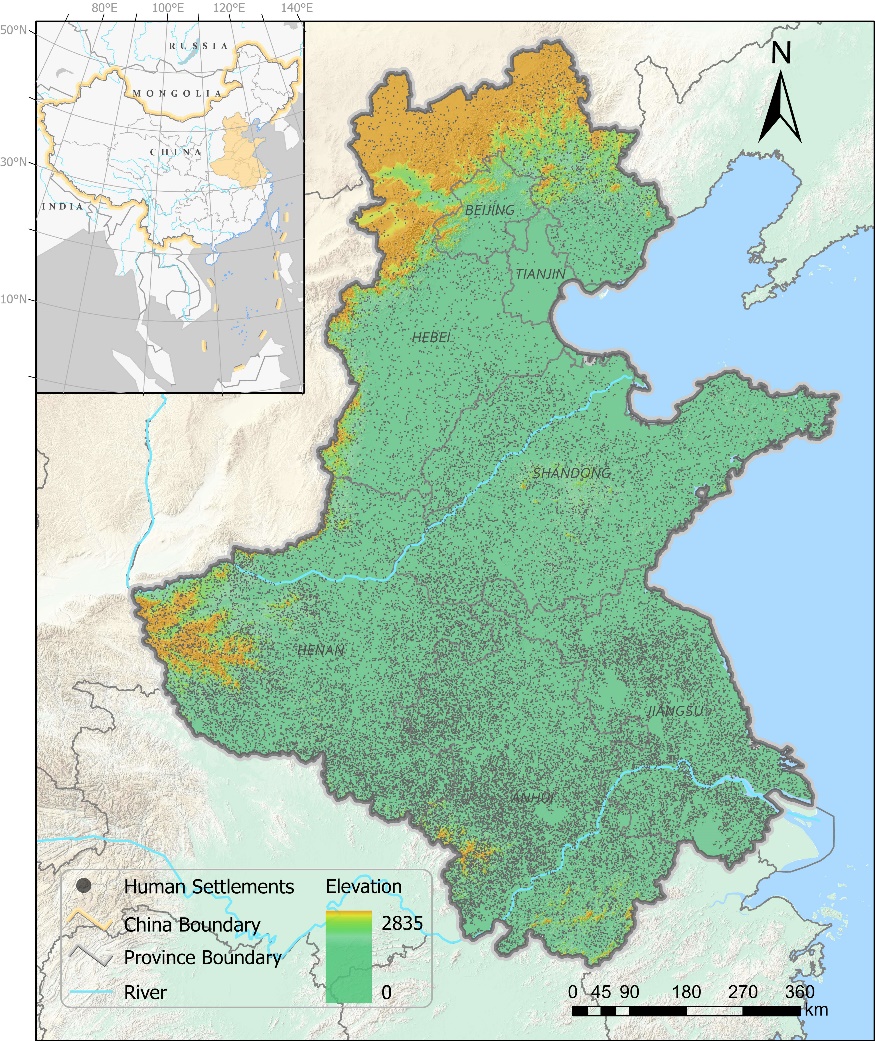


**Figure 5. Framework for sample collection.**

## 2.2 Collection of built-up samples

### 2.2.1 Raw built-up samples

The raw built-up samples were taken from the National Settlements Database of China (<http://www.resdc.cn/>). These records were generated in 2000 and comprised two types of settlement: government sites (including the department offices of provinces, cities, districts, counties, towns, and villages) and the offices of nationally owned companies (Figure 6). The total number of National Settlement points of the study area is 751,411, exceeding the analysis capacity in this study. We randomly subset 5,000 points from the total dataset, then used historical Landsat images to visually check each point and further diminish the number to 4,000 by excluding low-quality points (e.g., those near water bodies or in hilly areas).



**Figure 6. Records of the National Settlements Database of China, where each point represents a government office.**

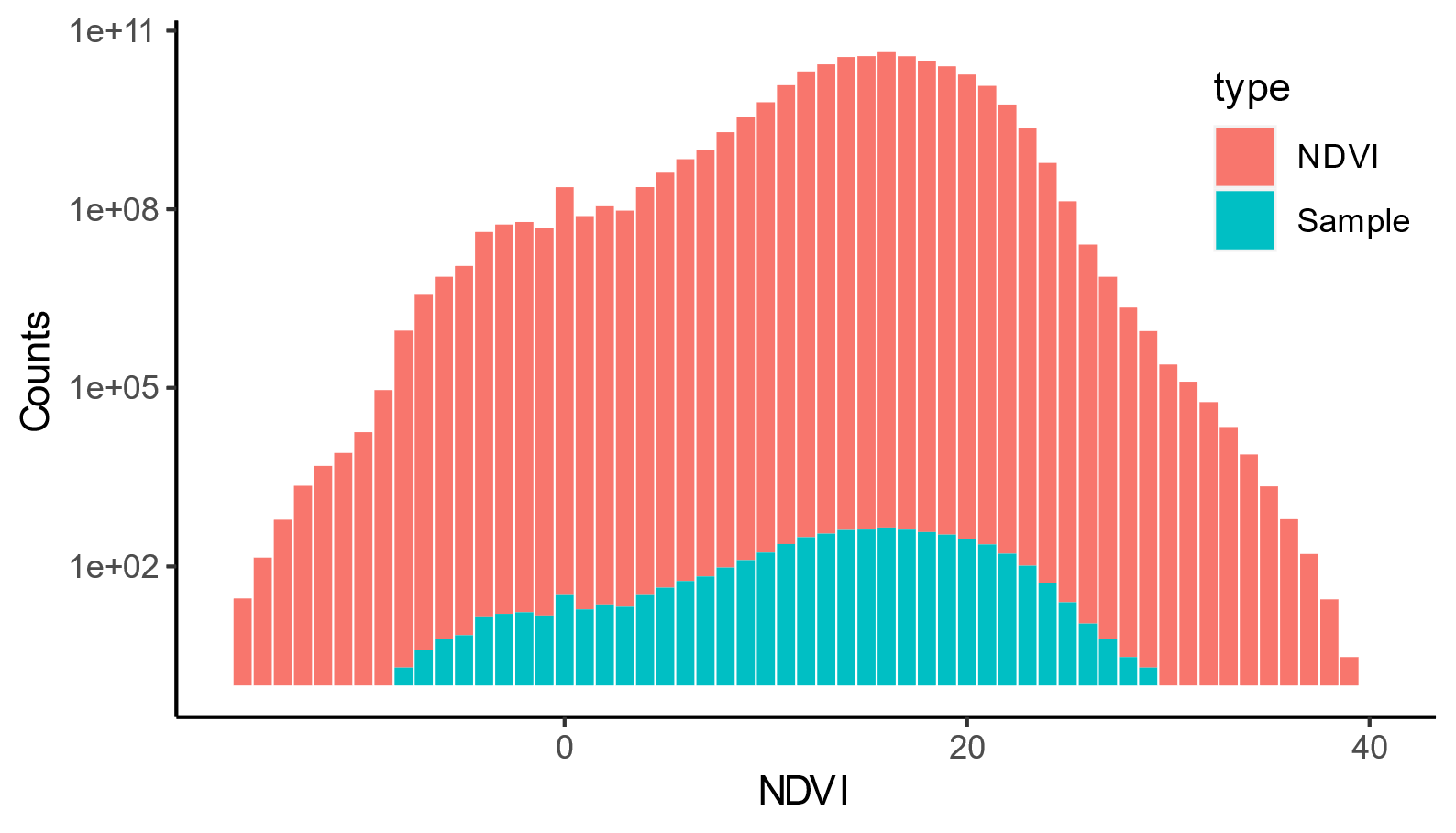
### 2.2.2 Visual inspection of built-up samples

Because of the low quality of Landsat data from 1990 to 1992, two false-color base maps (one map created using NDVI, NDBI, and EVI; the other map created from the coefficients of the temporal features) were used to assist with a visual inspection. Each sample point was inspected against all three base maps. We manually nudged their position to the center of nearby built-up land for some points located at positions that could be easily misclassified (such as the edge of a village or a skim road).

## 2.3 Collection of non-built-up samples

### 2.3.1 Raw non-built-up samples

A stratified sampling scheme was used to create the raw non-built-up samples. NDVI was used to stratify raw samples because it can distinguish different land covers effectively (Li et al. 2015), thus promoting even distributed non-built-up samples among different land covers. The raw non-built-up samples were produced with the following procedures. First, NDVI data were produced from the cloud-free image of the research area in 2017–2019. Then 50,000 random points were generated to extract the value of NDVI. Next, the random points were reduced to 5,000, where the histogram of NDVI data was used to stratify the reduction (Figure 7). Finally, these 5,000 points were visually checked.



**Figure 7. Histogram of NDVI data and non-built-up samples. NDVI was multiplied by 100 and converted to integer format. The sample values have a similar distribution to the NDVI data of the research area, suggesting they are evenly distributed among land covers.**

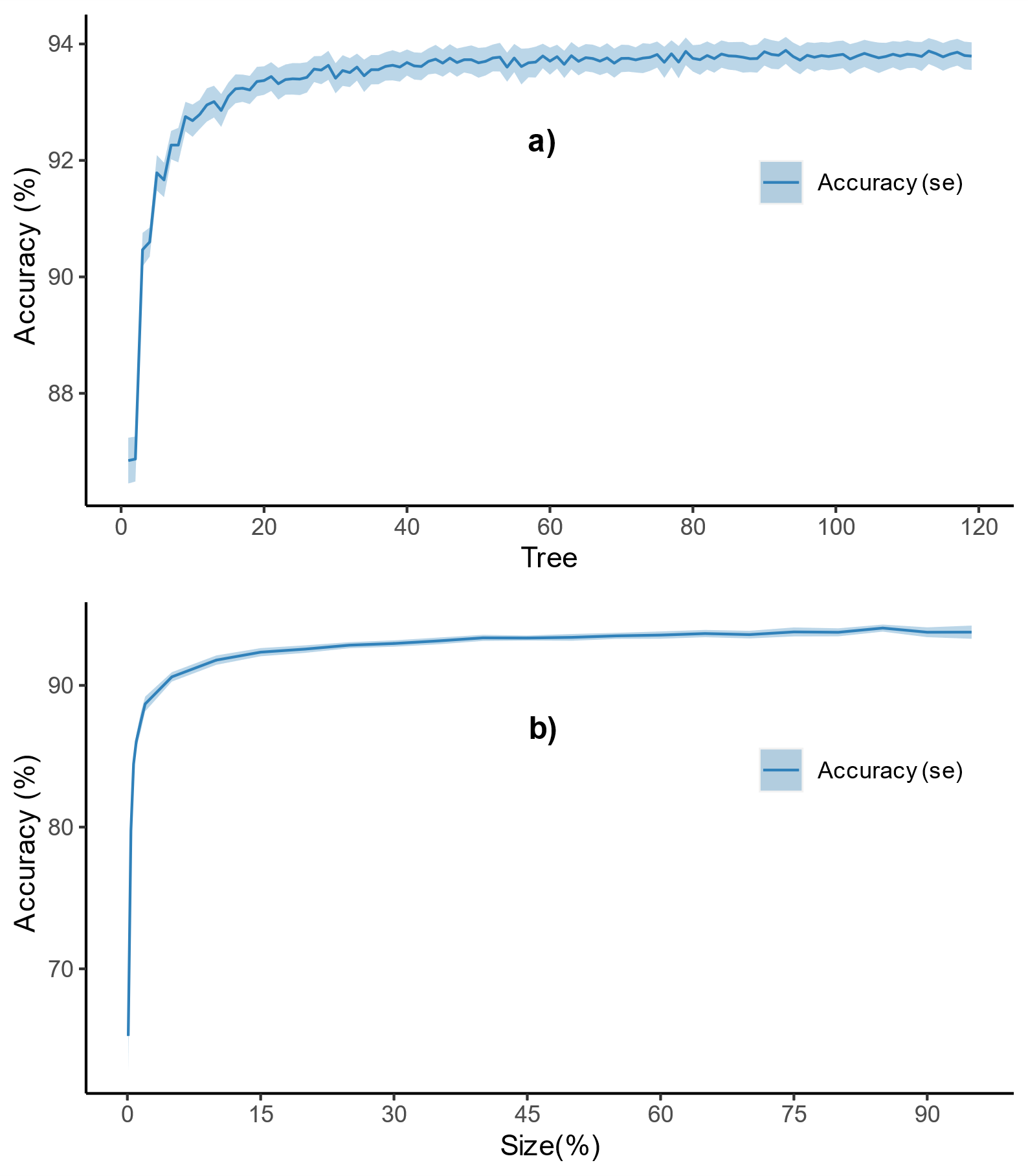
### 2.3.2 Visual inspection of non-built-up samples

Non-built-up samples were visually inspected using the Google Earth HD map of 2019. Points located in built-up lands were removed. Points located close to built-up lands were manually nudged to nearby non-built-up land to avoid interference.

# 3. Random Forest training and sample variation

## 3.1 Training sample size and tree number in Random Forests

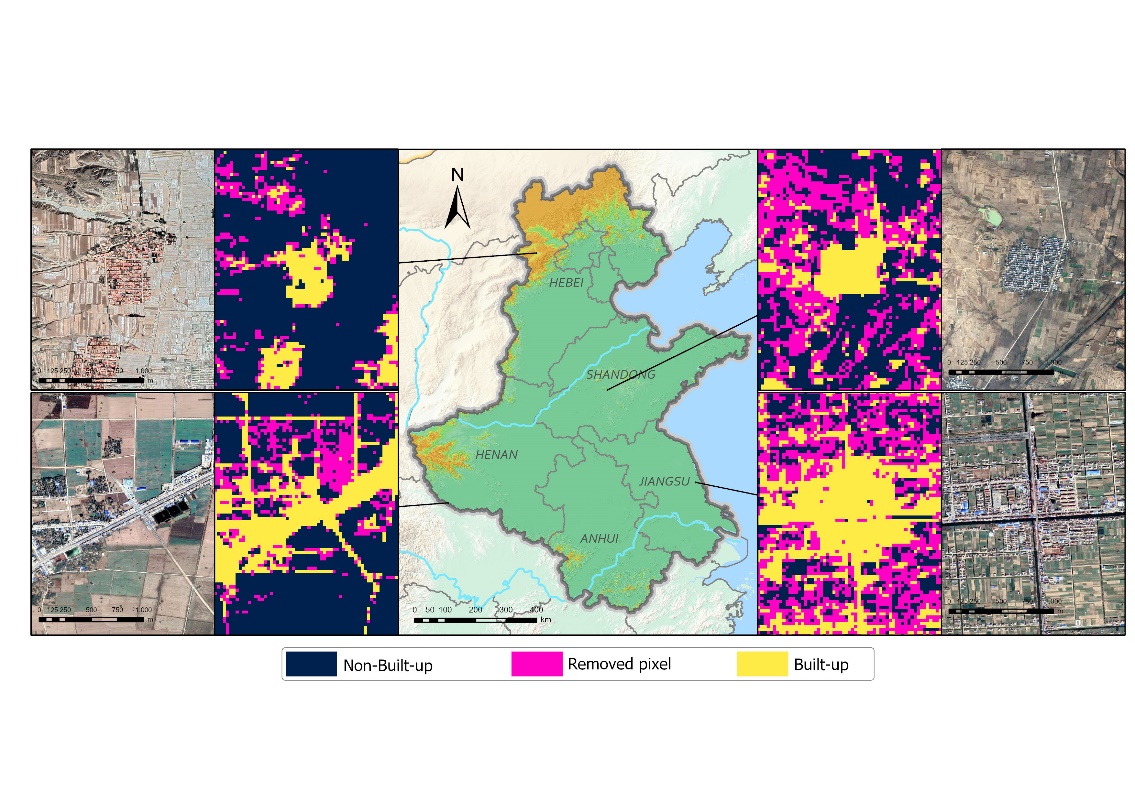
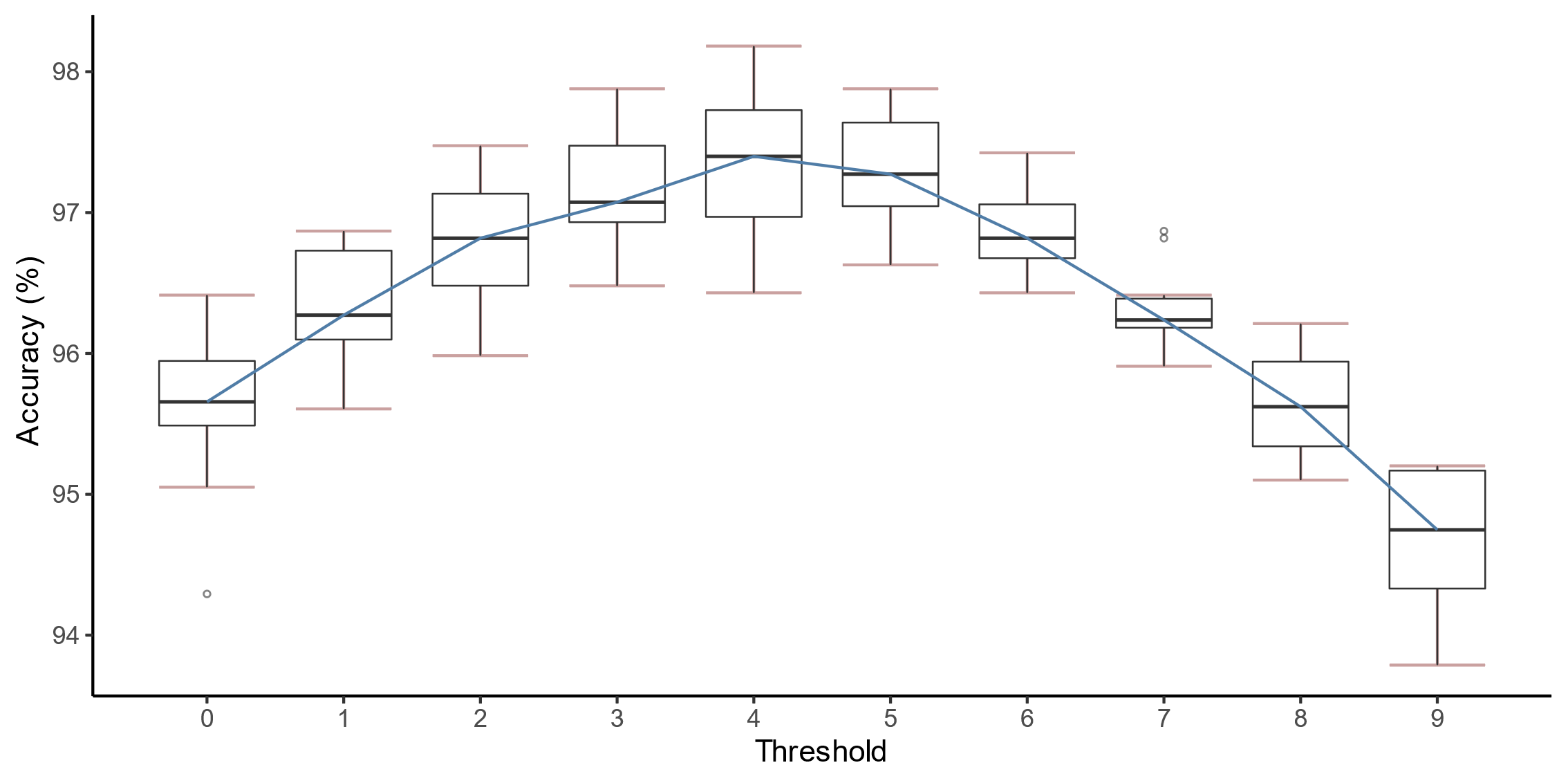
Before applying the Random Forest (RF) classifier for land mapping, we determined the optimal tree number. We used the *sklearn.model\_selection*.*GridSearchCV* module to test the impacts of tree number on accuracy (Figure 8a). We found no accuracy gains were achieved with more than 100 trees. Thus we set the tree number to 100. We also investigated control sample sizes from 0.5% to 99% of the sample and computed corresponding accuracy (Figure 8b). We found that ~50% of the control samples were sufficient to high accuracy. In this study, 75% of the control samples were used for built-up land mapping, among which 70% were used to train the RF classifier. As a result, 52.5% (75% × 70%) of control samples were used to train the RF classifier, which was sufficient for stable classification.



**Figure 8. Accuracy of built-up land mapping with different tree numbers and training sample sizes. Se: standard error computed from 10 classifications.**

## 3.2 Removing uncertainty caused by sample variation

Because different training samples produce different classifications, we repeated the classification 10 times with a different sample splitting state (i.e., a seed number set from 0 to 9). We summed the 10 classifications and created the final classification to be the pixels greater or equal to different thresholds. Figure 9a shows that a threshold of 4 led to the highest accuracy. Figure 9b shows that misclassifications from bare lands or farmland rotations were removed after applying the threshold.



**a)**

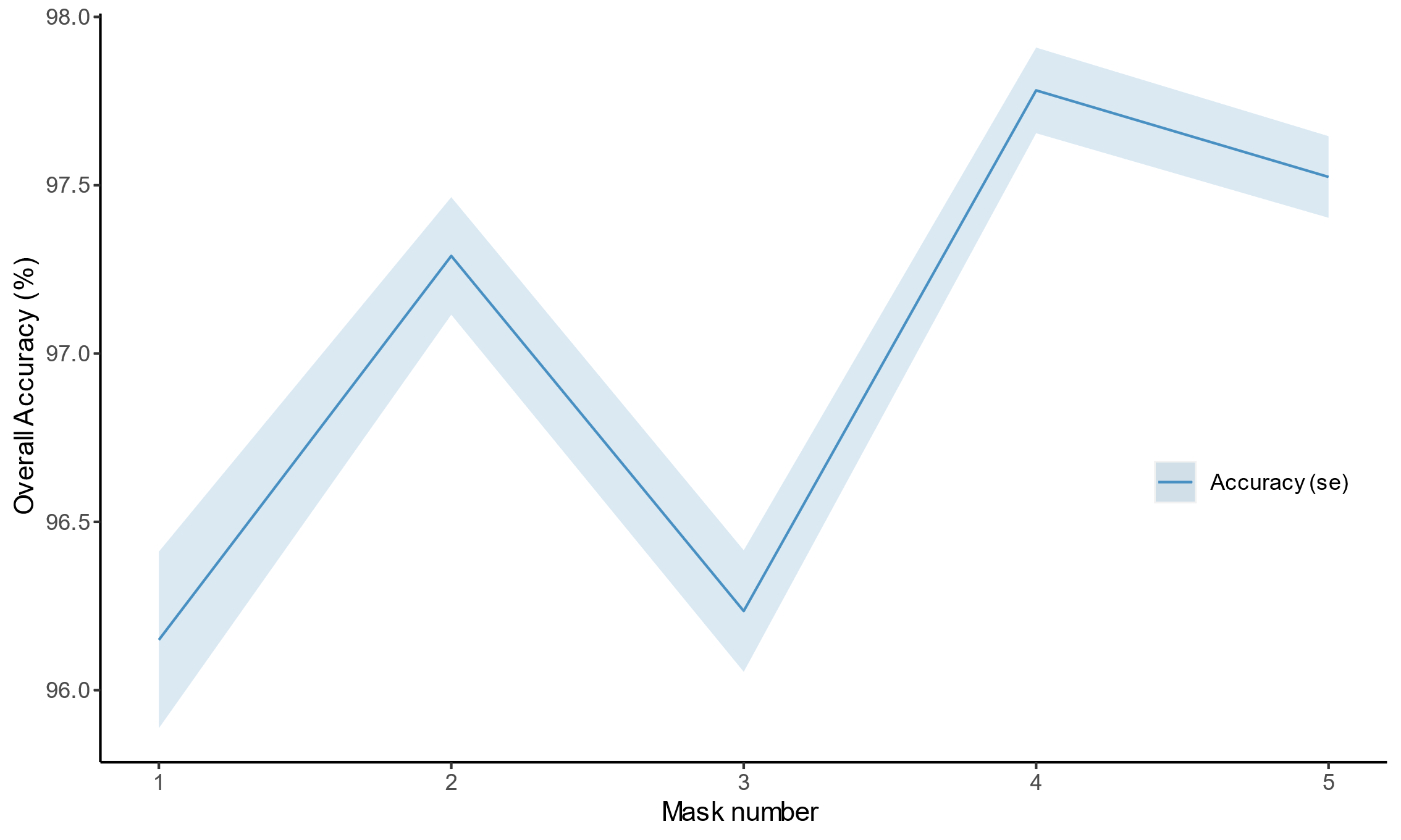
**b)**

**Figure 9. a) Changes in accuracy and b) spatial improvement after removing uncertainties caused by sample variation. Parameters in a) were computed from 10 classifications.**

# 4. Determining mask number and iteration number

## 4.1 Sensitivity analysis of mask number

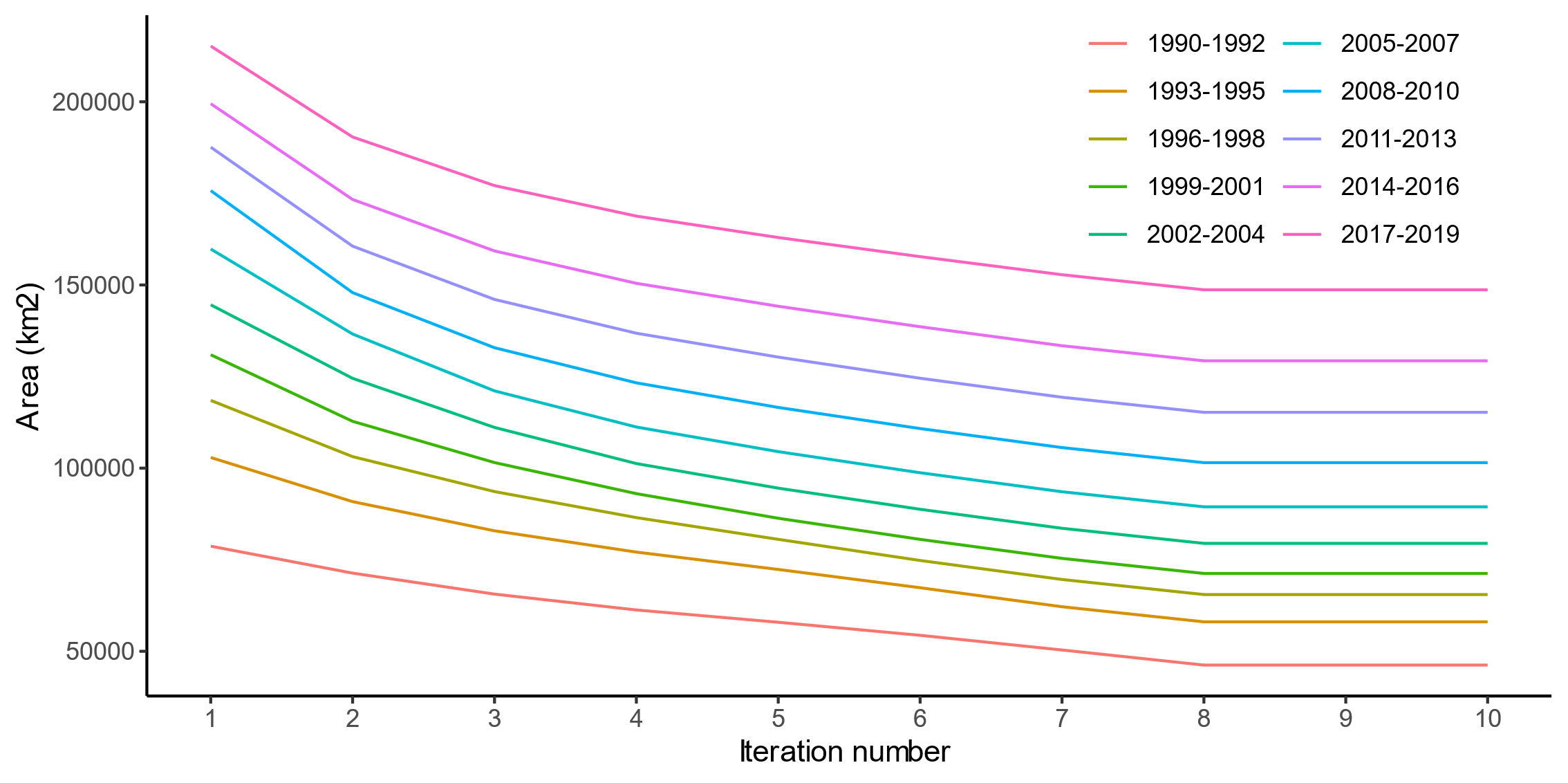
Before applying the temporal correction, the number of classifications (i.e., mask number) used to create the mask should be determined. We ran a sensitivity test to compute overall accuracies for 5 mask numbers (Figure 8). As the mask number increased, overall accuracy improved. We set the mask number to 2 because it is close to the second-highest overall accuracy with fewer classifications used in the mask. If we had set the mask number to 4, the maps of the last four periods would be used as a mask and could not be temporally corrected, which was not practical given there were 10 maps in total.



**Figure 10. The overall accuracy of the built-up land mapping with different mask numbers. The "se" refers to the standard error, which was computed from the ten classifications of 1990-2019.**

## 4.2 Sensitivity analysis of iteration number

After determining the mask number, we ran another sensitivity test to determine the temporal correction's iteration number (Figure 11). We found that after 8 iterations, the built-up area remained stable. This pattern can be found for all classifications from 1990 to 2019. As a result, we determined the iteration number to be 8.



**Figure 11. Built-up area change with different iteration numbers in the temporal correction.**

# References

Gong, P., Li, X., Wang, J., Bai, Y., Chen, B., Hu, T., Liu, X., Xu, B., Yang, J., Zhang, W., & Zhou, Y. (2020). Annual maps of global artificial impervious area (GAIA) between 1985 and 2018. *Remote Sensing of Environment, 236*

Li, X., Gong, P., & Liang, L. (2015). A 30-year (1984–2013) record of annual urban dynamics of Beijing City derived from Landsat data. *Remote Sensing of Environment, 166*, 78-90