Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

A label-noise robust active learning sample collection method for multi-temporal urban land-cover classification and change analysis

Jiayi Li[a], Xin Huang[a,b,*], Xiaoyu Chang[a]

[a] *School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, PR China*
[b] *State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, PR China*

ARTICLE INFO

Keywords:
Machine learning
Multi-temporal change detection
Landsat satellite imagery
Label-noise robust active learning
Multi-date classification
Automatic sample collection

ABSTRACT

Timely and reliable land-use/land-cover (LULC) change dynamic monitoring is the basis of urban understanding and planning. However, either the training sample shortage or the error accumulation in the multi-temporal processing inevitably restricts the monitoring performance. In this paper, to overcome these problems, we present a label-noise robust active learning method, which automatically collects reliable and informative samples from the images and builds a unified classification system with these augmented samples. In more detail, a Bayesian sample collection process that fuses the unsupervised transition information and the multi-temporal land-cover information is designed to provide candidate samples with “from-to” labels. A reliability-based multi-classifier active learning method is then proposed to adaptively allocate the more reliable samples to the classes that are difficulty to classify. Finally, a fusion of the multiple multi-date classifications trained by the selected samples is implemented to identify the change type of interest. The dynamic monitoring results for Shanghai, Shenzhen, and Shiyan in China, two megacities with rapid and obvious urbanization and a small city with relatively slow urbanization, indicate that the proposed method achieves a significantly higher accuracy than the current state-of-the art methods. The sample accuracy verified by the high spatial resolution reference maps endorses the applicability of the sample collection, while the reliability-based active learning further ensures the robustness of the proposed method in the label-noise situation. The presented method was tested in two difficult situations (a small training sample case and a training sample set without joint labeling), so that the robustness and accuracy of the approach can be expected to be of a similar or better quality in cases with more training samples. Given its effectiveness and robustness, the proposed method could be widely applied in LULC change dynamic monitoring.

1. Introduction

Timely and reliable geo-information on the extent of a city is of significant importance for urban growth studies (Li et al., 2015b). In addition to the time and location (Jabari et al., 2019), the categorization of the land-use/land-cover (LULC) dynamic, i.e., the “from-to” types of LULC changes, is important in many applications (Healey et al., 2018). Efficient and accurate multi-temporal land-cover monitoring is vital to facilitate a better understanding of the interaction between mankind and nature, especially under the trend of global urbanization.

Remote sensing has become the primary data source for multi-temporal LULC dynamic monitoring (Gómez et al., 2016). From the perspective of remote sensing processing, the mainstream from-to types of LULC change detection (also called multi-temporal classification) (Li et al., 2015b; Schneider, 2012; Wu et al., 2017; Xu et al., 2018; Yu et al., 2016) can be divided into post-classification and multi-date classification techniques. The post-classification methods produce multi-temporal LULC maps from independent classifications, and the changes are identified by comparing these maps (Masek et al., 2008). Due to its flexibility in organizing (or fusing) the existing multi-temporal classification maps, this kind of technique is popular, especially when dealing with LULC change detection tasks at a national or global scale (Masek et al., 2008; Xian et al., 2009). However, the disadvantages of the post-classification methods can be summarized as follows. Firstly, misclassification in any of the images will be compounded in the final change detection map, and these errors (including illogical land-cover change events) will be further amplified with an increase in the time series. Secondly, as urban LULC changes usually account for only a small part of the study region and are scattered in different locations, the confidence level of the change trajectory

[*] Corresponding author at: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, PR China.
E-mail address: xhuang@whu.edu.cn (X. Huang).

https://doi.org/10.1016/j.isprsjprs.2020.02.022
Received 8 November 2019; Received in revised form 30 January 2020; Accepted 29 February 2020

generated by direct multi-temporal comparison can be low, which restricts the reliability of the further applications (Lu et al., 2011; Schneider, 2012). Although several temporal priors have been investigated to reduce the misclassification, it is still not easy to apply such priors to multi-temporal urban LULC change detection. One major technique in the bi-temporal change detection task is referred to as the consistency checking and updating/backdating approach, in which the threshold to acquire a binary change map cannot be automatically and accurately determined (Li et al., 2015b; Wu et al., 2017; Xian et al., 2009; Xu et al., 2018; Yu et al., 2016). For hypothesized trajectory-based multi-temporal change detection methods (Li et al., 2015b; Schneider and Mertes, 2014; Xue et al., 2014a), the rules that classify LULC changes by the hypothesized trajectory signatures cannot fully portray the heterogeneous urban environment.

The multi-date classification method first extracts training samples from remote sensing time-series data for places that have undergone certain kinds of change as one separate class, and later builds a unified classifier by these samples to identify the change type of interest from the stacked time-series data (Huang et al., 2008; Im and Jensen, 2005; Nemmour and Chibani, 2006; Schneider, 2012). Given an advanced supervised machine learning classification algorithm and enough high-quality training information, a unified classification system can yield a desirable change detection performance (Huang et al., 2008; Nemmour and Chibani, 2006), and can show robustness when the land-cover classes are not normally distributed (Lu et al., 2004), as in a heterogeneous urban environment (Liu and Lathrop Jr, 2002). However, previous surveys have indicated that there are still several issues to be addressed (Gómez et al., 2016; Hussain et al., 2013). Firstly, signature extension by generalization of LULC signatures across time might present additional challenges for training data sampling (Gómez et al., 2016). The collection of sufficient training samples is also costly, which prevents it from being widely used. To tackle this problem, one solution is to generate synthetic training data from laboratory spectra, field surveys, and spectral models (Okujeni et al., 2013; Roy et al., 2019). However, the effect of spectral model processing (e.g., the modeling errors that ignore changes in the vegetation structure, as described by Roy et al. (2019)) and the effects of remote sensing monitoring (e.g., electronic noise, ground topography, variations in the exoatmosphere solar spectra, differences in spectral and radiometric calibration, the bidirectional reflectance distribution function (BDRF), and the adjacency effects mentioned in Ben-Dor et al. (2004)) pose challenges when applying a classifier trained on synthetic data to a real-world scenario. The other solution is to collect samples from the real images directly (Li et al., 2019; Xue et al., 2014b). For instance, Xue et al. (2014b) utilized the phenology trajectory of natural land covers to augment the sample set size. However, a phenology trajectory requires dense time-series data, and cannot be applied to permanent features in an urban environment, such as built-up areas and lakes. Furthermore, the sample redundancy and the label noise of the augmented sample set are not considered, and there is still room for improvement to achieve the goal of using the smallest number of adequate samples to achieve the maximum accuracy.

In the meantime, as argued by Schneider (2012), in view of the dominant position of supervised classification methods in remote sensing research in recent years, developing semi-automated ways to speed up training sample collection holds great promise. Meanwhile, Gómez et al. (2016) also suggested that the advanced machine learning paradigms, such as active learning (Demir et al., 2010; Tuia et al., 2011), could help to reduce the high computational cost of redundant samples, although the current active learning methods in the field of remote sensing image interpretation are sensitive to label noise (Guo et al., 2015; Huang et al., 2015). It should be noted that, although several advanced supervised classifiers (e.g., support vector machine (SVM), decision tree, and neural network classifiers) have been applied to LULC change detection (Gómez et al., 2016), there is still no optimal approach for urban LULC dynamic monitoring.

In this paper, to deal with the aforementioned problems, we present a label-noise robust active learning sample collection method for multi-temporal land-cover classification and change analysis. Firstly, the collection process for the from-to sample set is carried out by accumulating land-cover records of each single date, and then a Bayesian-based purification process is carried out, which uses unsupervised change information to ensure the reliability of the selected samples. Next, a discriminative subset of these newly selected samples is picked by the use of a reliability-based multi-classifier active learning method. Finally, the multi-date classification is implemented by fusing the multi-classifiers conducted from the union of the original training samples and the newly added discriminative ones. Taking a small city (Shiyan in China) with a low level of urbanization and two megacities (Shenzhen and Shanghai in China) with rapid and obvious urbanization in recent years as the study areas, three multi-temporal Landsat datasets were adopted to test the advantages of the proposed approach, in terms of both accuracy and efficiency.

2. Study areas and data description

2.1. Study sites, data sources, and preprocessing

In this study, the three Chinese cities of Shiyan, Shenzhen, and Shanghai (Fig. 1), which feature diverse spatial sizes, were selected for the LULC change monitoring. The city of Shanghai (31°40′ N to 31°53′ N and 120°51 E' to 122°12′ E) is the undisputed leader in economic development in China, with the permanent resident population of Shanghai being 24.18 million at the end of 2017 and the GDP being $444.8 billion in 2017. Shanghai has become the engine of the economic growth for the Yangtze River Delta region, and its incredible run of success has been accompanied by complicated LULC transformation. The city of Shenzhen (113°46′–114°37′E, 22°27′–22°52′N) has transformed from an unknown fishing village to one of the largest cities in the Pearl River Delta since it became China's first special economic zone in 1979. Shenzhen's permanent resident population increased from less than 100,000 in 1979 to over 12 million in 2017, and its GDP reached $328.7 billion in 2017. As a window of China for economic, scientific, and technological exchanges, its natural land covers have a high probability of transition to construction land. In addition to these two typical coastal megacities experiencing rapid urbanization, the city of Shiyan (including Zhangwan and Maojian districts, 110°46′–111°00′E, 32°04′–32°36′N) was also taken as a study area. Shiyan has a GDP of less than $30 billion, with a permanent resident population of 819,100 at the end of 2018. The city of Shiyan is a mountainous city with scarce land for construction and limited urban space. Compared with Shenzhen and Shanghai, there has been much less LULC transformation in Shiyan. To thoroughly evaluate the performance of the proposed method for delineating multi-temporal land-cover dynamics, these three study area were chosen for their diversity in geo-spatial, land-cover transition, and socio-economic aspects (Table 1).

The multi-temporal L1T level Landsat data adopted in this study were downloaded from the U.S. Geological Survey website (USGS http://earthexplorer.usgs.gov/). Cloud cover of less than 90% and similar dates/seasons were set as the image selection criteria. The selection results are listed in Table 1. On the basis of the standard L1 level product, the image mosaicking for two adjacent scenes within one year was first undertaken in the ENVI 5.3 environment, and then the multi-temporal preprocessing, including relative radiometric correction and geometric registration, was further conducted. False-color composites of the three datasets are presented in Fig. 1.

2.2. Land-cover change types and reference sites

We categorized the landscapes in the study areas into four land-cover classes: water, vegetation, built-up, and bare soil. Table 2 lists the multi-temporal (i.e., three-date) land-cover change scheme for each

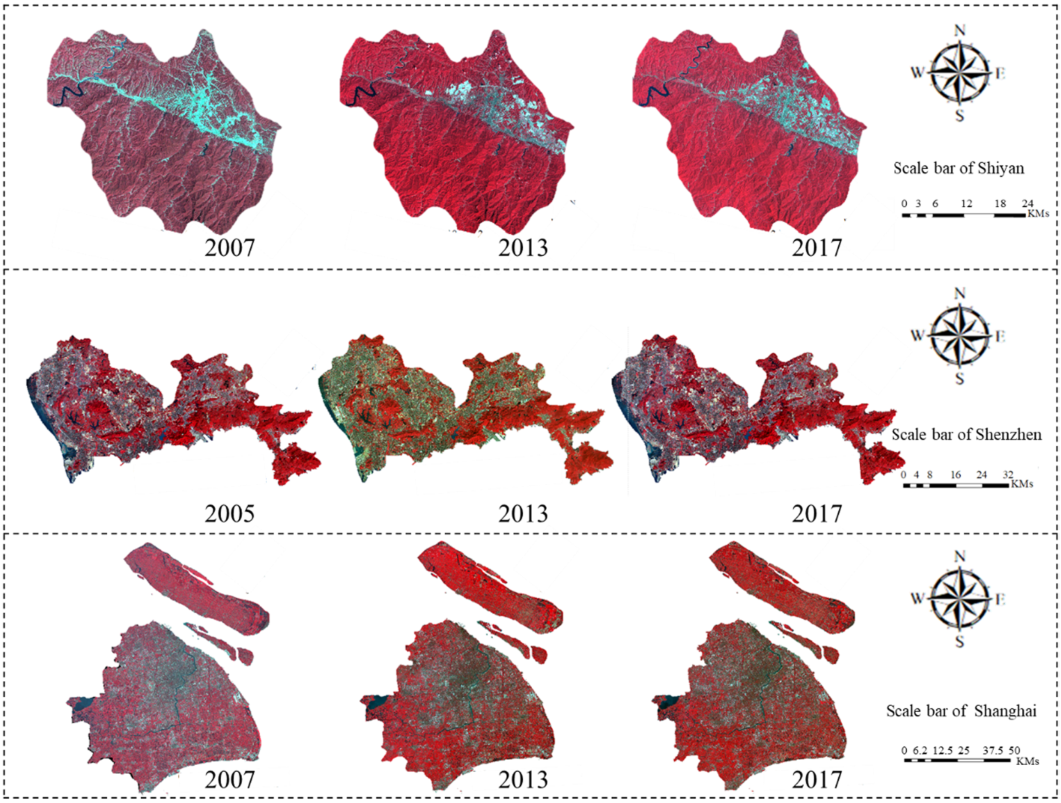

Fig. 1. False-color composites (R: near-infrared; G: red; B: green) of the study areas: Shiyan (the upper three images were acquired in 2007, 2013, and 2017), Shenzhen (the middle three images were acquired in 2005, 2013, and 2017), and Shanghai (the lower three images were acquired in 2007, 2013, and 2017). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

study area. The reference sites for each study area were selected in the following steps. Firstly, according to the semi-variogram based sampling criteria (Chen and Stow, 2002), three regular square grids (i.e., 300 × 300 m for Shiyan, 630 × 630 m for Shenzhen, and 1200 × 1200 m for Shanghai) were employed to divide the study areas into blocks, with the aim being to suppress the spatial autocorrelation. Secondly, the intersection points identified as pure (land-cover) samples for each date were collected as reference sites. Here a "pure" sample is defined as the associated land-cover class comprising > 60% of the pixel area (i.e., within 900 m^2). For Shanghai and Shiyan, the selected reference sites were visually identified by experts with at least two years of experience in remote sensing interpretation. In addition to the high-resolution images from Google™ Earth, a very high spatial resolution image covering Shanghai in 2017, made up of eight Gaofen-2 images (spatial resolution: 0.91 m, Table 3), was adopted as a reference for the manual inspection. For each date (i.e., 2005, 2013, and 2017) for Shenzhen, a thematic map covering the whole city at a 2-m resolution was adopted to determine the detailed land covers. A total of 21 high-resolution remote sensing images were collected to generate the thematic maps of Shenzhen. The satellite sensors considered in this experiment were QuickBird, Ziyuan-3, and Gaofen-2. More details can be found in Table 3. In the image preprocessing step, the raw digital number values of the remote sensing images were converted to surface reflectance with the QUick Atmospheric Correction (QUAC) algorithm in ENVI (Flaash, 2009). The QuickBird satellite images with the original spatial resolution of 1 m were resampled to 2 m. For the Gaofen-2 images, the high-resolution (1 m) panchromatic and lower-resolution (4 m) multispectral imagery were merged using the NNDiffuse pan-sharpening technique (Sun et al., 2014). The created high-resolution (1 m) multispectral imagery was then resampled to 2 m. Finally, the images were stitched together to cover the whole study area, using both edge feathering and reflectance correction (Chon et al., 2010). A supervised classification approach, as well as manual post-processing, was then conducted to transform the very high spatial resolution satellite imagery into thematic maps of the land cover (water, vegetation, bare soil, and built-up). Each thematic map was validated with at least 200 independent polygons, where, in each polygon, one pixel was randomly chosen as a test sample, and its overall accuracy was not less than 98%. Thirdly, a sample allocation scheme which strikes a balance between the concerns of this study (i.e., focusing on the changes related to urbanization) and the proportion of each class was adopted (Olofsson et al., 2014). In more detail, all the samples of the changed classes and all the *stable bare soil* samples (due to their rarity) were retained, and the numbers of *stable built-up*, *stable water surface*, and *stable vegetation* samples were proportionally reduced. Finally, on the basis of this sample allocation, random sampling for each class was carried out. The numbers of final reference sites for each study area are listed in Table 2. In the classification procedure, the reference sites of each city were randomly divided into disjoint sets for training and testing. In addition, please note that with the high-resolution thematic maps of Shenzhen, the accuracy of the automatically collected samples could also be assessed (see Section 4.3).

3. Methodology

The task of supervised classification is to identify the test samples for the most probable categories by learning rules from the training samples collected manually. The proposed method can be viewed as a combination of unsupervised learning, in that it first automatically collects samples from the unlabeled sample and assigns them with pseudo-labels (e.g., U in Fig. 2), and active learning, in that it continually picks the discriminative and reliable sample subset from these collected samples. With the aid of these newly added samples, the multi-date classification ability can be improved. Fig. 2 portrays the

Table 1
Characteristic of the cities and the Landsat data adopted in this study.

City	Geospatial			Socio-economic (Year 2017)			Land-cover transition	Sensor	Date (path/row)
	Coverage (km^2)	Location	Climate	GDP (billion)	Population urbanization rate	Pillars of the industrial structure			
Shiyan	1195	Inland	Subtropical marine	$30	55.1%	Automobile, agriculture, and tourism	Mountainous city with moderate transition from natural landscapes to artificial surfaces	Landsat 5 TM Landsat 8 OLI Landsat 8 OLI	2007-09-15 (125/37) 2013-08-09 (125/37) 2017-09-15 (125/37)
Shenzhen	1985	Coastal	Subtropical marine	$328.7	99.74%	High-tech, logistics, financial, cultural, and creative.	Transition from grassland, cultivated land, and unused land to built-up land.	Landsat 5 TM Landsat 8 OLI Landsat 8 OLI	2005-11-23 (122/44) and 2005-11-16 (121/44) 2013-11-29 (122/44) and 2013-10-05 (121/44) 2017-11-23 (122/44) and 2017-11-01 (121/44)
Shanghai	6800	Coastal	Subtropical monsoon	$444.8	88.06%	Financial, shipping, trade, and high-tech.	Reduction of water and unused land, accompanied with an increase of grassland, woodland, and built-up.	Landsat 5 TM Landsat 8 OLI Landsat 8 OLI	2007-07-28 (118/39) and 2007-07-28 (118/38) 2013-08-29 (118/39) and 2013-08-29 (118/38) 2017-08-24 (118/39) and 2017-08-24 (118/38)

flowchart of the proposed algorithm, which consists of two main steps: ① the Bayesian theory based from-to sample collection, which integrates the multi-temporal change intensity and the land-cover label possibilities (i.e., the soft classification result of each land-cover mapping); and ② a reliability-based multi-classifier active learning technique, which picks the discriminative and reliable sample subset ($\tilde{U}$) from the set (U). Please note that, for every Landsat Level-1 product, a Quality Assessment (QA) band is available, which describes the quality of the pixels within a scene. This can help the user determine the suitability of the scene for classification. Thus, the land parcels identified as unsuitable for land-cover classification in any observation can be excluded before step ②. After the automatic sample collection, the multi-date classification is implemented by the union of these newly added samples and the original ones. The multi-temporal spectral features of each test sample are then independently fed into each basic classifier trained by the augmented training set. The final label of each test sample can then be determined by majority voting on the three independent predictions.

3.1. Bayesian-based sample collection

Under the assumption that multi-temporal land-cover labels should be temporally correlated, the Bayesian-based approach can be formulated as: find the multi-temporal land-cover labels that provide the maximum *a posteriori* probability for the multi-temporal spectral features:

$$max_{c=\{l1,lt,\cdots,lT\}}\{p(l_1,\ l_t,\cdots,l_T|s_1,\ s_t,\cdots,s_T)\} \tag{1}$$

where l_t and $\boldsymbol{s}_t$ refer to the land-cover label (e.g., water) and the spectral feature at the tth date, respectively. c is the from-to class derived from the multi-temporal land-cover transition (e.g., *Water–built up 13* in Table 2, which means water to built-up from the first date of interest to 2013, and unchanged since 2013).

Using Bayesian theory, (1) can be reformulated as:

$$\max_{c=\{l_1,\dots,l_T\}}\left\{\frac{p((s_1,\ \dots,s_T)|(l_1,\ \dots,l_T))p(l_1,\ \dots,l_T)}{p(s_1,\ \dots,s_T)}\right\} \tag{2}$$

where $p(\boldsymbol{s}_1,\dots,\boldsymbol{s}_T)$, which is the probability distribution of the multi-temporal spectral features, is independent of the land-cover labels and can be ignored, as it makes no contribution to the determination of the land-cover transition; $p(\boldsymbol{s}_1,\dots,\boldsymbol{s}_T|l_1,\dots,l_T)$ is the probability of the multi-temporal spectral features under the condition of land-cover transition; and $p(l_1,\dots,l_T)$ is the multi-temporal land-cover transition probability.

In terms of $p(\boldsymbol{s}_1,\dots,\boldsymbol{s}_T|l_1,\dots,l_T)$, by assuming that the probability distribution of the spectral feature only correlates to its land-cover type at the current date, $p(\boldsymbol{s}_1,\dots,\boldsymbol{s}_T|l_1,\dots,l_T)$ can be shortened to $p(l_1|\boldsymbol{s}_1)\dots p(l_T|\boldsymbol{s}_T)$, where $p(l_T|\boldsymbol{s}_T)$, which is the land-cover class probability at the Tth date, is determined by averaging the label land-cover posterior probability of all the basic classifiers.

In terms of $p(l_1,\dots,l_T)$, by assuming that the future land-cover state of a sample can be modeled purely on the basis of the immediate preceding state, Markov chain theory (Kasetkasem and Varshney, 2002) transforms $p(l_1,\dots,l_T)$ to $p(l_T|l_{T\text{-}1})\dots p(l_t|l_{t\text{-}1})\dots p(l_2|l_1)$, where $p(l_t|l_{t\text{-}1})$ is the land-cover transition probability from the $(t-1)$th date to the tth date, which is estimated by the change probability $\rho_{(t-1)\sim t}$:

$$p(l_t|l_{t-1}) = \begin{cases} 1-\rho_{(t-1)_t\sim}, & \text{if } l_t = l_{t-1} \\ \rho_{(t-1)_t\sim}, & \text{if } l_t \neq l_{t-1} \end{cases} \tag{3}$$

In this study, an effective unsupervised change detector named iteratively reweighted multivariate alteration detection (IR-MAD; (Nielsen, 2007)) was utilized to calculate each bi-temporal change possibility $\rho_{(t-1)\sim t}$. IR-MAD is designed to measure the difference of the spectral characteristics acquired at two points in time and covering the same geographical region (e.g., $\boldsymbol{s}_{t-1}$, and $\boldsymbol{s}_t$). As mentioned above, the multi-temporal land-cover transition probability $p(l_1,\dots,l_T)$ is highly

Table 2
The land-cover change types used in each study area and the numbers of reference sites for each change type.

Classification scheme	Caption	Reference sites per class		
Unchanged, stable classes		Shiyan	Shenzhen	Shanghai
Stable water	Stable water surface during the study period	55	177	150
Stable vegetation	Stable vegetation during the study period	78	172	200
Stable bare soil	Stable bare soil during the study period	18	19	10
Stable built up	Stable built-up during the study period	70	123	170
Changed classes				
Water–built up 13	Water to built-up from the first date to 2013, and unchanged since 2013	–	36	–
Water–built up 17	Water to built-up 2013–2017	–	11	–
Vegetation–built up 13	Vegetation to built-up from the first date to 2013, and unchanged since 2013	103	132	231
Vegetation–built up 17	Vegetation to built-up 2013–2017	50	84	134
Bare soil–built up 13	Soil to built-up from the first date to 2013, and unchanged since 2013	52	106	63
Bare soil–built up 17	Soil to built-up 2013–2017	20	59	101

Table 3
List of the high spatial resolution satellite images used in this study.

Image	Date	Sensor	Image	Date	Sensor
Shenzhen 2005	2005/11/29 2005/11/16 2005/12/17 2005/05/02 2005/12/22 2005/09/05 2003/12/07 2002/08/31	QuickBird	Shenzhen 2017	2017/10/29 (4) 2017/02/15 (2) 2016/11/28 (2) 2016/10/10	Gaofen-2
Shenzhen 2013	2013/03/08 2014/02/20 (2) 2014/10/09	Ziyuan-3	Shanghai 2017	2017/03/01 (2) 2017/04/29 (2) 2017/12/12 (4)	Gaofen-2

correlated with the continuous multiplication of adjacent change probabilities, and any possible error in each change probability estimation will be compounded in the final result. Furthermore, if the land-cover classification posterior probability at each date is very high (e.g., 0.7 for each date), the continuous multiplication, which is at risk of degrading the multi-temporal transition possibility (e.g., $0.7^3 = 0.343$), is useless.

Thus, (2) can be reformulated as:

$$\begin{cases} \max_c\{[p(l_1|s_1)\cdot\ldots p(l_T|s_T)]\cdot[p(l_T|l_{T-1})\cdot\ldots p(l_2|l_1)]\} \\ \max_c\{[p(l_1|s_1)\cdot\ldots p(l_T|s_T)]\} \quad \text{if} \min_{t=1,\ldots,T} p(l_t|s_t) > thr \end{cases} \tag{4}$$

where $thr\in(0,1)$ is a threshold to ensure that the minima of the land-cover class probabilities is credible.

3.2. Reliability-based multi-classifier active learning

The core idea of active learning is to iteratively select the most beneficial (i.e., informative) subset from a large sample set. In practice, given a set of samples with known labels, one of the approaches that can maximize the information gain is to select the samples mislabeled by the current classifier as incremental training samples (Tuia et al., 2011). Nevertheless, with regard to the automatically collected samples with pseudo-labels, the risk of importing confused samples calls for the design of a specific active learning approach.

In this study, a large number of samples with pseudo-labels could be collected by the aforementioned approaches. Thus, we built a multi-date multi-classifier model by the use of the training samples, and then iteratively imported reliable and informative samples from the pseudo-set to refine the model until the desired classification performance was reached. During the iteration, the proposed sample inclusion process takes the following aspects into consideration: (1) which samples can be included; (2) how to allocate these samples to each class of interest; (3) are there any techniques to stop the samples from having a negative effect on the multi-temporal image interpretation system; and (4) when to stop the iteration. In the following, the sample selection scheme is described in detail, including the sample inclusion and checking method in each iteration, and the termination conditions.

(1) Which samples can be included?

According to the basic idea of active learning (Tuia et al., 2011), the inclusion of the pixels in the areas of uncertainty of the current multi-date classification model into the training set is able to force the model

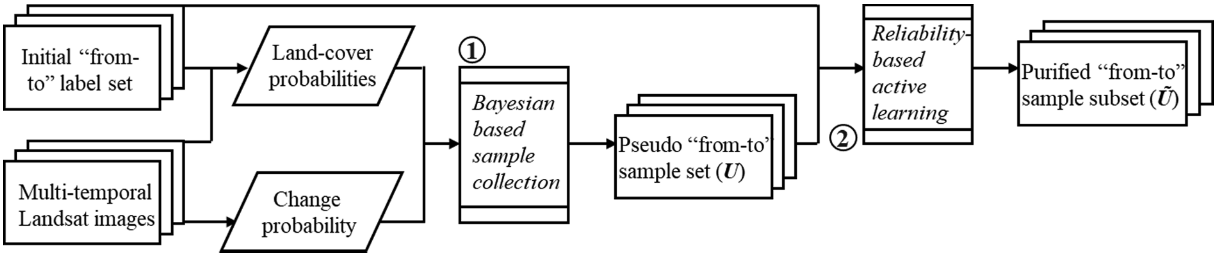
Initial "from-to" label set
Multi-temporal Landsat images
Land-cover probabilities
Change probability
①
Bayesian based sample collection
Pseudo "from-to" sample set (U)
②
Reliability-based active learning
Purified "from-to" sample subset ($\tilde{U}$)

Fig. 2. Flowchart of the proposed automatic sample collection method.

into solving the regions of low classification confidence. In the case of candidate sample set U with label noise, the samples with both high reliability and abundant information are preferred. Thus, in each iteration, high reliability is ensured by consistent labeling by the basic from-to classifiers (i.e., random forest (RF), SVM, and SoftMax), and abundant information is indicated by the conflict between the current multi-date classification result and the pseudo-label, which indicates that the incorporation of this sample should be informative for further refining the multi-date classification model. In this way, the pseudo-samples with both reliability and rich information can be incorporated.

(2) How to allocate the incorporated samples to each class of interest?

Given a set of candidate samples with labels, efficient active learning refers to collecting a small high-quality subset under a low computational cost. In the proposed method, instead of selecting the single most informative sample per iteration, a novel and efficient adaptive batch-mode active learning approach is proposed (i.e., a batch of samples is included at each iteration), which increases the speed of the sample selection and reduces the iterations.

In general, it can be expected that the classification system performance will improve with the iterations (i.e., more and more samples can be identified as reliable) and the growth rate will gradually slow down. Thus, more samples are required in the early stage of the active learning process. In this study, a new adaptive sample set size selection method was designed. For the kth iteration, the sample set reliability (i.e., Rs_k for short) identified by the current classification system is formulated as the number of reliable samples in U divided by the total number of samples in U. The sample set reliability should increase with the iterations (i.e., the growth of reliable samples) and gradually stabilize. By assuming that Rs_k is inversely correlated to the number of newly added samples at the current iteration (i.e., num_k for short), it can be formulated as follows:

$$num_k = (1 - Rs_k^{\hat{\text{A}}½}) \times Card(U) \times C \times N \tag{5}$$

where $Card(U)$ refers to the number of samples in U, C means the number of from-to classes, and N (set as 5 in this study) is the batch scale parameter used to control the size.

Meanwhile, for multi-class active learning, in an iteration, the labeling difficulty of each class can be taken into account in the sample allocation. For instance, a heterogeneous class (e.g., from bare soil to built-up) may need more help from newly added samples than a homogenous class (e.g., stable water surface). Thus, the number of newly added samples for each class (e.g., $num_{k,c}$ for class c) is inversely proportional to the class-wise sample set reliability ($Rs_{k,c}$ for short, which is defined as the number of reliable samples in U of class c divided by the total number of samples in U of class c), and should be summed up to num_k.

We now introduce the technique used to pick the most informative $num_{k,c}$ samples from all the reliable-pseudo samples in U of class c. Firstly, all of these samples are sorted in descending order according to the reliability level, which is formulated as follows (Han et al., 2018):

$$r(s) = \sum_{c=1}^{C} 1/c(\widehat{p}\,(c|s) - \widehat{p}\,(c+1|s)) \tag{6}$$

where $r(s)$ means the transition reliability level of sample s, which belongs to the reliable samples in U of class c; $\widehat{p}\,(c|s)$ represents the average from-to transition posterior probabilities in descending order; and $\boldsymbol{s} = [\boldsymbol{s}_1',\ldots, \boldsymbol{s}_t',\ldots, \boldsymbol{s}_T']$ represents the multi-temporal spectral feature. The first $num_{k,c}$ samples from these reliable samples in U of class c are then selected to be the newly added ones.

(3) Which techniques can stop the samples from having a negative effect on the classification system?

To further ensure the reliability of the newly added samples, a rechecking technique is designed. After the inclusion of these newly added samples, both the multi-classifier system and the current selected sample subset (i.e., $\tilde{U}_k$ for short) are updated. To avoid possible classification system degradation from the newly added but wrongly identified samples, the reliability of the current multi-classifier system (i.e., Rc_k, which is equal to the summation of $r(s)$, $s \in U$) and the reliability of the current selected sample subset (i.e., Rs_k) are simultaneously used to exclude these undesirable samples. In general, during active learning, both reliabilities increase with the growth of the sample set. Otherwise, if both $Rs_{k+1} < Rs_k$ and $Ru_{k+1} < Ru_k$ are met, samples that are added from the kth iteration should be abandoned, and the samples numbered $num_{k,c}$ should be recollected from the rest of U in a similar manner to the above.

(4) When to stop the iteration?

During the iterative process, the reliability of the selected sample subset (Rs) and the multi-classifier system (Rc) is improved and gradually becomes stable. Mathematically, a small value of $|| Rs_{k+1} - Rs_k ||_2/|| Rs_k||_2$, which refers to the difference of reliability between two successive iterations, indicates high stability of the selected sample subset. The difference of Rc can also be modeled in a similar fashion. Accordingly, the termination condition of the iteration refers to the relative stability of both reliabilities:

$$||Rs_{k+1} - -Rs_k||_2/||Rs_k||_2 < \sigma \text{ and } ||Rc_{k+1} - -Rc_k||_2/||Rc_k||_2 < \sigma \tag{7}$$

where σ is a small constant (set as 0.001 in this study) to measure the relative difference. At this point, a lightweight and purified alternative $\tilde{U}$ to U can be successfully collected.

4. Results

4.1. Experimental environment and compared method

In the classification procedure, a random stratification procedure was applied to the reference sites of each city (Table 2) to produce disjoint datasets for training and testing. The total sizes of these two sets followed the ratio of approximately 1:9. When allocating a sample set size to each class, the ratios of the rare change types were set as larger to mitigate the training data imbalance problem. In particular, due to the extreme scarcity of *stable bare soil* in all the reference sites, this kind of training data sample set was about half the size of the others. As suggested by Li et al. (2015a), in the small training sample set task, 10 independent trials were conducted to reduce the possible bias induced by the random sampling.

All the experiments were carried out using MATLAB R2018a on a PC with a single 3.20 GHz processer and 32.0 GB of RAM. To evaluate the performance of the proposed method, the classical multi-date classification technique was carried out as a benchmark. For a fair comparison, each classification step in this benchmark method was implemented with the majority vote from the three trained classifiers of SVM, RF and SoftMax, which were implemented in LibSVM (version 3.23; (Chang and Lin, 2011)) and two MATLAB built-in functions, respectively. As suggested by Li et al. (2019), the kernel of SVM was set as the radial basis function (RBF), and a random subset of $\sqrt{n}$ features was used for the RF classifier at each node, where n is the number of features. In addition, each hyper-parameter of these three classifiers was automatically tuned by 10-fold cross-validation. The classical multi-date classification method using the original from-to labels (hereinafter referred to as FT) identifies the test samples by building a classifier with the original training samples. For the proposed method using the active learning enhanced from-to samples (hereinafter referred to as AL-FT), the sample purification threshold *thr* and the batch active learning scale N were set to 0.7 and 5 for all three Landsat datasets. The sensitivity analyses for these parameters are provided in Section 5.4. The number of iterations was set to 30 for each study area.

4.2. Performance assessment

For each independent trial, the estimated transition error matrix, which records the multi-temporal change detection results and reference labels for every from-to transition type, was employed for the accuracy assessment (Olofsson et al., 2014). In addition to the user's accuracy (UA) and the producer's accuracy (PA), which were used to assess the class-wise performance, the overall accuracy (OA) and the macro-average of the F1 score (MF1 for short; (Zhong et al., 2019)) induced from the transition error matrix were also determined to present a general evaluation. Based on the area proportion of each type in the classification map and the proportion of correct identifications in all the reference sites, the OA estimates the proportion of correct identifications in the multi-date classification map. MF1 is the simple average of all the F1 scores of a single class ($\mathrm{F1}_{\mathrm{class}}$ for short in Table 5), which is the harmonic mean of the PA and UA. As a supplement to the OA, MF1 highlights the identification capability on relatively rare transition types. The mean values of the former three class-wise terms are presented, while the mean ± standard deviation records of the latter two metrics are listed for an overall evaluation. For each metric, a higher value indicates a better performance, and a lower standard deviation signifies a more robust result. In the meantime, the accuracies of the automatically collected samples (i.e., U and $\tilde{U}$) were evaluated with these metrics. To reduce the random bias, all the accuracies were averaged over 10 independent runs.

For estimating the area for each land-cover transition, the tri-temporal change maps of the Shiyan, Shenzhen, and Shanghai Landsat images were generated by majority voting with the 10 results. The estimated area and its approximate 95% confidence interval, which is also on the basis of the transition error matrix mentioned above, were calculated in accordance with the recommendations of Olofsson et al. (2014).

4.3. Assessment of the pseudo-sample set

The evaluation of the pseudo-sample sets is addressed. Firstly, in terms of the Bayesian-based work, it can be seen that more than 55% of each study area can be automatically collected as pseudo-samples, and thousands of samples can be collected for most classes, except for the *stable bare soil* class, with only two or three hundred samples in the Shiyan and Shenzhen datasets (Table 4). In terms of the following active learning work, less than 0.1% of the study area can be picked up, which can reduce the similar and redundant samples collected by the previous Bayesian-based work.

Furthermore, the high spatial resolution (i.e., 2-m resolution) thematic maps of Shenzhen were aggregated to the same resolution as the Landsat images, and the pure samples were used as reference for the sample accuracy assessment. On average, 57.14% ± 0.8% samples of U can be considered as pure, while the ratio for the purified subset $\tilde{U}$ *is* 24.71% ± 2.0%. Considering the uncertainty of the evaluation from the small ratio and the limited size of the purified subset $\tilde{U}$ (i.e., 2039 in Table 4), only the pure samples of U were assessed. The average sample size, PA, UA, and $\mathrm{F1}_{\mathrm{class}}$ for every class and the OA and MF1 for a general assessment are reported in Table 5. From the general assessment, the majority of the collected pure samples are desirable, but there is still moderate label noise. From the class-wise sample assessment, the accuracy of the stable classes (except for *stable bare soil*) is superior to that of the changed classes, while the accuracy of the *stable bare soil* and *water–built up 17* classes is inferior. The impact of these pseudo-samples with some label noise on the multi-temporal change detection is further reported in Section 4.4.

Table 4
Sample set sizes of pseudo-sets U and $\tilde{U}$, by averaging the 10 independent trials.

Class	Shiyan		Shenzhen		Shanghai	
	U	$\tilde{U}$	U	$\tilde{U}$	U	$\tilde{U}$
Stable water	3.51e+03	5	5.74e+4	27	2.19e+5	224
Stable vegetation	8.33e+05	10	6.81e+5	30	1.78e+6	492
Stable bare soil	2.79e+02	21	2.63e+2	74	1.81e+3	338
Stable built up	1.26e+04	41	3.01e+5	292	1.28e+6	1270
Water–built up 13	1.41e+04		1.78e+4	206		
Water–built up 17	1.55e+04		8.87e+3	178		
Vegetation–built up 13	6.43e+03	51	7.18e+4	317	2.42e+5	786
Vegetation–built up 17	6.64e+03	29	1.28e+5	289	5.40e+5	641
Bare soil–built up 13	8.92e+05	63	4.78e+4	250	4.08e+4	988
Bare soil–built up 17	3.51e+03	106	3.81e+4	376	1.04e+5	1301
Total	8.33e+05	326	1.35e+6	2039	4.21e+6	6040
Ratio of set size to study area (%)	67.20	0.03	61.21	0.08	55.89	0.08

Table 5
Accuracy assessment of the pure samples in the pseudo-sample set (U) for the Shenzhen Landsat images.

Class	Num.	PA	UA	$\mathrm{F1}_{\mathrm{class}}$
Stable water	8	93.48%	96.54%	0.9499
Stable vegetation	18	99.81%	88.22%	0.9366
Stable bare soil	5	25.12%	0.32%	0.0063
Stable built up	99	80.93%	80.20%	0.8054
Water–built up 13	37	20.96%	74.17%	0.3164
Water–built up 17	27	8.11%	1.61%	0.0261
Vegetation–built up 13	76	12.72%	73.16%	0.2164
Vegetation–built up 17	43	16.25%	73.45%	0.2641
Bare soil–built up 13	75	12.21%	57.04%	0.2004
Bare soil–built up 17	55	38.88%	50.18%	0.4364
Overall assessment	OA	85.11% ± 0.68%	MF1	0.4158 ± 0.0100

4.4. General change detection results

Table 6 lists the statistical results (PA, UA, $\mathrm{F1}_{\mathrm{class}}$, OA, and MF1) of the change detection results of the FT and AL-FT methods. Meanwhile, the student's *t*-test (Box, 1987), which is a statistical significance test of the difference between two from-to change detection results, is also reported for each metric. "+" denotes that AL-FT performs significantly better, "–" denotes that FT performs significantly better, and "*n*" means no significant difference between AL-FT and FT.

From the perspective of the overall evaluation (Table 6), both the OA and MF1 indicate the significant superiority of AL-FT over FT in all the study areas, and the gap grows dramatically as the size of the study area increases. From the perspective of the class-wise assessment, in addition to UA, PA, and $\mathrm{F1}_{\mathrm{class}}$, the area estimation for each class, which is measured on the tri-temporal change maps generated by majority voting with the 10 independent trials, is adopted as an auxiliary evaluation criteria (Table 7). Figs. 3–5 present the class-wise comparison between FT and AL-FT, from the 10 independent trials.

In terms of stable classes, the proposed method can achieve a satisfactory performance in the *stable vegetation* and *stable water* identification (Fig. 3), while the identification results for the *stable built-up* class are slightly worse, which can be attributed to the higher diversity of the spectral characteristics of artificial structures (Fig. 4). With regard to the terrain materials, both water and dark built-up have strong reflectivity in the green band and strong absorption in the near-infrared band, and both bare soil and bright built-up have strong shortwave infrared channel reflectance but low near-infrared channel reflectance (Li et al., 2019). The overestimation of the *stable water* by FT can be mitigated by the proposed AL-FT method, especially for the Shiyan and Shenzhen datasets (see the left column of Fig. 3 and Table 7). Compared with FT, the diverse artificial structures can be better identified by the proposed AL-FT method, as can be seen in the old residences (the parcel bordered in black in Fig. 4a), the large low-rise industrial area, and the compact middle-rise industrial area (the two local parcels in Fig. 4c).

Table 6
Accuracy assessment of the multi-temporal change detection classification results.

City		PA			UA			$F1_{class}$	
	FT	AL-FT	*t*-test[a]	FT	AL-FT	*t*-test	FT	AL-FT	*t*-test
Shiyan									
Stable water	83.35%	68.64%	*n*	94.96%	99.11%	+	0.8684	0.7669	*n*
Stable vegetation	99.81%	99.57%	*n*	95.12%	97.32%	+	0.9741	0.9843	+
Stable bare soil	92.88%	28.20%	–	29.16%	45.75%	*n*	0.4416	0.2659	–
Stable built up	46.72%	55.96%	*n*	78.52%	93.29%	+	0.5771	0.6926	+
Vegetation–built up 13	58.85%	65.67%	*n*	94.31%	95.64%	*n*	0.7113	0.7752	*n*
Vegetation–built up 17	77.39%	100.0%	+	99.30%	91.62%	–	0.8486	0.9538	*n*
Bare soil–built up 13	34.15%	59.51%	+	67.63%	61.64%	*n*	0.4201	0.5895	+
Bare soil–built up 17	54.08%	84.09%	+	28.90%	21.94%	*n*	0.3535	0.3417	*n*
OA	89.95% 1.30%	91.79% 1.19%	+	MF1			0.6493 0.0217	0.6712 0.0500	+
Shenzhen									
Stable water	93.52%	80.24%	–	96.17%	99.49%	+	0.9470	0.8866	–
Stable vegetation	97.62%	96.55%	*n*	97.30%	100.0%	+	0.9744	0.9824	*n*
Stable bare soil	79.35%	56.38%	–	49.01%	90.00%	+	0.5945	0.7513	+
Stable built up	89.05%	88.69%	*n*	79.02%	91.38%	+	0.8366	0.9000	+
Water–built up 13	64.81%	77.33%	+	59.42%	62.85%	*n*	0.6050	0.6909	+
Water–built up 17	65.53%	92.37%	+	35.50%	24.14%	–	0.5011	0.3774	–
Vegetation–built up 13	49.70%	79.54%	+	88.44%	87.38%	*n*	0.6325	0.8309	+
Vegetation–built up 17	92.13%	98.05%	+	78.17%	80.26%	*n*	0.8421	0.8809	*n*
Bare soil–built up 13	50.47%	75.50%	+	88.06%	88.84%	*n*	0.6303	0.8115	+
Bare soil–built up 17	74.23%	88.40%	+	66.82%	77.70%	+	0.6951	0.8252	+
OA	86.21% ± 1.43%	90.49% ± 1.46%	+	MF1			0.7286 ± 0.025	0.7942 ± 0.025	+
Shanghai									
Stable water	100.0%	97.07%	–	98.73%	99.43%	+	0.9936	0.9820	*n*
Stable vegetation	99.62%	97.56%	–	95.77%	99.84%	+	0.9764	0.9868	+
Stable bare soil	86.24%	80.78%	*n*	43.50%	71.67%	+	0.5492	0.7395	+
Stable built up	90.32%	98.95%	+	79.31%	89.92%	+	0.8435	0.9421	+
Vegetation–built up 13	46.95%	83.87%	+	92.44%	95.72%	+	0.6083	0.8930	+
Vegetation–built up 17	82.52%	98.35%	+	97.03%	95.06%	*n*	0.8880	0.9664	+
Bare soil–built up 13	54.88%	70.51%	+	53.99%	86.41%	+	0.5333	0.7692	+
Bare soil–built up 17	46.22%	64.64%	+	55.84%	74.92%	+	0.4968	0.6880	+
OA	86.34% ± 1.90%	94.33% ± 0.59%	+	MF1			0.7361 ± 0.030	0.8709 ± 0.023	+

[a] whether there is a significant difference (with 95% CI) between FT and AL-FT: " + " denotes that AL-FT performs significantly better, "–" denotes that FT performs significantly better, and "n" means no significant difference.

Table 7
Area estimation of each class in the multi-temporal change detection maps generated by majority voting with the 10 independent trials.

	City	Shiyan		Shenzhen		Shanghai	
Class		FT	AL-FT	FT	AL-FT	FT	AL-FT
Stable classes (km^2)	Stable water	54.77 ± 2.74	35.01 ± 4.25	113.80 ± 3.55	85.94 ± 8.17	418.52 ± 7.71	387.73 ± 5.07
	Stable vegetation	912.16 ± 39.71	909.01 ± 32.51	848.24 ± 16.47	811.36 ± 13.57	3010.90 ± 83.77	2538.18 ± 30.36
	Stable bare soil	31.13 ± 10.69	9.26 ± 4.91	26.79 ± 6.68	10.08 ± 4.24	46.86 ± 16.20	11.85 ± 5.11
	Stable built up	77.79 ± 25.46	78.28 ± 24.02	474.70 ± 35.20	448.21 ± 27.87	1780.03 ± 99.68	1752.66 ± 92.71
Changed classes (km^2)	Water–built up 13			32.63 ± 9.90	48.62 ± 12.72		
	Water–built up 17			7.60 ± 2.70	16.59 ± 6.68		
	Vegetation–built up 13	50.73 ± 23.29	64.38 ± 23.71	144.48 ± 28.89	168.73 ± 22.71	447.99 ± 83.71	642.15 ± 48.76
	Vegetation–built up 17	27.79 ± 23.19	46.35 ± 2.54	171.51 ± 17.36	218.19 ± 17.10	410.82 ± 63.88	888.93 ± 30.36
	Bare soil–built up 13	34.68 ± 10.25	34.16 ± 7.66	105.56 ± 24.39	100.44 ± 15.58	247.38 ± 53.73	182.54 ± 42.76
	Bare soil–built up 17	6.09 ± 3.83	18.71 ± 6.79	60.22 ± 7.16	77.36 ± 12.33	438.98 ± 79.21	397.44 ± 80.49

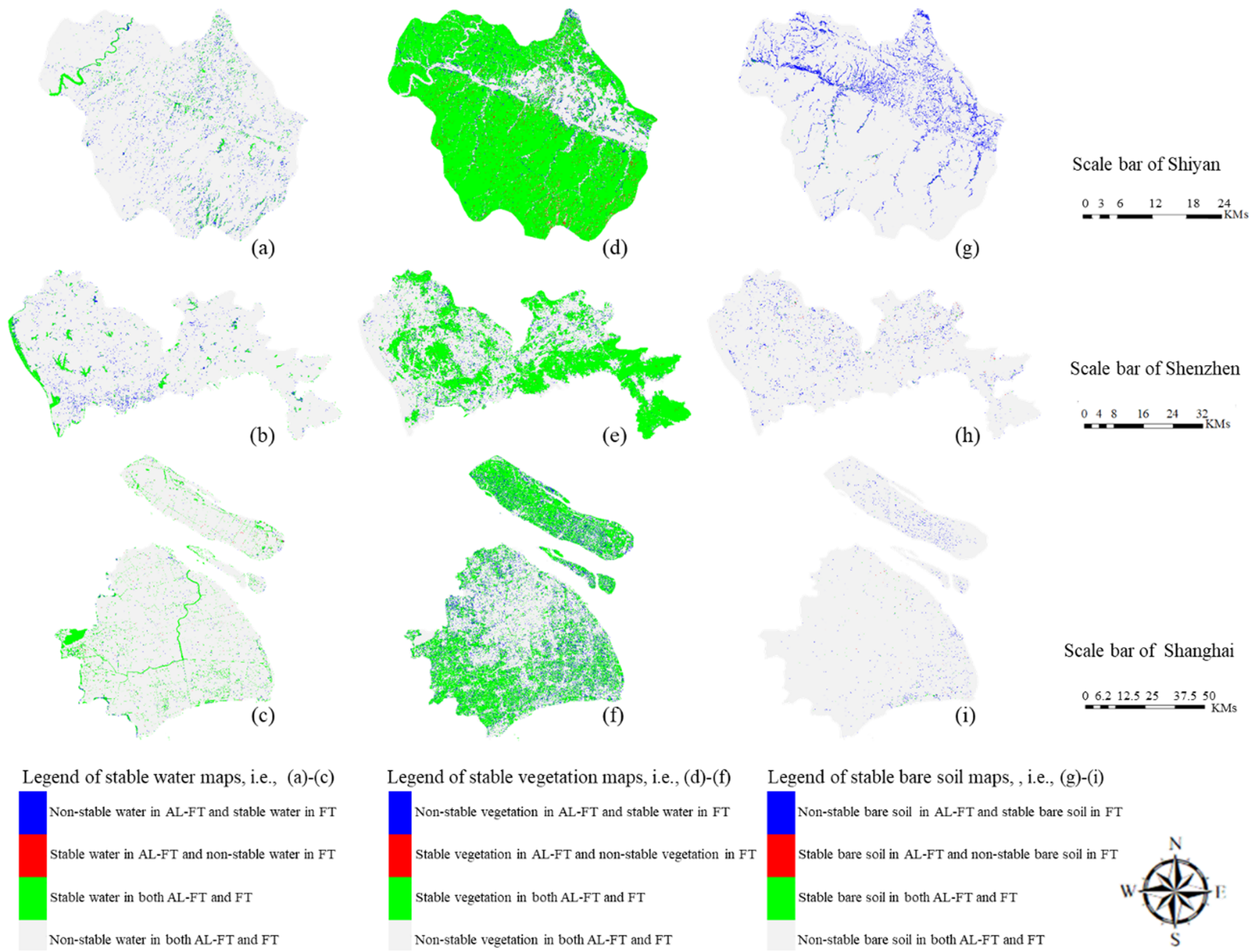
(a)
(d)
(g)
Scale bar of Shiyan
0 3 6 12 18 24 KMs
(b)
(e)
(h)
Scale bar of Shenzhen
0 4 8 16 24 32 KMs
(c)
(f)
(i)
Scale bar of Shanghai
0 6.2 12.5 25 37.5 50 KMs
Legend of stable water maps, i.e., (a)-(c)
Non-stable water in AL-FT and stable water in FT
Stable water in AL-FT and non-stable water in FT
Stable water in both AL-FT and FT
Non-stable water in both AL-FT and FT
Legend of stable vegetation maps, i.e., (d)-(f)
Non-stable vegetation in AL-FT and stable water in FT
Stable vegetation in AL-FT and non-stable vegetation in FT
Stable vegetation in both AL-FT and FT
Non-stable vegetation in both AL-FT and FT
Legend of stable bare soil maps, , i.e., (g)-(i)
Non-stable bare soil in AL-FT and stable bare soil in FT
Stable bare soil in AL-FT and non-stable bare soil in FT
Stable bare soil in both AL-FT and FT
Non-stable bare soil in both AL-FT and FT
N
W
E
S

Fig. 3. The identification of the *stable water* (left column), *stable vegetation* (middle column), and *stable bare soil* (right column) classes for the Shiyan (upper), Shenzhen (middle), and Shanghai (lower) datasets.

The last stable class is *stable bare soil*, which is a rare but important class in such megacities with a high cost of land. For the proposed AL-FT method, the recognition ability for *stable bare soil* also grows as the size of study area increases, which can be attributed to the growth of the number of automatically collected samples of this class (Table 4). Although the correctness of these added *stable bare soil* samples is not high (Table 5), the accuracy gains by these samples (Table 6) validate the applicability of AL-FT for the recognition of the *stable bare soil* class. For the FT method, the inferior UA and the overestimation of the *stable bare soil* class (see the blue parcels in the right column of Fig. 3) partly results from the overweighting of the training samples of *stable bare soil* (i.e., the ratio between training and testing is 1:1, which is much larger than the ratio for the other types). With the aid of the pseudo-samples, the proposed AL-FT method can overcome this shortcoming and obtain a robust area estimation (Table 7). For instance, the mislabeling of the maintenance road and the ridge of the paddy field (covered by *stable bare soil*, see the upper examples of Fig. 4a and 4b) and the mislabeling of wetland (see the lower example of Fig. 4b) by FT can be mitigated by AL-FT.

For both methods, the accuracies (i.e., PA, UA, and F1$_{\text{class}}$ in Table 6) and the coefficient of variation values of the area estimation (i.e., the ratio of the standard deviation to the mean values in Table 7) for the changed classes are inferior to those for *stable water* and *stable vegetation*, which can be attributed to the feature diversity of these changed types, i.e., the combination of the spectral confusion mentioned previously and the multi-temporal transition. Thus, the identification of the transition from water to built-up and from bare soil to built-up is challenging. In terms of the transition from vegetation to built-up, most metrics (Table 6) and the visual displays (Fig. 5b–d) show the desirable improvement achieved by the incorporation of the pseudo-samples. Although the proposed method suffers more from the stable dark built-up false alarms (Fig. 5a), in the comprehensive consideration of both the performance of the transition from water to built-up and that of the stable built-up (Table 6), the proposed method still performs well. In terms of the transition from bare soil to built-up, the proposed method can obtain a more accurate temporal accuracy, as in the industrial land expansion in Shiyan (Fig. 5e), the harbor construction in Shenzhen (Fig. 5g), and the airport construction in Shanghai (Fig. 5g). In summary, with the aid of the pseudo-samples, the results demonstrate the robustness of the proposed approach to label noise (see Table 5) and its applicability in both changed and stable regions.

5. Discussion

One of the main bottlenecks of multi-date classification based from-to change detection is the lack of efficient multi-temporal joint labeling training samples. In addition to the comparisons presented in the previous section, we further discuss the solution to this problem in this section, from the aspects of the necessity for multi-temporal joint labeling, the adaptability of the sample allocation, and the robustness of the parameters used in the proposed method.

5.1. Does the proposed multi-date classification based method have to use the multi-temporal joint labeling training set?

Considering the high cost of multi-temporal joint labeling, we further constructed another training set to investigate the necessity of

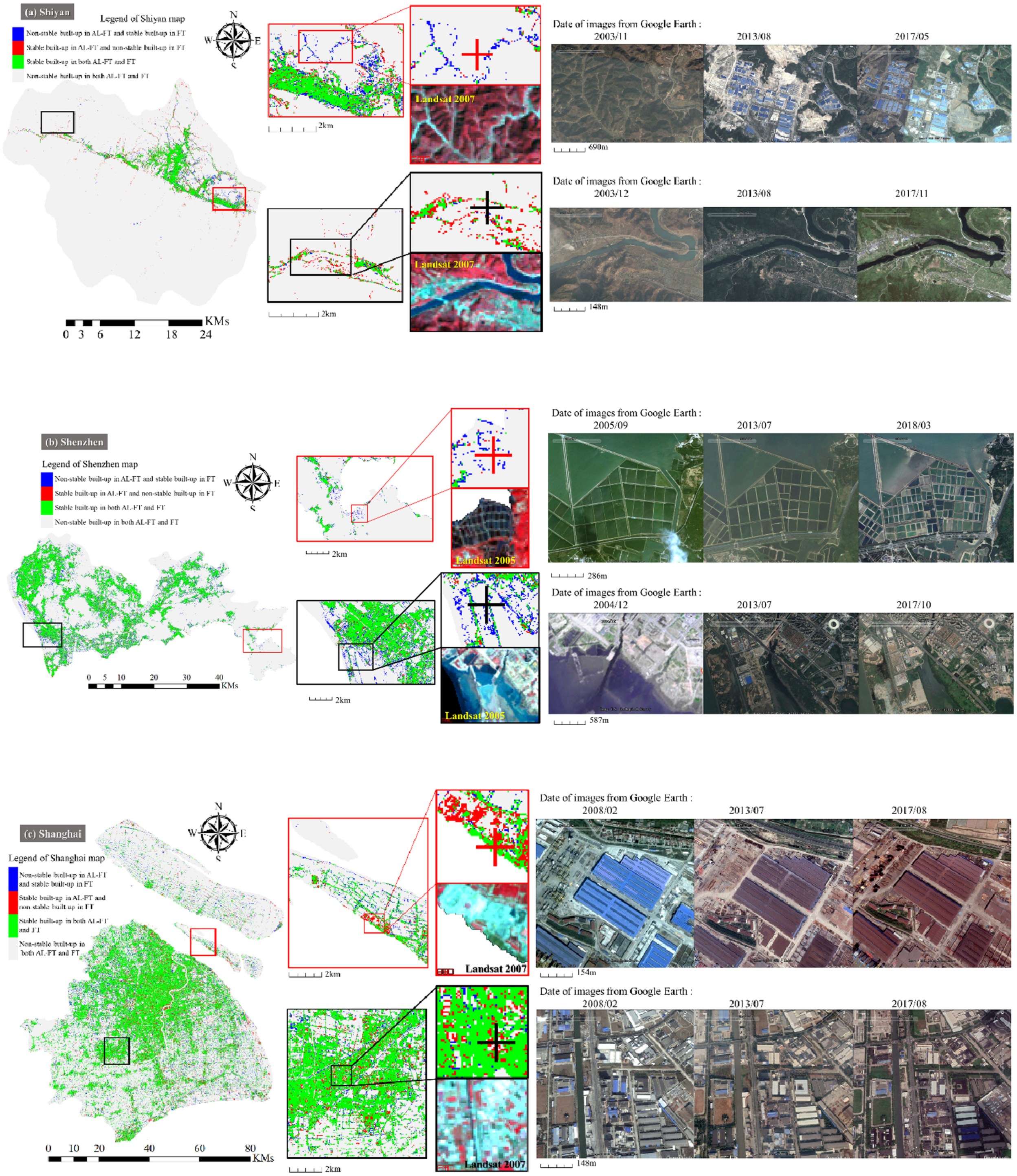
(a) Shiyan
Legend of Shiyan map
Non-stable built-up in AL-FT and stable built-up in FT
Stable built-up in AL-FT and non-stable built-up in FT
Stable built-up in both AL-FT and FT
Non-stable built-up in both AL-FT and FT
N W E S
KMs 0 3 6 12 18 24
2km
Landsat 2007
Date of images from Google Earth :
2003/11 2013/08 2017/05
690m
2003/12 2013/08 2017/11
148m
(b) Shenzhen
Legend of Shenzhen map
Non-stable built-up in AL-FT and stable built-up in FT
Stable built-up in AL-FT and non-stable built-up in FT
Stable built-up in both AL-FT and FT
Non-stable built-up in both AL-FT and FT
0 5 10 20 30 40 KMs
Landsat 2005
2005/09 2013/07 2018/03
286m
2004/12 2013/07 2017/10
587m
(c) Shanghai
Legend of Shanghai map
Non-stable built-up in AL-FT and stable built-up in FT
Stable built-up in AL-FT and non-stable built-up in FT
Stable built-up in both AL-FT and FT
Non-stable built-up in both AL-FT and FT
0 10 20 40 60 80 KMs
Landsat 2007
2008/02 2013/07 2017/08
154m
148m

Fig. 4. The identification of the *stable built-up* class for the (a) Shiyan, (b) Shenzhen, and (c) Shanghai datasets, each of which is marked with two small patches as examples. Left: identification results in the study area; middle: examples with zoomed-in local regions and the aligned false-color composite of the Landsat image at the start date; and right: the very high spatial resolution Google Earth images, whose spatial coverage is the area marked by the cross in the zoomed-in region, which were taken on the nearest date to the associated Landsat images.

multi-temporal joint labeling. This new training set without joint labeling was constructed by randomly selecting the samples from the reference sites, in which the size of each land-cover type at each date was the same as the corresponding size of the joint labeling training set used in the results section. Thus, both training sets had the same amount of LULC label information at each date, and the major difference was the location restriction. In other words, the joint labeling samples should be labeled with the land-cover type on each date, and the non-joint labeling samples are only required to be labeled with the land-cover type on a specific date. Considering the unclear observations of Landsat imagery (e.g., clouds, cloud shadows, snow/ice, and SLC-off data, (Hu et al., 2018)) and the rare existence of some important categories (e.g., *stable bare soil*), the sample collection for the joint labeling set is more costly.

While the traditional multi-date classification method (FT) was not applicable when using the non-joint labeling set, we now discuss

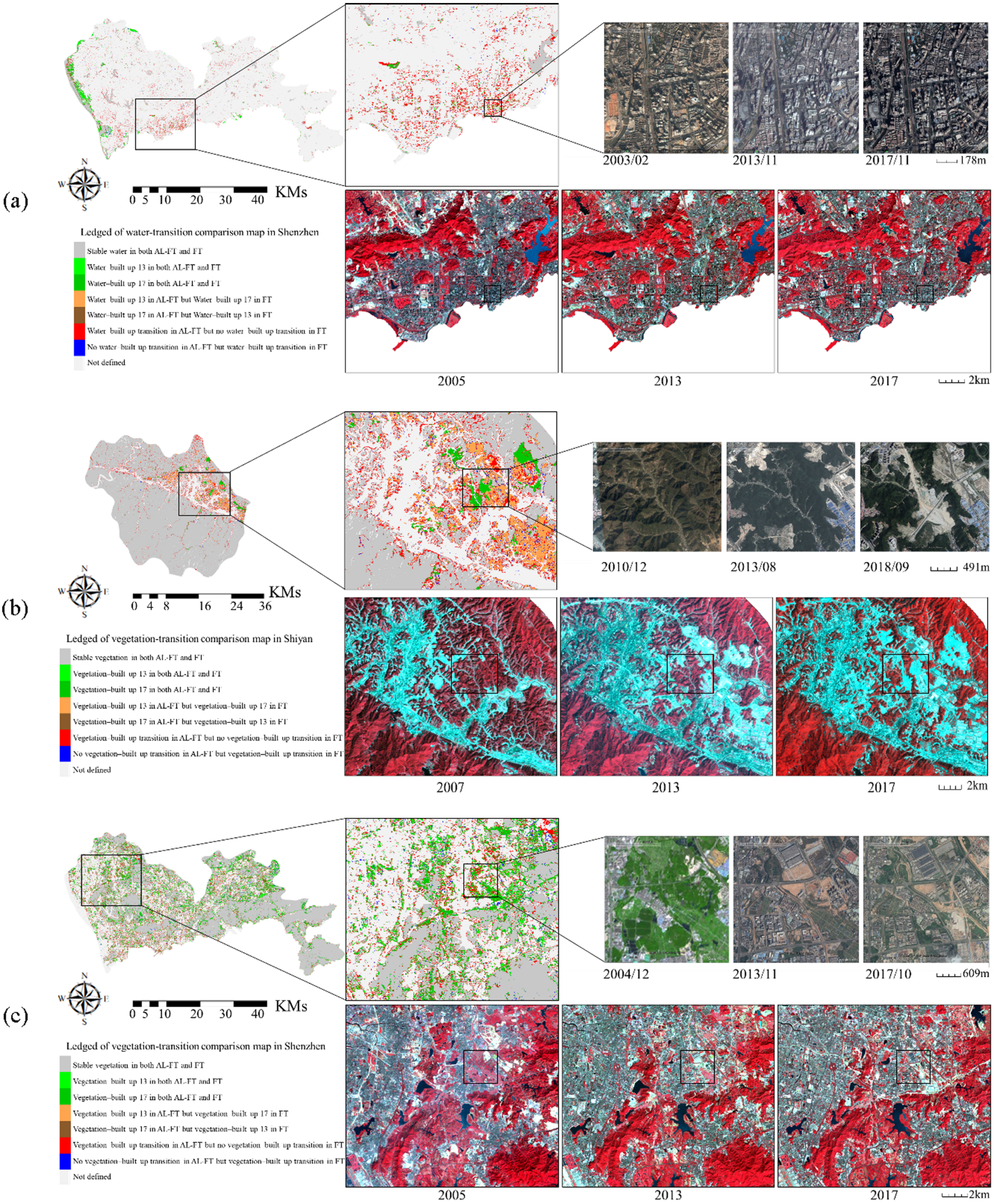
(a)
N
W
E
S
0 5 10 20 30 40
KMs
Ledged of water-transition comparison map in Shenzhen
⬜ Stable water in both AL-FT and FT
🟩 Water built up 13 in both AL-FT and FT
🟩 Water–built up 17 in both AL-FT and FT
🟧 Water built up 13 in AL-FT but Water built up 17 in FT
🟫 Water–built up 17 in AL-FT but Water–built up 13 in FT
🟥 Water built up transition in AL-FT but no water built up transition in FT
🟦 No water–built up transition in AL-FT but water–built up transition in FT
⬜ Not defined
2003/02
2013/11
2017/11
178m
2005
2013
2017
2km
(b)
N
W
E
S
0 4 8 16 24 36
KMs
Ledged of vegetation-transition comparison map in Shiyan
⬜ Stable vegetation in both AL-FT and FT
🟩 Vegetation–built up 13 in both AL-FT and FT
🟩 Vegetation–built up 17 in both AL-FT and FT
🟧 Vegetation–built up 13 in AL-FT but vegetation–built up 17 in FT
🟫 Vegetation–built up 17 in AL-FT but vegetation–built up 13 in FT
🟥 Vegetation–built up transition in AL-FT but no vegetation–built up transition in FT
🟦 No vegetation–built up transition in AL-FT but vegetation–built up transition in FT
⬜ Not defined
2010/12
2013/08
2018/09
491m
2007
2013
2017
2km
(c)
N
W
E
S
0 5 10 20 30 40
KMs
Ledged of vegetation-transition comparison map in Shenzhen
⬜ Stable vegetation in both AL-FT and FT
🟩 Vegetation built up 13 in both AL-FT and FT
🟩 Vegetation–built up 17 in both AL-FT and FT
🟧 Vegetation built up 13 in AL-FT but vegetation built up 17 in FT
🟫 Vegetation–built up 17 in AL-FT but vegetation–built up 13 in FT
🟥 Vegetation built up transition in AL-FT but no vegetation built up transition in FT
🟦 No vegetation–built up transition in AL-FT but vegetation–built up transition in FT
⬜ Not defined
2004/12
2013/11
2017/10
609m
2005
2013
2017
2km

Fig. 5. The identification of the changed classes for all three Landsat datasets. (a) The transition from water to built-up for the Shenzhen dataset; (b)–(d) the transition from vegetation to built-up for the three Landsat datasets; and (e)–(g) the transition from bare soil to built-up for the three Landsat datasets. In each subfigure, left: identification of the changed classes for every Landsat dataset. Upper-middle: the transition result in the local region, which is marked by the dark square in the left subfigure. Lower: the false-color composites of the Landsat images acquired at three dates in the local region. Upper right: the zoomed-in examples with the very high spatial resolution Google Earth images taken on the nearest date to the associated Landsat images.

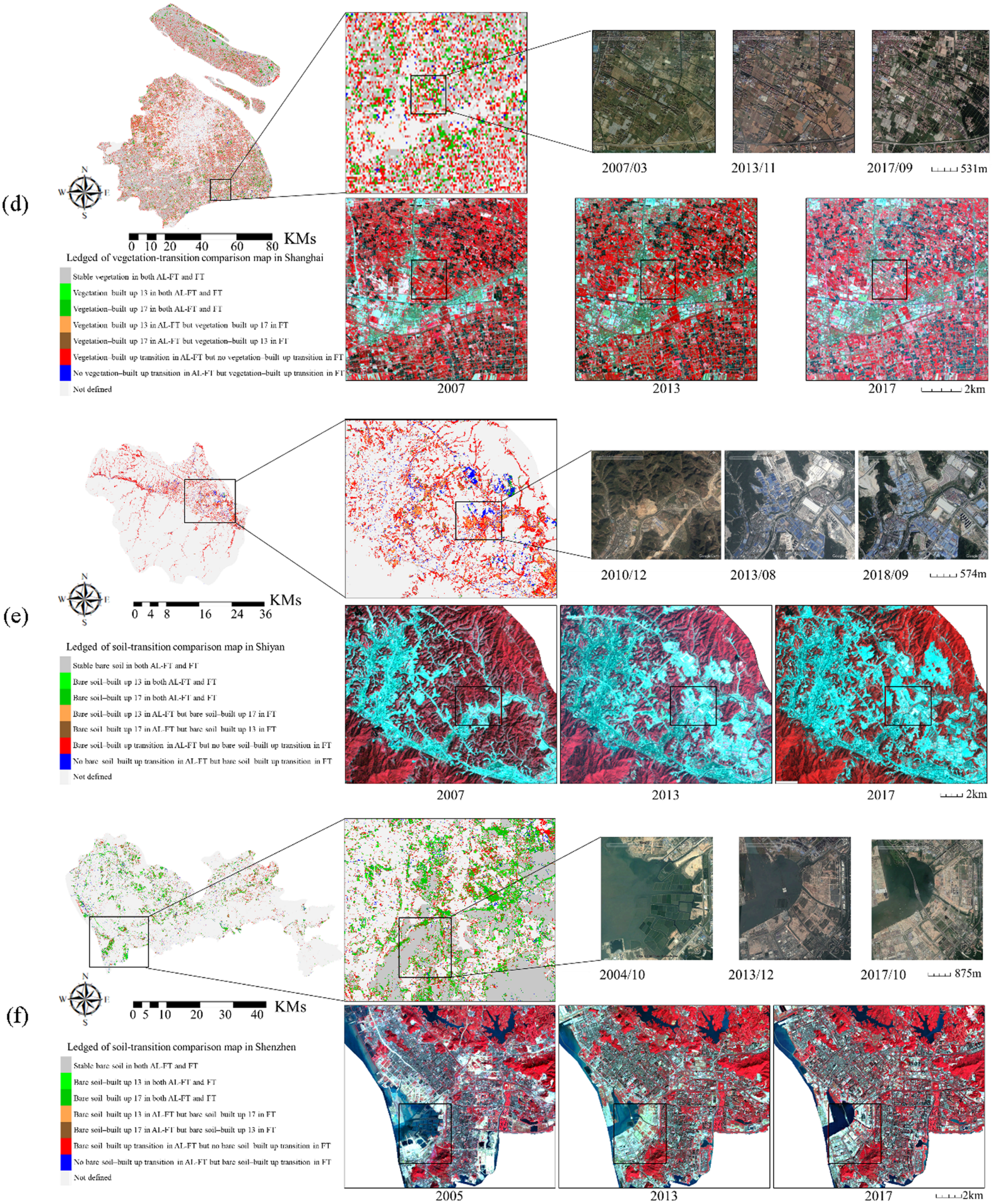
(d)
N
W
E
S
0 10 20 40 60 80 KMs
Ledged of vegetation-transition comparison map in Shanghai
Stable vegetation in both AL-FT and FT
Vegetation built up 13 in both AL-FT and FT
Vegetation–built up 17 in both AL-FT and FT
Vegetation built up 13 in AL-FT but vegetation built up 17 in FT
Vegetation–built up 17 in AL-FT but vegetation–built up 13 in FT
Vegetation–built up transition in AL-FT but no vegetation–built up transition in FT
No vegetation–built up transition in AL-FT but vegetation–built up transition in FT
Not defined
2007/03
2013/11
2017/09
531m
2007
2013
2017
2km
(e)
N
W
E
S
0 4 8 16 24 36 KMs
Ledged of soil-transition comparison map in Shiyan
Stable bare soil in both AL-FT and FT
Bare soil–built up 13 in both AL-FT and FT
Bare soil–built up 17 in both AL-FT and FT
Bare soil–built up 13 in AL-FT but bare soil–built up 17 in FT
Bare soil built up 17 in AL-FT but bare soil built up 13 in FT
Bare soil–built up transition in AL-FT but no bare soil–built up transition in FT
No bare soil built up transition in AL-FT but bare soil built up transition in FT
Not defined
2010/12
2013/08
2018/09
574m
2007
2013
2017
2km
(f)
N
W
E
S
0 5 10 20 30 40 KMs
Ledged of soil-transition comparison map in Shenzhen
Stable bare soil in both AL-FT and FT
Bare soil–built up 13 in both AL-FT and FT
Bare soil built up 17 in both AL-FT and FT
Bare soil built up 13 in AL-FT but bare soil built up 17 in FT
Bare soil–built up 17 in AL-FT but bare soil–built up 13 in FT
Bare soil built up transition in AL-FT but no bare soil built up transition in FT
No bare soil–built up transition in AL-FT but bare soil–built up transition in FT
Not defined
2004/10
2013/12
2017/10
875m
2005
2013
2017
2km

Fig. 5. (*continued*)

whether the proposed method can overcome this limitation. In addition to FT, two post-classification techniques were also taken as comparison methods, which were both applicable when using either training set. The first classical method directly stacks three single-date land-cover records (hereinafter referred to as PCC), and the second method refers to a multi-temporal extension of a recent method based on bi-temporal change probability analysis and Bayesian soft fusion (Wu et al., 2017) (hereinafter referred to as BPCC). The results for both training sets were then evaluated by a unified test set (Table 8), which was a complementary one to the union of both training sets. Please note that the samples selected for the non-joint labeling set at each date were not included in the complementary set.

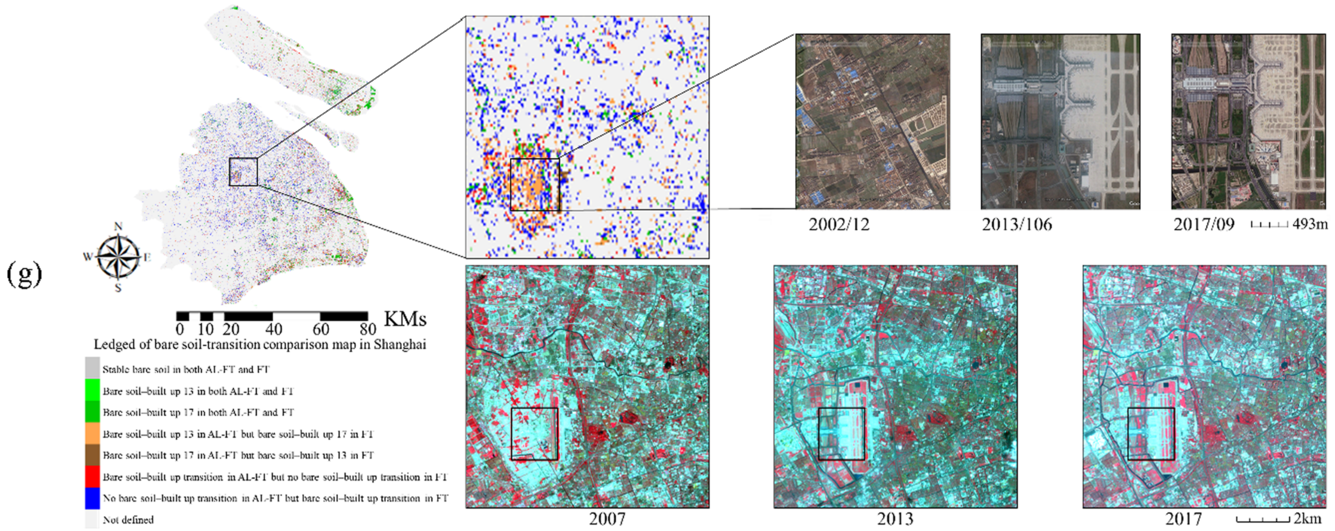
(g)
N
W
E
S
0 10 20 40 60 80 KMs
Ledged of bare soil-transition comparison map in Shanghai
Stable bare soil in both AL-FT and FT
Bare soil–built up 13 in both AL-FT and FT
Bare soil–built up 17 in both AL-FT and FT
Bare soil–built up 13 in AL-FT but bare soil–built up 17 in FT
Bare soil–built up 17 in AL-FT but bare soil–built up 13 in FT
Bare soil–built up transition in AL-FT but no bare soil–built up transition in FT
No bare soil–built up transition in AL-FT but bare soil–built up transition in FT
Not defined
2002/12
2013/106
2017/09
493m
2007
2013
2017
2km

Fig. 5. (*continued*)

Table 8
Accuracy assessment of the multi-temporal change classification results on two training sets.

City		Joint labeling training set				Non-joint labeling training set			
		PCC	FT	BPCC	AL-FT	PCC	FT	BPCC	AL-FT
Shiyan	OA	84.89% ± 1.75%	87.94% ± 1.66%	89.78% ± 1.55%	90.94% ± 1.38%	83.72% ± 0.64%	– –	91.91% ± 0.74%	92.19% ± 0.69%
	MF1	0.5958 ± 0.0691	0.5808 ± 0.0269	0.6353 ± 0.0721	0.6482 ± 0.0606	0.6230 ± 0.0249	– –	0.7120 ± 0.0250	0.6929 ± 0.0242
Shenzhen	OA	84.37% ± 1.06%	87.52% ± 2.08%	89.85% ± 1.81%	91.03% ± 1.95%	84.27% ± 0.45%	– –	91.10% ± 0.58%	91.48% ± 1.04%
	MF1	0.7268 ± 0.0500	0.7236 ± 0.0314	0.7288 ± 0.0467	0.7875 ± 0.0607	0.7345 ± 0.0383	– –	0.7498 ± 0.0360	0.7846 ± 0.0531
Shanghai	OA	80.63% ± 1.88%	87.77% ± 2.43%	92.87% ± 1.27%	93.94% ± 0.63%	78.70% ± 0.47%	– –	93.15% ± 0.34%	94.12% ± 0.65%
	MF1	0.7209 ± 0.0617	0.7486 ± 0.0505	0.8298 ± 0.0572	0.8421 ± 0.0529	0.7168 ± 0.0564	– –	0.8321 ± 0.0536	0.8439 ± 0.0505

In Table 8, all cases are shown, to demonstrate the overall classification performances. When using the same training set, except for the MF1 record of BPCC using the non-joint labeling set, most of the metrics show that the proposed AL-FT method is superior to the post-classification based approaches. When using different training sets, except for the OA values of PCC, most of the metrics show that the non-joint labeling set presents a better performance, and the AL-FT method is still superior to the three compared methods. Although the training samples in the non-joint labeling set may have no land-cover record in some dates, the location flexibility of this set enables more sample sites. Thus, filling these land-cover labels by the proposed automatic sample collection technique can make better use of these training sites. In short, the proposed AL-FT method does not require costly joint labeling, it can make better use of the training sample set without joint labeling, and it is superior to the most recent post-classification based approaches.

5.2. The function of the designed reliability-based sample allocation scheme

In view of the size of U (see the U columns in Table 4), it is impractical to combine such a huge pseudo-set with the original training sample set. The proposed active learning approach aims to iteratively pick the most reliable and informative subset (i.e., $\tilde{U}$) to assist the original training sample set, and focuses more on the difficult categories (e.g., the changed classes). Thus, in this subsection, Fig. 6 displays the iterative processes of the proposed method, from the perspective of both the class-wise assessment and the general performance. Without loss of generality, the number of iterations was set to 30 for each study area. The results of 10 independent trials were statistically recorded to reduce the possible bias induced by the random sampling. In each bubble chart of Fig. 6, the horizontal axis is the number of iterations, the vertical axis shows the class-wise sample set reliability (i.e., $Rs_{k,c}$, which is defined as the number of reliable samples in U of class c divided by the total number of samples in U of class c for the kth iteration), and the bubble size denotes the average sample set ($\tilde{U}_k$) size of the associated category at the beginning of the current iteration. In the lower-right subfigure of Fig. 6, for all the samples in the pseudo-sample set U, the blue curves show the average sample set reliability (Rs_k) and its standard deviation, and the sepia curves show the average reliability of the current multi-classifier system (i.e., Rc_k for the kth iteration, which is equal to the summation of $r(s)$, $s \in U$) and its standard deviation.

It can be seen that the general reliability increases rapidly in the early iterations, and then reaches a maximum and becomes stable after several iterations, which is similar to most of the class-wise reliabilities and sample set size growth trajectories (except for *stable bare soil*). By paying more attention to the heterogeneous classes in the sample allocation, the sample set sizes of the changed classes and the *stable built-up* class are much larger than those of the homogenous classes, and the reliabilities of these heterogeneous classes increase rapidly in the first few iterations. The reliability of *stable bare soil* is slightly reduced, which can be attributed to the following aspects. On the one hand, the scarcity of reliable *stable bare soil* samples (see Table 4) cannot support an enhancement of the *stable bare soil* recognition ability. On the other hand, the rapid growth of the heterogeneous class samples increases the

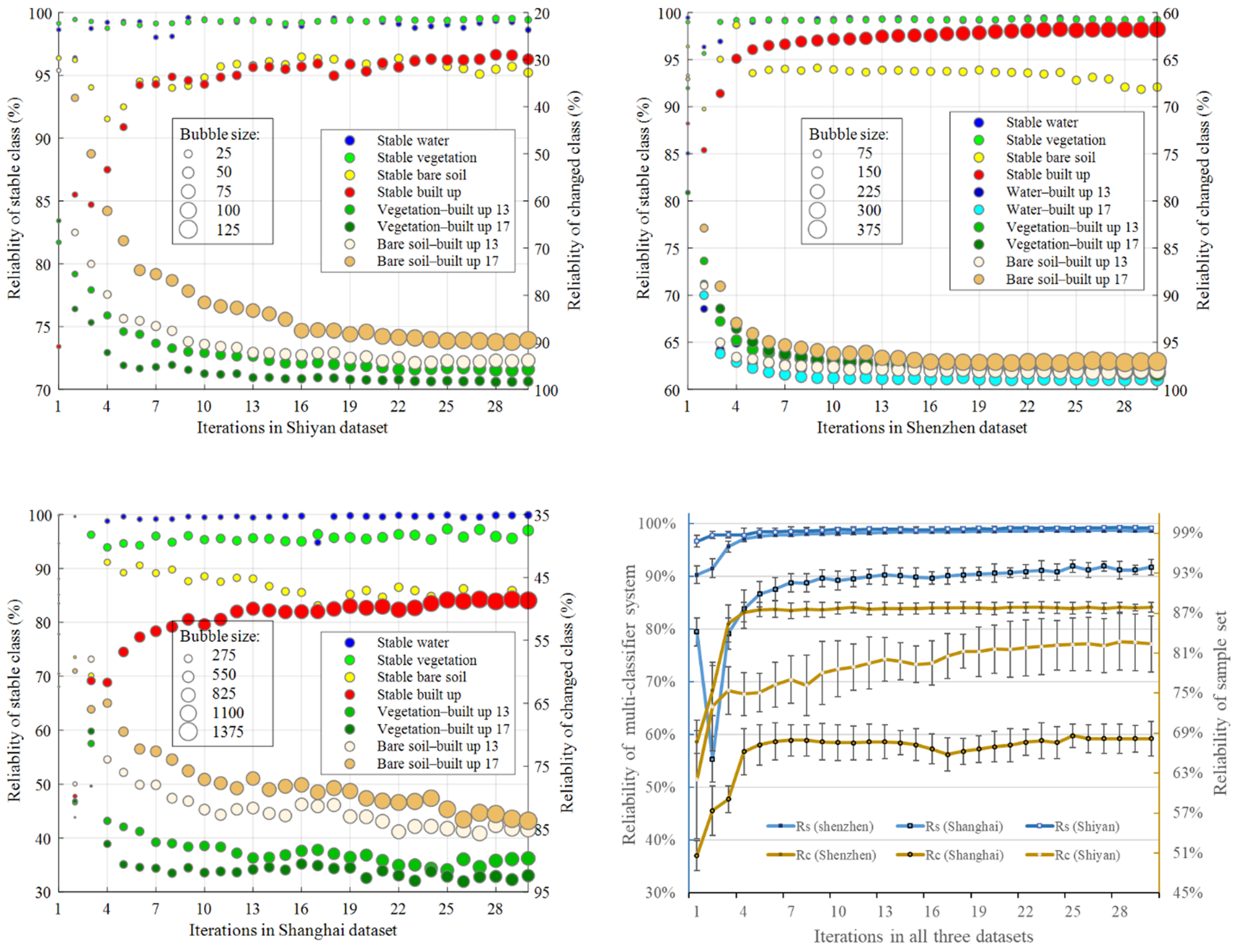

Fig. 6. Iterative processes of the proposed active learning method, including a class-wise assessment for each dataset and a general assessment on all three datasets.

risk of mislabeling *stable bare soil* as other categories. Meanwhile, with regard to the multi-temporal change detection classification results (Table 6), the interpretation ability for *stable bare soil* in the megacities is significantly improved.

Furthermore, we analyzed the function of the adaptive sample allocation process, by taking the following control groups as a comparison. For each control group, a number of pseudo-samples were randomly chosen and combined with the original training sample set to train the multi-classifier system. For the purpose of a fair comparison, the numbers of selected samples for each class were kept equal, and 10 to 1000 samples per class were randomly chosen to train the classifier. Please note that the numbers of adaptively selected pseudo-samples per class by the proposed sample allocation strategy (i.e., AL-FT in Fig. 7) are listed in Table 5. With an eye on the scarcity of the *stable bare soil* class in the Shiyan dataset and the Shenzhen dataset (Table 5), the Shanghai Landsat dataset was chosen in this experiment. The horizontal axis in Fig. 7 is the number of randomly selected samples per class, and the vertical axes show the OA value and the MF1 value of the mapping result. The accuracy was averaged over 10 runs to reduce the possible bias induced by the random sampling. In terms of the accuracy, it can be seen that the accuracy increases with the increment of the pseudo-sample set size, i.e., a rapid growth in the early stage and a slight growth after more than 200 samples per class. It is also noted that the best accuracies of the control groups are inferior to those of the proposed method, which uses much fewer pseudo-samples (see the last column in Table 5) in a class-adaptive allocation manner. Therefore, considering the computational cost (i.e., the sample set size) and mapping performance, the class-wise adaptive sample allocation technique in the proposed active learning method can be deemed as effective.

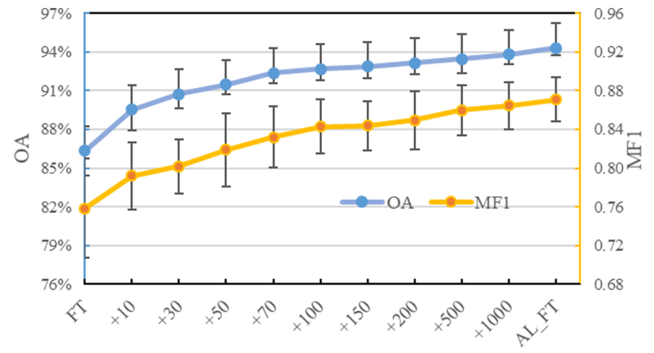

Fig. 7. The accuracy versus the number of randomly added samples per class from the pseudo-sample set U for Shanghai. "$+m$" in the horizontal axis indicates the multi-date classification trained by the union of the original training sample set and a subset from U, which contains m pseudo-samples per class.

5.3. Parameter sensitivity

Using all three Landsat datasets, sensitivity analyses of the land-cover class probability threshold *thr* in the Bayesian based sample

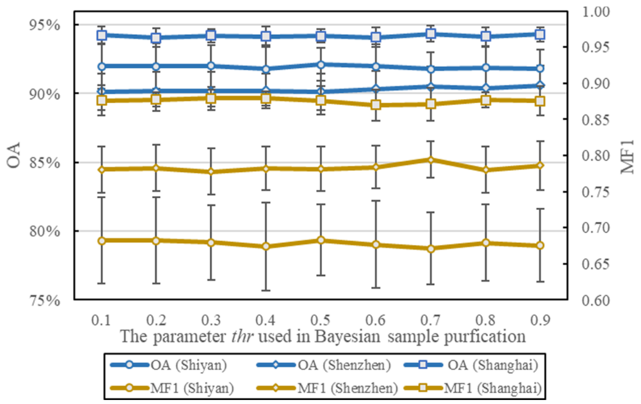

Fig. 8. The classification accuracy versus the threshold *thr* for the Shiyan, Shenzhen, and Shanghai Landsat images.

collection and the batch scale *N* parameter in the active learning (see Eq. (5)) were carried out. By balancing the tradeoff between the land-cover class probability and the adjacent change probability, *thr* aims to purify the pseudo-sample set. Fig. 8 represents the effect of threshold *thr* on the multi-date classification accuracy. The horizontal axis in Fig. 8 denotes the value of *thr* (i.e., from 0.1 to 0.9, with an interval of 0.1), and the vertical axes show the OA and MF1 of the mapping result. The accuracy was again averaged over 10 runs. From these figures, it can be clearly seen that the performances for all the datasets are relatively stable as the threshold increases.

For the sensitivity analysis of the batch scale *N*, Fig. 9 shows the statistical performances for 10 independent trials. In these three bubble charts, the horizontal axis is the number of iterations and the vertical axis shows the pseudo-sample set reliability, which is equal to the number of reliable samples in *U* divided by the total number of samples in *U*. The bubble size denotes the total selected pseudo-sample number at the beginning of the current iteration. The bubble color denotes the value of the batch scale *N*, which is followed by the final multi-temporal change detection result. From both charts, in view of the accuracy, although there are differences in the iterative process of the active learning steps, the batch scale *N* has little impact on the classification results. In view of the computational cost, despite the fact that cases with a large *N* can stabilize in earlier iterations, the cases with large numbers of selected samples will be more burdensome for mapping the multi-temporal change detection result. According to Fig. 9, setting *N* as 5 should be a suitable option to strike a balance. Thus, we set *N* as 5 and *thr* as 0.7 for all the experiments described in the previous subsections, considering both the computational cost and the performance.

6. Conclusion

In this paper, we have proposed a multi-classifier active learning sample collection method for urban land-cover multi-temporal from-to

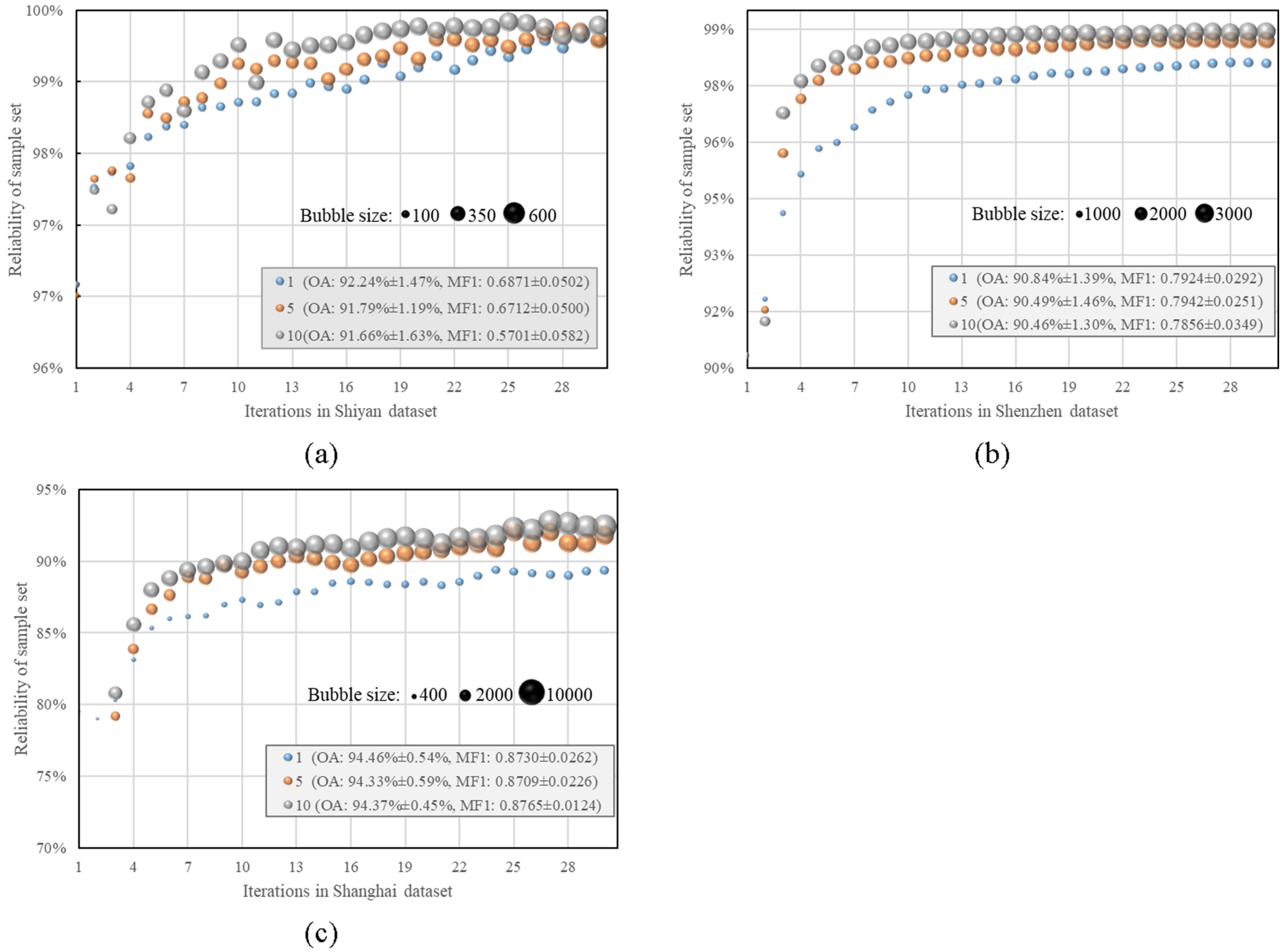

Fig. 9. The iterative process versus the batch scale *N* for (a) the Shiyan, (b) the Shenzhen, and (c) the Shanghai Landsat images.

change detection. The proposed method was effectively tested in one small city with a low urbanization level and two rapidly developing megacities of China, in which the urban environment features challenging spectral-spatial-temporal heterogeneity. The proposed method successfully labeled and selected informative and reliable samples to improve the change detection performance. The collected samples, which were assessed by high spatial resolution reference maps with good evaluation records, were taken into account to validate the effectiveness of the proposed method. When compared with post-classification and the classical multi-date classification methods, the proposed method showed a significant advantage in change detection performance.

Training samples are crucial for multi-temporal change detection, especially for heterogeneous urban areas. Despite the fact that multi-date classification based techniques have the merit of being task-oriented, due to the unclear observations of Landsat imagery (i.e., clouds, cloud shadows, snow/ice, and SLC-off data) and the rare existence of some important classes (e.g., *stable bare soil*), manual multi-temporal joint sample collection is costly. Although the post-classification based approaches are not subject to the multi-temporal joint labeling of training samples, the diversity of the built-up class, the scarcity of the bare soil class in each independent land-cover mapping, and the error accumulation (including illogical land-cover change events) from multi-temporal land covers still restrict their performance and further application. The proposed AL-FT method has advantages over these two above-mentioned techniques, while also overcoming their disadvantages. AL-FT purifies the independent land-cover mappings, selects reliable and informative samples, and conducts task-oriented multi-date classification with sufficient augmented samples. The presented method was tested in two difficult situations (including a small training sample set case and a training sample set without joint labeling), so that the robustness and accuracy of the proposed approach can therefore be expected to be of a similar or better quality in the case of more training samples. Moreover, as a positive attempt in using active learning in a label-noise environment, the proposed method broadens the field of active learning in remote sensing, and will be beneficial to multi-temporal change detection and other applications.

Declaration of Competing Interest

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.

Acknowledgement

This research was supported by the National Natural Science Foundation of China under grants 41701382, 41771360, and 41971295; the National Program for Support of Top-notch Young Professionals; the Hubei Provincial Natural Science Foundation of China under grants 2017CFA029 and 2017CFB188; the National Key R&D Program of China under grant 2016YFB0501403; and the Shenzhen Science and Technology Program under grant JCYJ20180306170645080.

References

Ben-Dor, E., Kindel, B., Goetz, A.F.H., 2004. Quality assessment of several methods to recover surface reflectance using synthetic imaging spectroscopy data. Remote Sens. Environ. 90, 389–404.

Box, J.F., 1987. Guinness, gosset, fisher, and small samples. Statist. Sci. 45–52.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) 2, 1–27.

Chen, D., Stow, D., 2002. The effect of training strategies on supervised classification at different spatial resolutions. Photogramm. Eng. Remote Sens. 68, 1155–1162.

Chon, J., Kim, H., Lin, C.-S., 2010. Seam-line determination for image mosaicking: a technique minimizing the maximum local mismatch and the global cost. ISPRS J. Photogramm. Remote Sens. 65, 86–92.

Demir, B., Persello, C., Bruzzone, L., 2010. Batch-mode active-learning methods for the interactive classification of remote sensing images. IEEE Trans. Geosci. Remote Sens. 49, 1014–1031.

Flaash, U.s.G., 2009. Atmospheric Correction Module: QUAC and Flaash User Guide v. 4. 7. ITT Visual Information Solutions Inc.: Boulder, CO, USA.

Gómez, C., White, J.C., Wulder, M.A., 2016. Optical remotely sensed time series data for land cover classification: a review. ISPRS J. Photogramm. Remote Sens. 116, 55–72.

Guo, Y., Ma, L., Zhu, F., Liu, F., 2015. Selecting training samples from large-scale remote-sensing samples using an active learning algorithm, International Symposium on Computational Intelligence and Intelligent Systems. Springer, pp. 40–51.

Han, X., Huang, X., Li, J., Li, Y., Yang, M.Y., Gong, J., 2018. The edge-preservation multi-classifier relearning framework for the classification of high-resolution remotely sensed imagery. ISPRS J. Photogramm. Remote Sens. 138, 57–73.

Healey, S.P., Cohen, W.B., Yang, Z., Brewer, C.K., Brooks, E.B., Gorelick, N., Hernandez, A.J., Huang, C., Hughes, M.J., Kennedy, R.E., 2018. Mapping forest change using stacked generalization: an ensemble approach. Remote Sens. Environ. 204, 717–728.

Hu, T., Huang, X., Li, J., Zhang, L., 2018. A novel co-training approach for urban land cover mapping with unclear Landsat time series imagery. Remote Sens. Environ. 217, 144–157.

Huang, C., Song, K., Kim, S., Townshend, J.R., Davis, P., Masek, J.G., Goward, S.N., 2008. Use of a dark object concept and support vector machines to automate forest cover change analysis. Remote Sens. Environ. 112, 970–985.

Huang, X., Weng, C., Lu, Q., Feng, T., Zhang, L., 2015. Automatic labelling and selection of training samples for high-resolution remote sensing image classification over urban areas. Remote Sens. 7, 16024–16044.

Hussain, M., Chen, D., Cheng, A., Wei, H., Stanley, D., 2013. Change detection from remotely sensed images: from pixel-based to object-based approaches. ISPRS J. Photogramm. Remote Sens. 80, 91–106.

Im, J., Jensen, J.R., 2005. A change detection model based on neighborhood correlation image analysis and decision tree classification. Remote Sens. Environ. 99, 326–340.

Jabari, S., Rezaee, M., Fathollahi, F., Zhang, Y., 2019. Multispectral change detection using multivariate Kullback-Leibler distance. ISPRS J. Photogramm. Remote Sens. 147, 163–177.

Kasetkasem, T., Varshney, P.K., 2002. An image change detection algorithm based on Markov random field models. IEEE Trans. Geosci. Remote Sens. 40, 1815–1823.

Li, J., Huang, X., Hu, T., Jia, X., Benediktsson, J.A., 2019. A novel unsupervised sample collection method for urban land-cover mapping using landsat imagery. IEEE Trans. Geosci. Remote Sens. 57, 3933–3951.

Li, J., Zhang, H., Zhang, L., 2015a. A nonlinear multiple feature learning classifier for hyperspectral images with limited training samples. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 2728–2738.

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

Liu, X., Lathrop Jr, R., 2002. Urban change detection based on an artificial neural network. Int. J. Remote Sens. 23, 2513–2518.

Lu, D., Mausel, P., Brondizio, E., Moran, E., 2004. Change detection techniques. Int. J. Remote Sens. 25, 2365–2401.

Lu, D., Moran, E., Hetrick, S., 2011. Detection of impervious surface change with multitemporal Landsat images in an urban–rural frontier. ISPRS J. Photogramm. Remote Sens. 66, 298–306.

Masek, J.G., Huang, C., Wolfe, R., Cohen, W., Hall, F., Kutler, J., Nelson, P., 2008. North American forest disturbance mapped from a decadal Landsat record. Remote Sens. Environ. 112, 2914–2926.

Nemmour, H., Chibani, Y., 2006. Multiple support vector machines for land cover change detection: an application for mapping urban extensions. ISPRS J. Photogramm. Remote Sens. 61, 125–133.

Nielsen, A.A., 2007. The regularized iteratively reweighted MAD method for change detection in multi-and hyperspectral data. IEEE Trans. Image Process. 16, 463–478.

Okujeni, A., van der Linden, S., Tits, L., Somers, B., Hostert, P., 2013. Support vector regression and synthetically mixed training data for quantifying urban land cover. Remote Sens. Environ. 137, 184–197.

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57.

Roy, D.P., Huang, H., Boschetti, L., Giglio, L., Yan, L., Zhang, H.H., Li, Z., 2019. Landsat-8 and Sentinel-2 burned area mapping-A combined sensor multi-temporal change detection approach. Remote Sens. Environ. 231, 111254.

Schneider, A., 2012. Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach. Remote Sens. Environ. 124, 689–704.

Schneider, A., Mertes, C., 2014. Expansion and growth in Chinese cities, 1978–2010. Environ. Res. Lett. 9, 024008.

Sun, W., Chen, B., Messinger, D., 2014. Nearest-neighbor diffusion-based pan-sharpening algorithm for spectral images. Opt. Eng. 53, 013107.

Tuia, D., Volpi, M., Copa, L., Kanevski, M., Munoz-Mari, J., 2011. A survey of active learning algorithms for supervised remote sensing image classification. IEEE J. Sel. Top. Signal Process. 5, 606–617.

Wu, C., Du, B., Cui, X., Zhang, L., 2017. A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion. Remote Sens.

Environ. 199, 241–255.

Xian, G., Homer, C., Fry, J., 2009. Updating the 2001 National Land Cover Database land cover classification to 2006 by using Landsat imagery change detection methods. Remote Sens. Environ. 113, 1133–1147.

Xu, Y., Yu, L., Zhao, F.R., Cai, X., Zhao, J., Lu, H., Gong, P., 2018. Tracking annual cropland changes from 1984 to 2016 using time-series Landsat images with a change-detection and post-classification approach: Experiments from three sites in Africa. Remote Sens. Environ. 218, 13–31.

Xue, X., Liu, H., Mu, X., Liu, J., 2014a. Trajectory-based detection of urban expansion using Landsat time series. Int. J. Remote Sens. 35, 1450–1465.

Xue, Z., Du, P., Feng, L., 2014b. Phenology-driven land cover classification and trend analysis based on long-term remote sensing image series. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7, 1142–1156.

Yu, W., Zhou, W., Qian, Y., Yan, J., 2016. A new approach for land cover classification and change analysis: integrating backdating and an object-based method. Remote Sens. Environ. 177, 37–47.

Zhong, L., Hu, L., Zhou, H., 2019. Deep learning based multi-temporal crop classification. Remote Sens. Environ. 221, 430–443.