Generating a series of fine spatial and temporal resolution land cover maps by fusing coarse spatial resolution remotely sensed images and fine spatial resolution land cover maps

Xiaodong Li\textsuperscript{a}, Feng Ling\textsuperscript{a,b,*}, Giles M. Foody\textsuperscript{b}, Yong Ge\textsuperscript{c}, Yihang Zhang\textsuperscript{a}, Yun Du\textsuperscript{a}

\textsuperscript{a} Key Laboratory of Monitoring and Estimate for Environment and Disaster of Hubei Province, Institute of Geodesy and Geophysics, Chinese Academy of Sciences, Wuhan 430077, China

\textsuperscript{b} School of Geography, University of Nottingham, University Park, Nottingham NG7 2RD, U.K.

\textsuperscript{c} State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences & Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

*Corresponding author at Key Laboratory of Monitoring and Estimate for Environment and Disaster of Hubei Province, Institute of Geodesy and Geophysics, Chinese Academy of Sciences, Wuhan 430077, China

E-mail address: lingf@whigg.ac.cn (F. Ling)
ABSTRACT

Studies of land cover dynamics would benefit greatly from the generation of land cover maps at both fine spatial and temporal resolutions. Fine spatial resolution images are usually acquired relatively infrequently, whereas coarse spatial resolution images may be acquired with a high repetition rate but may not capture the spatial detail of the land cover mosaic of the region of interest. Traditional image spatial–temporal fusion methods focus on the blending of pixel spectra reflectance values and do not directly provide land cover maps or information on land cover dynamics. In this research, a novel Spatial–Temporal remotely sensed Images and land cover Maps Fusion Model (STIMFM) is proposed to produce land cover maps at both fine spatial and temporal resolutions using a series of coarse spatial resolution images together with a few fine spatial resolution land cover maps that pre- and post-date the series of coarse spatial resolution images. STIMFM integrates both the spatial and temporal dependences of fine spatial resolution pixels and outputs a series of fine spatial–temporal resolution land cover maps instead of reflectance images, which can be used directly for studies of land cover dynamics. Here, three experiments based on simulated and real remotely sensed images were undertaken to evaluate the STIMFM for studies of land cover change. These experiments included comparative assessment of methods based on single-date image such as the super-resolution approaches (e.g., pixel swapping-based super-resolution mapping) and the state-of-the-art spatial–temporal fusion approach that used the Enhanced Spatial and Temporal Adaptive Reflectance Fusion Model (ESTARFM) and the Flexible Spatiotemporal DAta Fusion model (FSDAF) to predict...
the fine-resolution images, in which the maximum likelihood classifier and the automated land cover updating approach based on integrated change detection and classification method were then applied to generate the fine-resolution land cover maps. Results show that the methods based on single-date image failed to predict the pixels of changed and unchanged land cover with high accuracy. The land cover maps that were obtained by classification of the reflectance images outputted from ESTARFM and FSDAF contained substantial misclassification, and the classification accuracy was lower for pixels of changed land cover than for pixels of unchanged land cover. In addition, STIMFM predicted fine spatial–temporal resolution land cover maps from a series of Landsat images and a few Google Earth images, to which ESTARFM and FSDAF that require correlation in reflectance bands in coarse and fine images cannot be applied. Notably, STIMFM generated higher accuracy for pixels of both changed and unchanged land cover in comparison with other methods. Keywords: Spatial temporal fusion; Super-resolution mapping; Endmember extraction.
1. Introduction

Land cover maps are one of the most fundamental datasets used in many scientific fields and are often produced from remotely sensed images (Bartholome and Belward 2005; Friedl et al. 2002). A wide variety of remote sensing systems have been developed, and hence, images are available with different spatial and temporal resolutions, thereby allowing the production of land cover maps at different spatial and temporal scales. With most satellite remote sensing systems, a trade-off typically exists between spatial and temporal resolution. In general, fine spatial resolution remote sensors can acquire images that provide spatially detailed land cover information, but their relatively coarse temporal resolution limits their usage in monitoring rapid land cover changes. By contrast, coarse spatial resolution remotely sensed images can often be acquired at a fine temporal resolution that provides a repetition rate suitable for the detection of rapid land cover changes but are unable to represent the spatial detail of the land cover mosaic. To realize the full potential of remote sensing as a source of information on land cover change, a method that allows the production of land cover maps with both fine spatial and temporal resolutions is required. Such maps could be obtained by combining all available remotely sensed images of varying spatial and temporal resolution to form a series of fine-resolution land cover maps.

Recently, spatial–temporal image fusion, which aims to produce fine spatial and temporal resolution remotely sensed images from images with different spatial and temporal resolutions, has become a promising means to address the trade-off between spatial and temporal resolution (Gevaert and Garcia-Haro 2015; Zhu et al. 2016).
Spatial–temporal data fusion methods can be categorized into weighted function based methods, unmixing-based methods, and dictionary-pair learning based methods (Zhu et al. 2016). Among the weighted function based methods, the Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM) proposed by Gao et al. (2006) was developed first and is one of the most popular spatial–temporal image fusion methods. By fusing coarse spatial resolution Moderate Resolution Imaging Spectroradiometer (MODIS) and fine spatial resolution Landsat sensor images, STARFM can predict Landsat-like reflectance images with the spatial resolution of Landsat and the temporal resolution of MODIS. A number of studies have suggested improvements to STARFM, including studies of forest disturbance (Hilker et al. 2009a), in heterogeneous regions (Zhu et al. 2010), as well as in gap filling to reduce the negative effects of cloud (Gevaert and Garcia-Haro 2015). STARFM and the improved models based on it have been mainly used to detect reflectance changes caused by processes such as phenology over large areas, and used to generate dense time series of Landsat-like data (Hilker et al. 2009b), enhance land cover classification (Jia et al. 2014), and predict key environmental variations such as evapotranspiration (Anderson et al. 2011) and temperature (Hilker et al. 2009b). Other spatial–temporal image fusion models, such as the unmixing-based algorithm that extracts endmembers on the basis of linear spectral mixture model and assigns the unmixed reflectance to fine spatial resolution pixels (Huang and Zhang 2014; Zhukov et al. 1999; Zurita-Milla et al. 2009) and the dictionary-pair learning based methods, which capture features from the coarse- and fine-resolution image pairs used for predicting fine-resolution image (Huang and Song
2012), have also been proposed and applied to Landsat and MODIS images in recent years (Amoros-Lopez et al. 2013; Wu et al. 2012).

Generally, spatial–temporal image fusion models aim to generate a series of continuous reflectance values instead of discrete categorical values. A further image classification step is needed to produce from the reflectance images a corresponding series of land cover maps for the study of land cover class dynamics (Jia et al. 2014). The use of these methods for generating land cover maps and monitoring land cover changes often suffers from two important limitations.

First, most spatial–temporal image fusion algorithms assume that land cover type does not change during the data observation period (Fu et al. 2013; Gao et al. 2006; Zhu et al. 2010). Previous research has shown that STARFM does not deal well with abrupt land cover changes. Song and Huang (2013) showed that STARFM failed to fuse the pixel reflectance accurately in a study of land cover change in an urban area. The Enhanced STARFM (ESTARFM) is often better than STARFM for studies of heterogeneous landscapes (Zhu et al. 2010) but can be worse than STARFM for predicting abrupt changes of land cover type (Emelyanova et al. 2013). The Spatial Temporal Adaptive Algorithm for mapping Reflectance CHange (STAARCH) improves STARFM’s performance when land cover type change and disturbance exist, but it is more suitable for spatial–temporal fusion of forest land cover (Hilker et al. 2009a). The Flexible Spatiotemporal DAta Fusion model (FSDAF) can predict Landsat-like reflectance values with both gradual change and land cover type change, but it cannot capture tiny changes in land cover type, such as when only a few fine
pixels experienced land cover type change and the change is invisible in the coarse-resolution image (Zhu et al. 2016). Similar to STARFM, the unmixing-based spatial–temporal reflectance fusion methods consider only the change in endmember spectra but not in land cover types (Huang and Zhang 2014; Zhukov et al. 1999; Zurita-Milla et al. 2009).

Second, most spatial–temporal image fusion methods need one or more observed pairs of coarse- and fine-resolution images for training and require the coarse- and fine-resolution remotely sensed data from different satellite sensors to be mutually comparable and correlated. All the weighted function based methods, including STARFM, ESTARFM, STAARCH, and all the dictionary-pair learning-based methods need one or more observed pairs of coarse- and fine-resolution images, which have comparable reflectance bands, for training (Gao et al. 2006; Gevaert and Garcia-Haro 2015; Zhu et al. 2010). These methods mainly focus on predicting Landsat-like remotely sensed images with MODIS repetition rates. However, these methods cannot deal with other satellite images, which have uncorrelated reflectance bands, and are thus limited in the use of land cover change analysis. For instance, in regional-scale land cover analysis, the detection of very-high-resolution land cover changes at high temporal resolutions is required. In general, we can obtain a series of Landsat images and a few very-high-resolution images such as panchromatic aerial photograph. The weighted function based and dictionary-pair learning based methods cannot fuse these data because the very-high-resolution images usually have different reflectance bands compared with Landsat images.
The spatial–temporal image fusion methods aim to produce fine spatial–temporal resolution reflectance images rather than land cover maps. The fused fine-resolution images have many applications, such as phenology analysis. If the aim is to generate a sequence of land cover maps from the reflectance images from which land cover change trajectories may be extracted, then a further image classification analysis is still required, which may introduce uncertainty and error in the land cover maps. First, the classification of an image series can be complex and laborious. Training statistics are required to inform classification analysis, and these may vary in quality in time due to issues such as phenology. Moreover, the classification is also problematic, with the potential for different classifiers to generate dissimilar land cover maps from the same training data. Traditional classifiers applied to mono-temporal image may also ignore the temporal information contained in a series of images and thereby produce a classification of sub-optimal accuracy. The spatial–temporal–based image classifier has the advantage in taking both the spatial and temporal links between neighboring pixels (Cai et al. 2014), but is challenging to use for voluminous image series (Liu and Cai 2012; Liu et al. 2006). Finally, the spatial–temporal image fusion models generate a large volume of fine spatial–temporal resolution reflectance images as the intermediate data to be used for the production of land cover maps. This situation may represent practical challenges in terms of data access and storage.

Given the concerns with the spatial–temporal reflectance fusion model for producing land cover maps, a more appropriate fusion approach could be based on directly downscaling the coarse spatial resolution image series to fine spatial resolution.
land cover maps rather than reflectance images, with the aid of information derived from a few fine spatial resolution images that may be available. Chen et al. (2015) updated land cover maps from downscaled Normalized Difference Vegetation Index (NDVI) time-series data from MODIS, a current Landsat image, and a Landsat image that pre-dates it. The NDVI time-series data are used as ancillary data to extract changed pixels in the Landsat images, and the labels of changed pixels are determined using the current Landsat image. Thus, this method can update fine-resolution land cover maps with Landsat repetition rates based on available Landsat images, but cannot predict fine-resolution land cover maps with MODIS repetition rates. In addition, a major problem with this approach is that a large proportion of coarse spatial resolution image pixels may be of mixed land cover composition. A possible solution of this problem is to use the fractional land cover class composition images that can be extracted via a spectral unmixing analysis. A comparison of the obtained fraction images indicates the change, if any, in land cover that has occurred in the time period between the dates of image acquisition (Lu et al. 2004). This approach can potentially reveal important temporal land cover information, such as land cover modification and conversion (Foody 1999; Lu et al. 2011). Unfortunately, these approaches show only the change in the fraction of the area that is represented by each coarse-resolution pixel and not the geographical location of the change. Information on the location of change might be obtained through a super-resolution analysis (Li et al. 2016; Wang et al. 2015).

Super-resolution land cover mapping (SRM) is a promising technique used to generate land cover maps with a finer spatial resolution than the input data and is
typically viewed as a post-processing approach applied after a spectral unmixing analysis. SRM predicts the spatial distribution of each land cover class in the area represented by each coarse spatial resolution pixel and provides more fine spatial resolution land cover information than spectral unmixing (Atkinson 2009). Various approaches have been proposed (Atkinson 2005; Ge et al. 2016; Kasetkasem et al. 2005; Ling et al. 2016; Tatem et al. 2003), and SRM has been used in many fields, including the extraction of waterlines (Foody et al. 2005), rural land cover (Tatem et al. 2003), refining the estimation of ground control point location (Foody 2002), land cover change detection (Wang et al. 2015), land cover map updating (Li et al. 2015b), and wetland inundation analysis (Li et al. 2015a).

Traditionally, SRM is applied to a mono-temporal coarse spatial resolution image dataset. The SRM solution space is large because SRM predicts land cover maps with finer spatial resolution than the input data, and it can provide multiple plausible solutions that satisfy the constraints of the SRM analysis. A fine spatial resolution land cover map that pre- or post-dates the coarse spatial resolution image could be used to provide fine spatial resolution information to constrain and enhance the SRM solution (Li et al. 2015b; Ling et al. 2011; Wang et al. 2015; Xu and Huang 2014). Although the accuracy of SRM may be increased through the use of multi-temporal data, challenges remain, especially if a time series of images are used. Often, a sequence of coarse spatial resolution images together with a few fine spatial resolution images that pre- and post-date the coarse-resolution images are available. Applying existing SRMs to each coarse-resolution image without or with only one fine spatial resolution map
that pre- or post-dates it fails to account for the temporal dependence in the image series and fails to fully exploit the available information. The construction of SRM that considers the temporal dependences of a coarse-resolution image from fine-resolution maps that pre- and post-date it is necessary for a fuller reconstruction of land cover dynamics.

The objective of this paper is to propose a Spatial–Temporal remotely sensed Images and land cover Maps Fusion Model (STIMFM). The inputs to STIMFM are a series of coarse spatial resolution multi-spectral remotely sensed images and few fine spatial resolution land cover maps that pre- and post-date the coarse spatial resolution image series. The fine spatial resolution land cover maps can be obtained from various data sources, such as through classification of remotely sensed images or maps produced conventionally from field survey. As a result, the input to STIMFM is more general than that of other spatial–temporal image fusion models. Critically, STIMFM outputs a series of fine spatial–temporal resolution land cover maps, not reflectance images. In addition, STIMFM takes information on class temporal dependence that exists in different images into account and is able to deal with land cover change. STIMFM was compared with a set of alternative methods. The latter includes two SRM methods that use a mono-temporal coarse spatial resolution image as input and two spatial–temporal image fusion methods, namely, the ESTARFM which adopts the coarse spatial resolution image and two fine and coarse spatial resolution image pairs that pre- and post-date the coarse-resolution image as input, and the FSDAF which adopts the coarse spatial resolution image and one fine and coarse spatial resolution
image pair that pre- or post-date the coarse-resolution image as input.

2. Methods

2.1. STIMFM

A coarse spatial resolution image $Y$ contains $I \times J$ pixels. Fine spatial resolution land cover maps of the same geographical region are $X_{pre}$ and $X_{post}$, which are temporally close to and pre- or post-date $Y$, respectively. $X_{pre}$ and $X_{post}$ contain $I \times s \times J \times s$ pixels, where $s$ is the scale or zoom factor and each coarse spatial resolution pixel contains $s \times s$ fine spatial resolution pixels. $C$ land cover classes are present in $X_{pre}$ and $X_{post}$. The STIMFM predicts a fine spatial resolution land cover map $X$ at the time of coarse-resolution image $Y$ observation, and has $I \times s \times J \times s$ pixels, each of which has a land cover class label in $C$. STIMFM produces a series of fine spatial and fine temporal resolution land cover maps. It uses a series of coarse spatial resolution remotely sensed images and a few fine spatial resolution land cover maps as input (Fig. 1). STIMFM comprises several main steps, including spectral endmember estimation, analysis of land cover class fraction temporal change, objective function construction, and model optimization. The STIMFM flowchart is shown in Fig. 2.
Fig. 1 Production of a series of fine spatial and temporal resolution land cover maps from a series of coarse spatial resolution remotely sensed images and a few fine spatial resolution land cover maps in STIMFM.

Fig. 2 Flowchart of STIMFM.

2.2. Spectral endmember estimation
In STIMFM, endmembers that are representative of the spectra of pure land cover classes are estimated for coarse spatial resolution remotely sensed images. Endmember spectra need to be extracted for each coarse spatial resolution image in the dataset as differences may be expected in a time series due to issues such as phenology or variation in image acquisition properties (e.g., angular viewing geometry). Although many endmember extraction algorithms are available, they are not directly used in STIMFM because spectral endmembers are difficult to extract accurately from coarse spatial resolution remotely sensed images due to the small proportion of pure pixels that are typically contained. Information for the estimation of endmembers is instead provided by the fine spatial resolution land cover maps that pre- and post-date the coarse spatial resolution image time series.

The land cover classes are defined in the fine spatial resolution land cover maps. For each coarse spatial resolution remotely sensed image, the linear mixture model (LMM) is applied in STIMFM to estimate endmember spectra. With the LMM, the spectral response of each coarse spatial resolution pixel is viewed as being composed of a weighted linear sum of the endmember spectra within that pixel, in which the weights are determined by the relative areal proportions of each endmember (Settle and Drake 1993). On the basis of the linear mixing assumption, the spectral signature $y_{ij}$ for the coarse spatial resolution pixel $(i,j)$ in $Y$ can be represented by

$$y_{ij} = Ef_{ij} \quad (1)$$

where $y_{ij}$ is a $B \times 1$ spectral vector, $B$ is the number of spectral bands, $E$ is a $B \times C$ matrix that represents the endmembers used for $Y$, $f_{ij}$ is the $C \times 1$ vector that represents
fractions of all endmembers in the pixel \((i,j)\) in \(Y\).

Theoretically, to solve for \(E\) with \(B \times C\) unknown variables, at least \(B \times C\) equations are required. \(l\) \((l>C)\) coarse pixels are collected to compose a system of linear mixture equations

\[
[y_1, y_2, \ldots, y_l] = E [f_1, f_2, \ldots, f_l]
\]

where \(y_i\) is the spectral signature for the \(l\)-th coarse spatial resolution pixel in \(Y\), and \(f_i\) is the fraction vector of different classes in the \(l\)-th coarse spatial resolution pixel in \(Y\). \(E\) can be solved on the basis of the inversion of Eq. (2) by computing a least squares best fit solution

\[
[E] = \arg \min \left( \sum_{n=1}^{l} \| y_n - E f_n \|^2 \right)
\]

where \(y_n\) is the \(n\)-th coarse spatial resolution pixel's spectral signature in \(Y\), and \(f_n\) is the fraction vector in the \(n\)-th coarse spatial resolution pixel in \(Y\). \(\| \|_2^2\) is the L2 norm of the residual vector. "argmin" means the minimizing argument of the function.

A number of coarse spatial resolution pixels in \(Y\) with known endmember fractions are sought to solve Eq. (3). For each class, the focus is a set of coarse-resolution pixels that have the least changed fractions of that class during the time period covered by \(X_{pre}\) and \(X_{post}\). To avoid the collinearity problem in the use of LMM (van der Meer and Jia 2012), \(m\) coarse-resolution pixels that have the highest fraction of a given class (i.e., the \(m\) purest coarse-resolution pixels of the class) among the selected set of coarse-resolution pixels are used. All the \(m \times C\) coarse spatial resolution pixels are used for endmember estimation in Eq. (3), which can be solved by computing a least squares best fit solution. Assuming the fractions of the \(m \times C\) coarse spatial resolution pixels
are unchanged, the fractions of these coarse pixels in $X_{\text{pre}}$ or $X_{\text{post}}$ are used as a substitute of those in $Y$. The fractions in $X_{\text{pre}}$ and $X_{\text{post}}$ are produced through a spatial degradation process by dividing the number of fine spatial resolution pixels of each class by the total number of fine spatial resolution pixels in a coarse spatial resolution pixel (i.e., $s^2$).

2.3. Analysis of land cover class fraction temporal change

With the estimated endmembers, class fraction images that represent the area percentage of a pixel occupied by each endmember can be extracted from coarse spatial resolution image $Y$ using the estimated endmember spectra $E$ and on the basis of the mean square error minimization criterion of the LMM

$$[\tilde{f}_y] = \arg \min \left\| y_y - E f_y \right\|^2$$

$$0 \leq f_{ijc} \leq 1, \quad c = 1, \ldots, C$$

$$\sum_{c=1}^{C} f_{ijc} = 1$$

where $f_y = [f_{ij1}, f_{ij2}, \ldots, f_{ijC}]^T$, and $f_{ijc}$ is the fraction value of the $c$-th endmember in coarse spatial resolution pixel $(i,j)$ in $Y$.

The fraction images produced from the coarse spatial resolution image by spectral unmixing, as well as those produced by spatially degrading the fine spatial resolution land cover maps that pre- and post-date the coarse spatial resolution image, provide the land cover trajectory at the acquisition times of $X_{\text{pre}}$, $Y$, and $X_{\text{post}}$. The change of class fractions in each coarse spatial resolution pixel represents the temporal transitions between classes in the period between the dates of image acquisition. If the class fractions remain unchanged between the coarse-resolution image and fine-resolution
map that pre- or post-dates it, then the fine spatial resolution pixel class labels may
probably also be unchanged during this period. In this case, the images are temporally
correlated. By contrast, if the class fractions changed drastically between two images,
then the fine spatial resolution pixels may have changed during this period. Thereby,
the images are temporally uncorrelated. As a result, the temporal dependence between
different images can be analyzed on the basis of the change in fractions in each coarse
spatial resolution pixel.

Assume \(a_{ijk}\) is the \(k\)-th \((k = 1, \ldots, s^2)\) fine spatial resolution pixel in the coarse
spatial resolution pixel \((i, j)\) \((i = 1, \ldots, I, \ j = 1, \ldots, J)\) in the land cover map \(X\), \(a_{ijk,pre}\)
and \(a_{ijk,post}\) are the \(k\)-th fine spatial resolution pixel in coarse spatial resolution pixel \((i, j)\)
in the maps \(X_{pre}\) and \(X_{post}\), and \(c(a_{ijk}), c(a_{ijk,pre}), \) \(and\) \(c(a_{ijk,post})\) \(are\) land cover class labels
for fine spatial resolution pixels \(a_{ijk}\), \(a_{ijk,pre}\), and \(a_{ijk,post}\), respectively. The temporal
dependence or correlation between fine spatial resolution pixels \(a_{ijk,pre}\) and \(a_{ijk}\) during
\(X_{pre}\) and \(Y\) observation period or between fine spatial resolution pixels \(a_{ijk}\) and \(a_{ijk,post}\)
during \(Y\) and \(X_{post}\) observation period, which is dependent on the class labels of \(a_{ijk,pre}\)
and \(a_{ijk}\) or the class labels of \(a_{ijk}\) and \(a_{ijk,post}\) [Eqs. (7)–(8)] and the change in fractions in
this coarse pixel measured by \(w_{ij,pre}\) and \(w_{ij,post}\) [Eqs. (9)–(10)], can be characterized as

\[
\delta(c(a_{ijk}), c(a_{ijk,pre})) \quad \text{or} \quad w_{ij,post} \times \delta(c(a_{ijk}), c(a_{ijk,post})).
\]

\[
\delta(c(a_{ijk}), c(a_{ijk,pre})) = \begin{cases} 1 & \text{if } c(a_{ijk}) = c(a_{ijk,pre}) \\ 0 & \text{otherwise} \end{cases}
\]

\[
\delta(c(a_{ijk}), c(a_{ijk,post})) = \begin{cases} 1 & \text{if } c(a_{ijk}) = c(a_{ijk,post}) \\ 0 & \text{otherwise} \end{cases}
\]

On the basis of the Kronecker delta function, Eqs. (7)–(8) return a value of 1 if the fine
spatial resolution pixel in different images have an unchanged class label, thereby
indicating that different image pixels are temporally correlated, and a value of 0 if the
fine spatial resolution pixel in different images have changed class labels, thereby
indicating that the different image pixels are temporally uncorrelated.

The changes in fractions in coarse-resolution pixel \((i,j)\) during \(X_{\text{pre}}\) and \(Y\)
observation period and during \(Y\) and \(X_{\text{post}}\) observation period are measured by \(w_{ij,\text{pre}}\)
and \(w_{ij,\text{post}}\) on the basis of the Gaussian model in Eqs. (9)–(10)

\[
w_{ij,\text{pre}} = \exp\left(-\left\|f_{ij} - f_{ij,\text{pre}}\right\|^2\right) \tag{9}
\]

\[
w_{ij,\text{post}} = \exp\left(-\left\|f_{ij} - f_{ij,\text{post}}\right\|^2\right) \tag{10}
\]

where \(f_{ij,\text{pre}}\) and \(f_{ij,\text{post}}\) are the land cover fraction vector in coarse pixel \((i,j)\) in \(X_{\text{pre}}\) and
\(X_{\text{post}}\) produced by spatially degrading \(X_{\text{pre}}\) and \(X_{\text{post}}\) according to the scale factor \(s\). \(w_{ij,\text{pre}}\)
and \(w_{ij,\text{post}}\) indicate the strength of temporal dependence between fine pixels in coarse
pixel \((i,j)\) during \(X_{\text{pre}}\) and \(Y\) observation period or during \(Y\) and \(X_{\text{post}}\) observation period.
\(w_{ij,\text{pre}}\) and \(w_{ij,\text{post}}\) decrease with the increase in the change of fractions in Eqs. (9)–(10).

2.4. Spatial–temporal SRM model

Given the coarse spatial resolution image \(Y\), the fine spatial resolution maps \(X_{\text{pre}}\)
and \(X_{\text{post}}\), STIMFM aims to predict the fine spatial resolution land cover map \(X\) at the
time of \(Y\) observation. The optimal STIMFM result \(X\), given \(Y\), \(X_{\text{pre}}\), and \(X_{\text{post}}\), can be
formulated by applying the maximum a posteriori rule in Bayesian framework, i.e., by
solving the maximization problem:
\[
\bar{X} = \arg \max \left\{ P_{\text{posterior}} \left( X | Y, X_{\text{pre}}, X_{\text{post}} \right) \right\} \\
= \arg \max \left\{ \frac{1}{Z} \exp \left[ -U_{\text{posterior}} \left( X | Y, X_{\text{pre}}, X_{\text{post}} \right) \right] \right\} 
\]

where \( Z \) is a normalizing constant. \( P_{\text{posterior}} \left( X | Y, X_{\text{pre}}, X_{\text{post}} \right) \) is the posterior probability of \( X \), given \( Y, X_{\text{pre}}, \) and \( X_{\text{post}} \). \( U_{\text{posterior}} \left( X | Y, X_{\text{pre}}, X_{\text{post}} \right) \) is the posterior energy function of \( X \), given \( Y, X_{\text{pre}}, \) and \( X_{\text{post}} \). The solving of (11) is complicated because it involves the optimization of a global distribution model of the entire image.

Based on the Markov random field approach, the searching of the optimal \( X \) is equivalent to minimization the posterior energy function, which can be specified to model the spatial and temporal dependencies of pixel on its spatial and temporal neighborhoods (Cai et al. 2014; Li et al. 2014).

\[
U_{\text{posterior}} \left( X | Y, X_{\text{pre}}, X_{\text{post}} \right) = U_{\text{spectral}} \left( Y | X \right) + U_{\text{spatial}} \left( X \right) + U_{\text{temporal}} \left( X | X_{\text{pre}}, X_{\text{post}} \right) 
\]

where \( U_{\text{spectral}} \left( Y | X \right) \) is spectral constraint function that represents the inconsistency between \( Y \) and \( X \), \( U_{\text{spatial}} \left( X \right) \) and \( U_{\text{temporal}} \left( X | X_{\text{pre}}, X_{\text{post}} \right) \) are the spatial and temporal constraint functions, respectively.

### 2.4.1 Spectral constraint function

The spectral constraint function is used to link the fine spatial resolution land cover map \( X \) with the observed coarse spatial resolution image \( Y \). The spectral response of a coarse spatial resolution pixel in \( Y \) is composed of a weighted linear sum of endmember spectral responses within that pixel in the fine spatial resolution map \( X \) on the basis of the LMM. A synthetic coarse spatial resolution pixel spectral signature is developed for a coarse spatial resolution pixel on the basis of the endmember spectral signatures and the fraction of each endmember according to the LMM. The STIMFM spectral
constraint function aims to minimize the L2 norm of the residual vector between the observed and synthetic coarse spatial resolution spectral signatures

$$U_{\text{spectral}}(Y|X) = \sum_{i=1}^{I} \sum_{j=1}^{J} \| y_{ij} - \mathbf{E}_{f_{ij}} \|^2$$  \hspace{1cm} (13)

where the class fraction vector $f_{ij}$ is calculated by dividing the number of fine-resolution pixels of different classes in coarse-resolution pixel $(i, j)$ by $s^2$ in $X$, which is estimated and iteratively updated from STIMFM. $\mathbf{E}_{f_{ij}}$ is the synthetic spectra for coarse-resolution pixel $(i, j)$ on the basis of the LMM.

2.4.2 Spatial constraint function

The spatial constraint function is used to describe the spatial pattern of land cover distribution. In STIMFM, the maximal spatial dependence model that aims to maximize the spatial dependence between neighboring fine spatial resolution pixels was used for its simplicity and effectiveness (Atkinson 2009). For a fine spatial resolution pixel $a_{ijk}$, the spatial dependence is quantified with respect to its neighboring fine spatial resolution pixels. The STIMFM spatial constraint function is computed as

$$U_{\text{spatial}}(X) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \left( -\lambda_s \times \sum_{l \in N(a_{ijk})} \frac{1}{d(a_{ijk}, a_l)} \cdot \delta(c(a_{ijk}), c(a_l)) \right)$$  \hspace{1cm} (14)

where $N(a_{ijk})$ is the spatial neighborhood that includes all fine spatial resolution pixels inside a square window whose center is $a_{ijk}$ ($a_{ijk}$ itself is not included), and $a_l$ is a neighboring fine spatial resolution pixel of $a_{ijk}$ in $N(a_{ijk})$. The size of the neighborhood $N(a_{ijk})$ is $W$. $d(a_{ijk}, a_l)$ is the Euclidean distance between $a_{ijk}$ and $a_l$. $c(a_l)$ is the land cover class label for fine spatial resolution pixel $a_l$. $\delta(c(a_{ijk}), c(a_l))$ equals 1 if $c(a_{ijk})$ and $c(a_l)$ are the same and 0 otherwise. $\lambda_s$ is the spatial weight parameter. $-\lambda_s$ is
multiplied because the STIMFM objective function seeks the minimal value as the optimal solution.

2.4.3 Temporal constraint function

The STIMFM temporal constraint function is used to measure the temporal dependence between the predicted fine spatial resolution map $X$ and the input fine spatial resolution maps $X_{pre}$ and $X_{post}$. The class label of fine spatial resolution pixel $a_{ijk}$ is temporally correlated to fine spatial resolution pixel $a_{ijk,pre}$ and $a_{ijk,post}$ in the maps $X_{pre}$ and $X_{post}$ depending on the class labels of $a_{ijk}$, $a_{ijk,pre}$ and $a_{ijk,post}$ and the strength of temporal dependences measured by the weights $w_{ij,pre}$ and $w_{ij,post}$. $\lambda_T$ is the temporal weight parameter. The STIMFM temporal constraint function is written as

$$U_{temporal}(X|X_{pre},X_{post}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (-\lambda_T \times (w_{ij,pre} \times \delta(c(a_{ijk}),c(a_{ijk,pre}))) + w_{ij,post} \times \delta(c(a_{ijk}),c(a_{ijk,post})))\) \quad (15)$$

2.5. Model initialization and optimization

An initial fine spatial resolution land cover map is used as input to STIMFM at the outset. The initialization map is produced according to the land cover class fraction images estimated from a coarse spatial resolution image. The fine spatial resolution pixels are randomly allocated class labels in a manner that maintains the class proportional information conveyed by a prior spectral unmixing analysis (Kasetkasem et al. 2005). The class labels in the initial fine spatial resolution land cover map are then updated iteratively. Here, the Iterative Conditional Mode (ICM) was applied to update the fine spatial resolution pixel class labels. ICM converges when no pixel class
labels change during two successive iterations or when a predefined number of
iterations have been undertaken.

3. Experiments and results

The proposed STIMFM was evaluated in three experiments. The first used Landsat
multi-spectral images and the National Land Cover Database (NLCD) land cover maps
(Landsat–NLCD). The second used MODIS and Landsat multi-spectral images
(MODIS–Landsat). The third used Landsat and Google Earth Images (Landsat–GEI).

For a rigorous assessment, several traditional approaches were used for comparison,
including the Pixel Swapping based SRM (PS_SRM) (Atkinson 2005), the Spatial
Regularization based SRM (SR_SRM) (Ling et al. 2014), the ESTARFM (Zhu et al.
2010), and the FSDAF model (Zhu et al. 2016).

PS_SRM and SR_SRM use only a mono-temporal coarse spatial resolution image
as input and hence do not exploit the temporal information in the land cover. By contrast,
ESTARFM uses a coarse spatial resolution image and pairs of coarse and fine spatial
resolution images that pre-and post-date it as input. ESTARFM is based on the
assumption that remotely sensed data from different satellite sensors observed on the
same, or at least very close, date are mutually comparable and correlated, and uses the
correlation to blend multi-source data and minimize the system biases. The FSDAF,
which is based on spectral unmixing analysis and a thin plate spline interpolator, is also
used for comparison. It requires only one pair of fine and coarse spatial resolution
images that pre- or post-date the coarse-resolution image.

ESTARFM and FSDAF output a fine spatial resolution reflectance image rather
than a land cover map. The fine spatial resolution image produced may then be classified. Two types of classification methods were applied. The first one is the Maximum Likelihood Classifier (MLC), which is one of the statistical classifiers that relies on the second-order statistics of a Gaussian probability density function for the distribution of the feature vector of each class. In MLC, a pixel is allocated to the class with which it has the highest likelihood of membership (Richards and Jia 1999). The produced final fine spatial resolution land cover maps by MLC, which are referred to as ESTARFM_MLC and FSDAF_MLC, were compared with STIMFM.

The second used classification method is the automated Land Cover updating approach based on integrated change detection and classification methods (LCupdating) produced by Chen et al. (2012). MLC used only the fused fine-resolution image as input but ignored the available fine-resolution image and land cover map that pre- or post-dates the coarse-resolution image. LCupdating was applied to the fused image from ESTARFM and FSDAF by incorporating the fine-resolution remotely sensed image and land cover map that pre-date the coarse-resolution image. LCupdating first detects changes between the input and fused fine-resolution images from ESTARFM or FSDAF and then predicts the changed pixel labels in the fused image based on the Markov random field based classifier. The ESTARFM and FSDAF incorporating LCupdating methods (ESTARFM_LCupdating and FSDAF_LCupdating) were compared with STIMFM.

The parameters of these different methods were set according to results reported in the literature and through trial and error. The STIMFM spatial weight parameter $\lambda_s$
and temporal weight parameter $\lambda_\tau$ were set to 0.05. The neighborhood window size $W$ in the STIMFM spatial constraint function was set to $2\times s-1$ (Tolpekin and Stein 2009). The number of unchanged coarse pixels, $m$, in STIMFM endmember estimation was set to 100.

3.1 Landsat–NLCD experiment

3.1.1 Data preparation

This experiment used Landsat Thematic Mapper (TM) multi-spectral images and NLCD land cover maps. The NLCD is a land cover classification scheme of Albers Equal Area projection, which has been applied consistently at a spatial resolution of 30 m across the conterminous USA primarily on the basis of Landsat satellite data. NLCD maps for the years 2001, 2006, and 2011 were used in this experiment. The NLCD 2001 was based primarily on a decision tree classification of 2001 Landsat satellite data. The NLCD 2006 and 2011 were based primarily on a decision tree classification from 2006 and 2011 Landsat satellite data, and also quantified land cover change from 2001 to 2006 and 2006 to 2011 (Homer et al. 2015; Jin et al. 2013; Xian et al. 2009). The original sixteen classes were reclassified into eight classes (Fig. 3). Subset land cover maps, each with a size of $2000 \times 2000$ pixels (centered at 34°40'0"N and 79°27'00"W), were acquired from NLCD 2001, 2006, and 2011 [Fig. 3(b–d)].
A Landsat TM image (path 016, row 036) acquired on April 9, 2006 in the study area was downloaded from the United States Geological Survey (USGS). This Landsat image was re-projected to the Albers Equal Area projection, and six spectral bands at the spatial resolution of 30 m (the 120 m thermal infrared band was excluded) were used to extract the same 2000 × 2000 pixel area that was identified in the NLCD maps [Fig. 3(a)]. The subset image was calibrated to surface reflectance (Gao et al. 2006; Masek et al. 2006) and then spatially degraded to simulate a coarse spatial resolution multi-spectral image using a scale factor s=8 [Fig. 3(f), 240 m] with a mean filter. The NLCD 2006 [Fig. 3(c)] was used as the reference map used for accuracy assessment. The pixels that changed land cover class from 2001 to 2011 accounted for 12.08% of all fine spatial resolution pixels.

For analyses with the PS_SRM and SR_SRM, only the degraded multi-spectral image [Fig. 3(f)] was needed as input. For the STIMFM, the required input included the degraded multi-spectral image [Fig. 3(f)] and the NLCD 2001 and NLCD 2011 land cover maps [Fig. 3(b), 3(d)]. For ESTARFM, pairs of fine and coarse spatial resolution multi-spectral images that temporally pre- and post-date the 2006 coarse-resolution
remotely sensed image were needed. To obtain the required data, a Landsat TM image acquired on April 17, 2001 and a Landsat TM image acquired on April 7, 2011 were also downloaded, re-projected, subsetted, and calibrated. The original 30 m spatial resolution reflectance images with six spectral bands (the 120 m thermal infrared band was excluded) were spatially degraded to simulate their corresponding coarse spatial resolution multi-spectral images at scale factors $s=8$, respectively. Therefore, the input to the ESTARFM_MLC and ESTARFM_LC updating included fine and coarse spatial resolution multi-spectral image pairs in 2001 and 2011 and the coarse spatial resolution multi-spectral image for 2006. The input to FSDAF_MLC and FSDAF_LC updating included fine and coarse spatial resolution multi-spectral image pairs in 2001 and the coarse spatial resolution multi-spectral image in 2006. In ESTARFM_LC updating and FSDAF_LC updating, the NLCD 2001 fine-resolution land cover map was also used as the base data.

3.1.2 Results

![Input and result maps for the zoomed area in the Landsat–NLCD experiment.](image)

**Fig. 4** Input and result maps for the zoomed area in the Landsat–NLCD experiment.
The land cover maps produced from the different methods are shown in Fig. 3 for the entire area and in Fig. 4 for the zoomed area [320 × 320 pixel area in Fig. 3(a)]. In the zoomed area, the PS_SRM map contained many speckle-like artifacts [Fig. 4(g)]. This situation arises because the spectral unmixing may determine a small fractional cover of a class that is actually absent in a coarse-resolution pixel, and this fraction must be maintained in the result. The SR_SRM map contained fewer speckle-like artifacts than PS_SRM, because SR_SRM relaxed the constraint of land cover fraction maintenance [Fig. 4(h)]. However, the maximal spatial dependence model used in SR_SRM also led to rounded land cover patches. Compared with PS_SRM and SR_SRM, more spatial detail of the land cover mosaic was retained in the ESTARFM, FSDAF, and STIMFM maps. Many speckle-like artifacts in the ESTARFM_MLC [Fig. 4(i)] and FSDAF_MLC [Fig. 4(k)] maps existed because MLC is a per-based classification method, and the spatial context information was not used. ESTARFM_MLC and FSDAF_MLC incorrectly classified cases that have similar reflectance values, such as “forest”, “herbaceous”, and “wetlands”, in the result maps [Figs. 4(i), (k)].

![Fig. 5 Landsat, ESTARFM and FSDAF images in the zoomed area for the Landsat–NLCD experiment.](image)

In contrast to ESTARFM_MLC and FSDAF_MLC, ESTARFM_LCupdating [Fig. 4(j)] and FSDAF_LCupdating [Fig. 4(l)] quantified the land cover changes between 2001 and 2006 and generated land cover maps that were more similar to the reference...
map [Fig. 4(c)]. The labels of pixels that were detected as unchanged by LCUpdation were preserved in the ESTARFM_LCUpdation and FSDAF_LCUpdation maps. The labels of changed pixels were determined based on the Markov random field based classifier, which considers the contextual information in classification. Thus, most speckle-like artifacts were eliminated in ESTARFM_LCUpdation and FSDAF_LCUpdation. However, many changed pixel labels were incorrectly predicted by ESTARFM_LCUpdation and FSDAF_LCUpdation. “Herbaceous” was incorrectly labeled as “developed” in the ESTARFM_LCUpdation highlighted by the black circle [Fig. 4(j)], and the linear-shaped “developed” in the FSDAF_LCUpdation highlighted by the black circle was disconnected [Fig. 4(l)]. The predicted reflectance for pixels of changed land cover for “herbaceous” and “planted/cultivated” in ESTARFM [e.g., those highlighted by the red circle in Fig. 5(f)] was dissimilar to that in the Landsat 2006 reference image [Fig. 5(b)] because ESTARFM cannot capture abrupt land cover changes (Zhu et al. 2010), and the predicted reflectance of linear-shaped “developed” land cover was similar to that of “planted/cultivated” in the FSDAF image highlighted by the red circle in Fig. 5(g), because FSDAF cannot capture tiny land cover changes (Zhu et al. 2016). By contrast, the STIMFM land cover map as shown in Fig. 4(m) was quite similar to the reference map, and the detailed land cover patterns were well represented. STIMFM correctly predicted the class labels not only for almost all pixels of unchanged land cover but also for most of those pixels for which land cover class had changed, such as those highlighted in the red circle in Fig. 5(d).
The overall accuracies of different methods are shown in Table 1. The result maps were compared with the NLCD 2006. The overall accuracy of STIMFM is higher than those obtained from the other methods. Table 1 also shows the accuracies of pixels of changed and unchanged land cover (PULC means the percentage of correctly labeled pixels of unchanged land cover among all pixels of unchanged land cover, and PCLC means the percentage of correctly labeled pixels of changed land cover among all pixels of changed land cover) obtained from the different methods. For PS_SRM and SR_SRM, which applied a mono-temporal remotely sensed image, no obvious difference was found between PULC and PCLC values. For ESTARFM, FSDAF, and STIMFM applied to multi-temporal data, the PULC values were higher than the PCLC values. These results indicate that extracting changed land cover information is more difficult than extracting unchanged land cover information from ESTARFM, FSDAF, and STIMFM. STIMFM integrates the temporal dependence model in its objective function, and the fine spatial resolution pixel class labels are temporally dependent on

<table>
<thead>
<tr>
<th>Method</th>
<th>OA</th>
<th>PULC</th>
<th>PCLC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS_SRM</td>
<td>41.61</td>
<td>41.54</td>
<td>42.13</td>
</tr>
<tr>
<td>SR_SRM</td>
<td>49.10</td>
<td>49.01</td>
<td>49.77</td>
</tr>
<tr>
<td>ESTARFM_MLC</td>
<td>33.57</td>
<td>35.10</td>
<td>22.44</td>
</tr>
<tr>
<td>ESTARFM_LCupdating</td>
<td>88.28</td>
<td>94.89</td>
<td>40.18</td>
</tr>
<tr>
<td>FSDAF_MLC</td>
<td>33.60</td>
<td>34.29</td>
<td>28.60</td>
</tr>
<tr>
<td>FSDAF_LCupdating</td>
<td>89.50</td>
<td>96.15</td>
<td>41.08</td>
</tr>
<tr>
<td>STIMFM</td>
<td>94.89</td>
<td>99.24</td>
<td>63.27</td>
</tr>
</tbody>
</table>
those in the pre- and post-dated fine-resolution land cover maps. If the fine-resolution pixel class labels are unchanged during the observation period, then STIMFM could make the best use of pixel class labels in the fine-resolution maps that pre- and post-date the coarse-resolution image. Thus, the accuracies for classes with unchanged class labels are high. By contrast, if the fine-resolution pixel class labels have changed during the observation period, then STIMFM could not make the best use of pixel class labels in the fine-resolution maps that pre- and post-date the coarse-resolution image, and the accuracies for classes with changed class labels are relatively low. The PULC was higher than 99%, and the PCLC was higher than 63% for STIMFM; these values are higher than those obtained from the other methods.

3.2 MODIS–Landsat experiment

3.2.1 Data preparation

The study area was located near Sorriso (12°33′00″S and 55°42′00″W) in Mato Grosso State, Brazil. This area was mainly covered by tropical forests but has suffered from deforestation in recent years (Hansen et al. 2008). This experiment used eleven coarse spatial resolution MODIS images and two fine spatial resolution land cover maps that pre- and post-date the coarse spatial resolution image series as input and outputs eleven fine-resolution land cover maps with MODIS repetition rates to show the fine spatial and temporal deforestation process in the study area. Landsat Enhanced Thematic Mapper Plus (ETM+) images (path 226, row 069) acquired on 2002/06/08 and 2002/09/12 were downloaded from USGS [Fig. 6(d) and (f)]. Data in six bands (the
120 m thermal infrared band was excluded) at the 30 m spatial resolution with the Universal Transverse Mercator projection were used and calibrated to surface reflectance values (Gao et al. 2006; Masek et al. 2006). One cloud-free Landsat ETM+ image acquired on 2002/07/10 was used for accuracy assessment [Fig. 6(e)]. A total of thirteen eight-day surface reflectance MODIS product (MOD09A1) datasets that comprise seven spectral bands (620 nm–2055 nm) with a spatial resolution of 463 m acquired from 2002/06/02 to 2002/09/13 were downloaded from USGS (Walker et al. 2012). The MODIS images were re-projected into the UTM coordinate system and resampled to a spatial resolution of 450 m using the nearest neighbor interpolation, and were adopted as the coarse spatial resolution multi-spectral images required for the analyses. The study area covers 300 × 300 MODIS pixels, which correspond to 4500 × 4500 Landsat pixels, with a scale factor s=15.

Fig. 6 MODIS, Landsat images, and reference maps in the MODIS–Landsat experiment from 2002/06/08 to 2002/09/12.

The three Landsat images were classified to produce land cover maps with a 30 m
spatial resolution [Fig. 6(g)–(i)]. Two land cover classes, forest and nonforest, were considered in this experiment. The endmembers of each class were manually selected from each Landsat image, and MLC was applied to generate the fine spatial resolution forest/nonforest reference maps. The fine-resolution change maps that produced by a per-pixel comparison of maps in Fig. 6(g)–(i) are shown in Fig. 6(j)–(k). The pixels that changed land cover class from 2002/06/08 to 2002/09/12 accounted for 4.30% of all fine spatial resolution pixels.

STIMFM used the MODIS multi-spectral image series from 2002/06/10 to 2002/09/05 and the 2002/06/08 and 2002/09/12 fine spatial resolution land cover maps in Fig. 6(g) and (i) as input and predicted a series of land cover maps at 30 m spatial resolution with MODIS repetition rates during this period. The accuracy was assessed using the 2002/07/10 land cover map [Fig. 6(h)]. The STIMFM was compared with PS_SRM, SR_SRM, ESTARFM_MLC, ESTARFM_LCupdating, FSDAF_MLC, and FSDAF_LCupdating using the 2002/07/10 land cover map in Fig. 6(h) for assessment. In these methods, the eight-day composite MODIS image [2002/07/04–2002/07/11 in Fig. 6(b)] was used as the coarse-resolution image. Aside from this data, ESTARFM used the eight-day composite MODIS images [2002/06/02–2002/06/09 in Fig. 6(a) and 2002/09/06–2002/09/13 in Fig. 6(c)] and Landsat multi-spectral images [2002/06/08 in Fig. 6(d) and 2002/09/12 in Fig. 6(f)] as input, and FSDAF used the eight-day composite MODIS image [2002/06/02–2002/06/09 in Fig. 6(a)] and Landsat multi-spectral image [2002/06/08 in Fig. 6(d)] as input. In ESTARFM and FSDAF, the MODIS bands 1–4 and 6–7 were used in ESTARFM and FSDAF, because no similar
spectral band of Landsat image was observed from MODIS band 5. The 2002/06/08 fine spatial resolution land cover map in Fig. 6(g) was also inputted in the ESTARFM_LC updating and FSDAF_LC updating.

3.2.2 Results
Table 2

Overall accuracies (OAs) and accuracies of different methods in predicting PULC and PCLC in the MODIS–Landsat experiment. The MODIS image used in different methods was the eight-day composite data from 2002/07/04 to 2002/07/11.

<table>
<thead>
<tr>
<th>Method</th>
<th>OA</th>
<th>PULC</th>
<th>PCLC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS_SRM</td>
<td>88.14</td>
<td>89.29</td>
<td>62.52</td>
</tr>
<tr>
<td>SR_SRM</td>
<td>89.28</td>
<td>90.42</td>
<td>63.73</td>
</tr>
<tr>
<td>ESTARFM_MLC</td>
<td>95.17</td>
<td>96.90</td>
<td>56.71</td>
</tr>
<tr>
<td>ESTARFM_LCupdating</td>
<td>96.32</td>
<td>98.07</td>
<td>57.42</td>
</tr>
<tr>
<td>FSDAF_MLC</td>
<td>95.07</td>
<td>96.72</td>
<td>58.23</td>
</tr>
<tr>
<td>FSDAF_LCupdating</td>
<td>96.61</td>
<td>98.37</td>
<td>57.38</td>
</tr>
<tr>
<td>STIMFM</td>
<td>98.27</td>
<td>99.69</td>
<td>66.67</td>
</tr>
</tbody>
</table>

The OA, PULC, and PCLC values obtained from the application of the different methods are shown in Table 2. The overall accuracies obtained from the PS_SRM and SR_SRM were lower than 90%, whereas the overall accuracies of ESTARFM_MLC, ESTARFM_LCupdating, FSDAF_MLC, and FSDAF_LCupdating were higher than 95%. These findings indicate that the classification from fine-resolution image extracted by spatial–temporal fusing of coarse and fine-resolution images can better improve the accuracy compared with SRM applied to a mono-temporal coarse-resolution image. The OA value for STIMFM was 98.27%, which is higher than all the other methods. The PCLC values were lower than the PULC values for ESTARFM, FSDAF, and STIMFM methods, which is similar to those in the Landsat–NLCD experiment. The STIMFM has the highest PULC value, which was 99.69%, and the highest PCLC value, which was 66.67%, among all the methods.
Fig. 7 Input, reference, and result images and maps for the zoomed area at different years for the MODIS–Landsat experiment. The MODIS image used in different methods was the eight-day composite data from 2002/07/04 to 2002/07/11.

The reference, input, and result images and maps in the zoomed area are shown in Fig. 7. A part of the forest patch (highlighted by a blue circle in regions A and B in Fig. 7) changed to nonforest from 2002/06/08 to 2012/07/10 [Fig. 7(j)], and a part of the forest patch (highlighted by a blue circle in region C in Fig. 7) changed to nonforest from 2002/07/10 to 2012/09/12 [Fig. 7(k)]. The PS_SRM map contained many speckle-like artifacts [Fig. 7(o)], and SR_SRM contained land cover patches with oversmoothed rounded boundaries [Fig. 7(p)]. In the ESTARFM and FSDAF fused images [Fig. 7(l), (q)], the pixels of unchanged land cover considerably resemble those in the reference
Landsat image, whereas the pixels of changed land cover (highlighted by blue circles in Fig. 7) were noticeably different from those in the reference Landsat image [Fig. 7(e)]. As a result, these pixels of changed land cover were erroneously classified in the ESTARFM_MLC, ESTARFM_LCupdating, FSDAF_MLC, and FSDAF_LCupdating results [Figs. 7(m), (n), (r), and (s)]. By contrast, most of the changed and unchanged pixels are correctly allocated by STIMFM [Fig. 7(t)], thereby showing the ability of the proposed STIMFM model in the reconstruction of land cover trajectories for pixels of both changed and unchanged land cover. The land cover changes in Fig. 8 were extracted by comparing the STIMFM predicted maps and input fine-resolution land cover map that pre-dates the coarse images [Fig. 6(g)]. The colors in Fig. 8 indicate the date when the pixels begin to change. The forest area decreased gradually, whereas the nonforest area increased in Fig. 9. With STIMFM, the detailed spatial extent information and the change of areas for different classes can be extracted, thereby showing the effectiveness of the proposed method.
Fig. 8 30 m spatial extent of land cover change with MODIS repetition rates derived from STIMFM. The colors represent the date when pixels begin to change. “unchan or prev chan” marked as white color means unchanged or previously changed before 2002/06/08.

Fig. 9 Forest and nonforest areas extracted using STIMFM in the MODIS–Landsat experiment.

3.3 Landsat–GEI experiment

The study area was located in Wuhan (30°27′30″N and 114°32′30″E), Hubei province, China. This area underwent rapid urbanization in 2010–2016. This experiment used eleven cloud-free 30 m spatial resolution Landsat-8 Operational Land Imager (OLI) multi-spectral images (path 123, row 039) from 2013 to 2015 and two
5 m spatial resolution land cover maps acquired in 2012 and 2016 as input. Eleven 5 m resolution land cover maps during 2013–2015 were predicted to show the fine spatial and temporal urbanization process in the study area. The acquired eleven Landsat OLI images were downloaded from USGS. The first seven bands of OLI image with a spatial resolution of 30 m were selected. Two GEIs acquired on 2012/04/26 and on 2016/02/20 [Figs. 10(a), (b)] with a spatial resolution of 5 m were re-projected into the UTM coordinate system and digitized into the 5 m land cover maps [Figs. 10(c), (d)]. Four land cover classes, namely, water, vegetation, bareland, and urban, were found in the fine-resolution maps. The study area covers 320 × 450 Landsat pixels, which correspond to 1920 × 2700 fine-resolution pixels in Figs. 10(c) and (d), with a scale factor $s = 6$. The land cover change map from 2012/04/26 to 2016/02/20 is shown in Fig. 10(e). The pixels that changed land cover class accounted for 23.49% of all fine spatial resolution pixels from 2012/04/26 to 2016/02/20.
STIMFM was used to produce the eleven 5 m resolution land cover maps with Landsat repetition rates during 2013–2015 using the eleven cloud-free Landsat images and two 5 m land cover maps on 2012/04/26 and 2016/02/20 as input. The STIMFM accuracy was assessed using a 5 m fine-resolution land cover map, which was produced according to a GEI at the spatial resolution of 5 m acquired on 2014/12/06 [Fig. 11(b)]. This GEI is the only fine-resolution one available in the study area during 2012–2016 and was re-projected into the UTM coordinate system and digitized to the reference land cover map [Fig. 11(c)]. STIMFM was compared with PS_SRM and SR_SRM, which were applied to a single-date Landsat OLI image acquired on 2014/10/06 [Fig. 11(a)]; this image is temporally closest to the GEI in 2014 [Fig. 11(b)]. ESTARFM and FSDAF were not used for comparison because they require the coarse- and fine-resolution images to have comparable and correlated reflectance bands, whereas Landsat and GEI have different spectral bands and the GEI can hardly be transformed.
into reflectance images, which are correlated to the Landsat images.

**Table 3**

Overall accuracies (OAs) and accuracies of different methods in predicting PULC and PCLC in the Landsat–GEI experiment. The Landsat image used for assessment in the different methods was acquired on 2014/10/06.

<table>
<thead>
<tr>
<th>Method</th>
<th>OA</th>
<th>PULC</th>
<th>PCLC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS_SRM</td>
<td>72.71</td>
<td>76.22</td>
<td>61.13</td>
</tr>
<tr>
<td>SR_SRM</td>
<td>73.73</td>
<td>77.29</td>
<td>61.99</td>
</tr>
<tr>
<td>STIMFM</td>
<td>94.31</td>
<td>99.61</td>
<td>76.81</td>
</tr>
</tbody>
</table>

The OA accuracies were lower than 74% for PS_SRM and SR_SRM and increased to 94.31% for STIMFM (Table 3). The PULC value was higher than 99%, and the PCLC value was higher than 76% for STIMFM; these values were obviously higher than those for PS_SRM and SR_SRM. The PULC values were higher than the PCLC values for STIMFM because STIMFM could make the best use of unchanged pixel labels in the fine-resolution maps that pre- and post-date the Landsat images in land cover mapping.
Fig. 1 Landsat image, Google Earth image, reference and result maps in the Landsat–GEI experiment.

The Landsat image used in the different methods was acquired on 2014/10/06.

The PS_SRM contained many speckle-like artifacts [Fig. 11(d)]. Many speckle-like artifacts were smoothed to rounded patches in SR_SRM [Fig. 11(e)]. The linear-shaped urban objects were discrete in PS_SRM and SR_SRM and connected in STIMFM [Fig. 11(f)]. In STIMFM, most speckle-like artifacts and rounded patches were eliminated, and the spatial pattern of most patches was close to the reference map [Fig. 11(c)].

The 5 m spatial extent of land cover change with Landsat repetition rates derived from STIMFM is shown in Fig. 12, in which the explicit time of land cover change and the detailed spatial extent of urbanization process at fine spatial and fine temporal resolutions are obvious. Fig. 13 shows the areas of different classes extracted using STIMFM. The water, vegetation, and bareland areas decreased, whereas the urban area increased from April 2012 to May 2013. The areas of different classes remained almost unchanged from May to September 2013. Since October 2013, the vegetation area decreased, whereas the water, bareland, and urban areas increased.
Fig. 12 5 m spatial extent of land cover change with Landsat repetition rates derived from STIMFM. The colors represent the date when pixels begin to change. “unchan or prev chan” marked as white color means unchanged or previously changed before 2012/04/26.

Fig. 13 Areas of different classes extracted using STIMFM in the Landsat–GEI experiment.
4. Discussion

Results show that STIMFM is a promising approach for the production of a series of fine spatial–temporal resolution land cover maps, which were achieved by fusing a series of coarse spatial resolution remotely sensed images with a limited set of fine spatial resolution land cover maps. The mapping accuracies of STIMFM reached relatively high levels in all three experiments. Compared with popular state-of-the-art SRM algorithms that are generally applied on mono-temporal remotely sensed image, STIMFM can produce land cover maps of a much higher accuracy as expected, because fine spatial resolution land cover temporal information is incorporated into its analysis.

Compared with ESTARFM and FSDAF, STIMFM predicted the labels of both changed and unchanged pixels with higher accuracy in the Landsat–NLCD and MODIS–Landsat experiments. In the Landsat–GEI experiment, STIMFM predicted a sequence of fine spatial–temporal resolution land cover maps from eleven Landsat images and two GEIs, to which ESTARFM and FSDAF that require correlation in reflectance bands in coarse and fine images cannot be applied.

Although ESTARFM, FSDAF, and STIMFM aim to extract high spatial–temporal resolution information, they have important differences that affect practical application. First, they have different assumptions and thus use different inputs. ESTARFM and FSDAF require coarse- and fine-resolution remotely sensed images from different satellite sensors observed at the same or similar date to have comparable and highly
correlated reflectance bands. Thus, only a limited set of images can be used in
ESTARFM and FSDAF, such as Landsat and MODIS, thereby limiting the application
of ESTARFM and FSDAF. For instance, panchromatic aerial photographs with a very
high spatial resolution cannot be used with Landsat image with a 30 m resolution to
generate fine-resolution land cover maps with Landsat repetition rates from ESTARFM
and FSDAF because the aerial photographs and Landsat images have different spectral
bands. By contrast, STIMFM does not require similar coarse and fine spatial resolution
images, but directly considers the relationship between the land cover classes
themselves and not their spectral response. The coarse spatial resolution images are
unmixed to land cover fractions, and STIMFM is built on the analysis of land cover
spatial and temporal dependences in the different images instead of analyzing the
relationship of pixel spectral values in different images. In addition, ESTARFM and
FSDAF require one or more observed pairs of coarse- and fine-resolution images
acquired at the same or similar date for training, whereas STIMFM does not need the
course-resolution images at the acquisition data of the fine-resolution maps as input.

Second, ESTARFM, FSDAF, and STIMFM have different outputs; the output of
ESTARFM and FSDAF are multi-spectral reflectance images, whereas the output of
STIMFM are land cover maps. If the aim is to generate spectral images, then
ESTARFM and FSDAF are suitable. For instance, unlike the STIMFM result, the
ESTARFM and FSDAF result can be used in the analysis of phenology change.
STIMFM produces land cover maps with discrete class labels and is more suitable in
monitoring the spatial distribution pattern and temporal change trajectory of land cover
classes at a fine spatial and temporal resolution. Although the reflectance images output from ESTARFM and FSDAF can be further classified to produce land cover maps, problems still exist. To generate fine spatial resolution land cover map series, ESTARFM and FSDAF should generate a series of fine spatial resolution multi-spectral reflectance images, thereby requiring massive storage for these intermediate data when the study area is large. In addition, the generation of land cover maps from these image series requires a large amount of training data, which are difficult to collect in practice. The classification of reflectance images is also often underdetermined and contains a large solution space. By contrast, STIMFM is modeled based on the spatial–temporal character of pixel class labels. It does not produce intermediate fine spatial resolution multi-spectral image series, and the endmembers could be automatically estimated for each coarse spatial resolution image on the basis of optimization approach. The STIMFM has a simple objective function and comprises only few parameters and is thus relatively easy to use. As a result, STIMFM is more suitable in the reconstruction of fine spatial and temporal resolution land cover maps compared with ESTARFM and FSDAF.

Although STIMFM provides a great opportunity to enhance studies of land cover and its dynamics, its performance is dependent on several factors. In STIMFM, the analysis of land cover class fraction temporal change is conducted by comparing the coarse-resolution fraction images produced from spatial degrading the input fine-resolution maps and from spectral unmixing of the coarse-resolution image. First, fraction images extracted from spectral unmixing probably have errors and
uncertainties, which will affect the land cover fraction temporal change analysis in STIMFM. The linear mixture model is used to estimate the endmembers and generate fraction images for class fraction change analysis in STIMFM. However, this approach is not always ideal, because the mixing may be nonlinear. Some nonlinear mixture models may be applied to decrease the fraction image error and improve the class fraction temporal change accuracy. Second, the accuracy of the fine-resolution land cover maps that pre- and post-dated the coarse-resolution images also affect the STIMFM accuracy. The overall accuracy of STIMFM decreases with the increase in the number of incorrect pixel labels in the input fine-resolution land cover maps because STIMFM labeled the unchanged pixels according to the labels in the fine-resolution maps. In addition, the incorrect pixel labels in the fine-resolution maps would decrease the accuracy in the class fraction temporal change analysis and thus decrease the STIMFM accuracy. Advanced classifiers such as object-based classifiers should be used to extract accurate land cover maps from the fine-resolution images used as STIMFM input. Third, the co-registration between the fine-resolution land cover maps and the coarse-resolution images plays a key role because misregistration would lead to inaccurate detection of fraction changes of each class in each coarse pixel. Advanced methods such as the sub-pixel scale co-registration method should be developed and applied in STIMFM.

The STIMFM performance is also affected by the model functions and parameters. First, in the STIMFM spatial constraint function, the \textit{a priori} land cover spatial distribution model has a major role in the prediction of fine spatial resolution land cover
spatial pattern. The land cover maximal spatial dependence model is used as the \textit{a priori} land cover spatial pattern information in STIMFM for its simplicity. However, this \textit{a priori} information is used for all classes although they may actually have different spatial patterns and is most suitable for the situation in which land cover patches are larger than the coarse spatial resolution pixel size. More \textit{a priori} information could be introduced to characterize the spatial pattern of the classes in STIMFM. Second, the STIMFM performance is dependent on the spatial and temporal weights $\lambda_S$ and $\lambda_T$. When the spatial weight $\lambda_S$ is relatively large, STIMFM would decrease the influence of temporal information, and the STIMFM result would be dominated by the spatial constraint function and resemble the SR_SRM result. By contrast, when the temporal weight $\lambda_T$ is relatively large, the STIMFM result would be dominated by the fine spatial resolution maps that temporally pre- and post-dated the coarse-resolution images, and the spatial pattern of land cover patches would be difficult to reconstruct in the result maps. The optimal $\lambda_S$ value can be automatically estimated through quantification of the effects of land cover class spectral separability (Li et al. 2016; Tolpekin and Stein 2009), whereas the estimation of optimal $\lambda_T$ value has not been studied to our knowledge. In this paper, the optimal $\lambda_S$ and $\lambda_T$ values were determined through many trials. In practice, a subset of coarse spatial resolution images and fine spatial resolution maps are usually available, and these data can be used to estimate the optimal $\lambda_S$ and $\lambda_T$ values in STIMFM. Finally, a selected number ($m \times C$) of purest coarse-resolution pixels are used to estimate the endmembers $E$ in $Y$. In practice, $m$ can be set in the range about 100–200 if $Y$ is a multi-spectral image.
5. Conclusion

In this paper, a novel spatial–temporal remotely sensed images and land cover maps fusion model was proposed. This model aims to produce a series of fine spatial–temporal resolution land cover maps from a series of coarse spatial resolution remotely sensed images and a few fine spatial resolution land cover maps. In STIMFM, the endmember spectra of different land cover classes are estimated automatically for each coarse spatial resolution image with the aid of available fine spatial resolution land cover maps. Using the estimated endmember spectra, an objective function, which incorporates the pixel spectral and land cover spatial and temporal information, is constructed. The output of STIMFM is achieved by solving the optimization problem.

The performance of STIMFM was explored using three experiments and compared with that of several popular state-of-the-art algorithms. The STIMFM has comparable efficiency with ESTARFM and FSDAF in terms of computing time. STIMFM can produce land cover maps with higher accuracies than those algorithms used for comparison. The overall accuracies of STIMFM are higher than 94% in all experiments reported. Results indicate that STIMFM is a promising approach for generating land cover maps and estimating land cover change at both fine spatial and temporal resolutions. Although issues that would benefit from further research exist, this novel land cover fusion method provides a great opportunity to enhance studies of land cover and its dynamics.
Acknowledgements

This work was supported in part by the Youth Innovation Promotion Association CAS (Grant No. 2017384), in part by the Natural Science Foundation of China (Grant No. 61671425), in part by the Distinguished Young Scientist Grant of the Chinese Academy of Sciences, and in part by the State Key Laboratory of Resources and Environmental Informational System. The authors would like to thank Xiaolin Zhu for providing the ESTARFM and FSDAF programs and Xuehong Chen for providing the LCupdating program.
References


LIST OF FIGURE CAPTIONS

Fig. 1 Production of a series of fine spatial and temporal resolution land cover maps from a series of coarse spatial resolution remotely sensed images and a few fine spatial resolution land cover maps in STIMFM.

Fig. 2 Flowchart of STIMFM.

Fig. 3 Input and result maps for the entire study area in the Landsat–NLCD experiment.

Fig. 4 Input and result maps for the zoomed area in the Landsat–NLCD experiment.

Fig. 5 Landsat, ESTARFM and FSDAF images in the zoomed area for the Landsat–NLCD experiment.

Fig. 6 MODIS, Landsat images, and reference maps in the MODIS–Landsat experiment from 2002/06/08 to 2002/09/12.

Fig. 7 Input, reference, and result images and maps for the zoomed area at different years for the MODIS–Landsat experiment. The MODIS image used in different methods was the eight-day composite data from 2002/07/04 to 2002/07/11.

Fig. 8 30 m spatial extent of land cover change with MODIS repetition rates derived from STIMFM. The colors represent the date when pixels begin to change. “unchan or prev chan” marked as white color means unchanged or previously changed before 2002/06/08.

Fig. 9 Forest and nonforest areas extracted using STIMFM in the MODIS–Landsat experiment.

Fig. 10 Google Earth images, land cover maps and change maps in the Landsat–GEI experiment.

Fig. 11 Landsat image, Google Earth image, reference and result maps in the Landsat–GEI experiment.

The Landsat image used in the different methods was acquired on 2014/10/06.

Fig. 12 5 m spatial extent of land cover change with Landsat repetition rates derived from STIMFM. The colors represent the date when pixels begin to change. “unchan or prev chan” marked as white color
means unchanged or previously changed before 2012/04/26.

**Fig. 13** Areas of different classes extracted using STIMFM in the Landsat–GEI experiment.