Missing Data and Regression Models for Spatial Images

Jun Zhang, Murray K. Clayton, and Philip A. Townsend

***Abstract*—In previous work, we have shown that a functional concurrent linear model (FCLM) can be used to model the relationship between two spatial images. In this paper, we provide two extensions of the use of the FCLM to address missing data problems in series of colocated spatial images. First, we show how to build an FCLM relating two images involving gypsy moth defoliation data when there are missing data in some regions of the images. Because there is interest in filling in the missing scan lines in Landsat 7 images, we then further extend this approach to provide an imputation method for Landsat 7 data when the focus is on repairing a single image, rather than in relating images. A side effect of our approach is that the FCLM appears to automatically select the best parts of different covariate images for repairing a target image.**

***Index Terms*—Functional concurrent linear model (FCLM), missing data, wavelet.**

I. Introduction

PREVIOUSLY [12], we introduced a functional concurrent linear model (FCLM) for 2-D spatial images. The FCLM was proposed as a tool for modeling the relationship between two colocated spatial images, an example was given wherein it was desired to use a spatial image of elevation as a covariate to explain a colocated spatial image of gypsy moth defoliation. In a sense, the FCLM can be thought of as a method for regressing one spatial image on another.

In the notation of [12], we write the FCLM as

$$Y = B_0 + X_1 \circ B_1 + X_2 \circ B_2 + \cdots + X_p \circ B_p + E \quad (1)$$

where Y is the response image; $X_1, X_2, \ldots,$ and X_p are p explanatory images; $B_0, B_1, \ldots,$ and B_p are $p+1$ coefficient matrices; and E is the error matrix. All matrices in (1) are K by N matrices. "$\circ$" stands for the Schur product or element-wise matrix multiplication.

In the example in [12], Y represented defoliation rates over a region in the Appalachian mountains from [1]; initially with one explanatory variable, X_1 represented elevation for the same region. The modeling showed that defoliation was indeed related to elevation, although it was useful to include a tree species classification as an additional covariate X_2.

Pixel-wise, the model in (1) is ill-defined. To address this, [12] used wavelets to represent the matrices $B_0, B_1, \ldots,$ and B_p, and then used LASSO to choose the wavelet coefficients (and corresponding wavelets) that would remain in the model. This has the effect of smoothing, or reducing the dimension, of the coefficient matrices. The LASSO problem can be expressed as

$$\min_{\beta} \quad \|Y - \hat{Y}\|_F^2 + \lambda \|\beta\|_1 \quad (2)$$

where $\|\cdot\|_F$ is the Frobenius matrix norm. To choose λ, [12] used Bayesian Information Criterion (BIC).

More explicitly, if $p = 1$, then in terms of the wavelet bases, the FCLM looks like

$$\begin{aligned} Y &= \sum_{k=1}^{L}\sum_{j=1}^{N_k} v_{kj}\phi_{kj} + X_1 \circ \left\{ \sum_{k=1}^{L}\sum_{j=1}^{N_k} w_{kj}\phi_{kj} \right\} + E \\ &= \sum_{k=1}^{L}\sum_{j=1}^{N_k} v_{kj}\phi_{kj} + \sum_{k=1}^{L}\sum_{j=1}^{N_k} w_{kj}\{X_1 \circ \phi_{kj}\} + E \end{aligned}$$

where L is the highest level of wavelet bases, N_k is the number of wavelet bases for level k, v_{kj}, and w_{kj} are the wavelet coefficients, and ϕ_{kj} is the wavelet basis function. If we remove the first or finest level from the model, then we remove three quarters of the parameters. We have found that this speeds up the computational aspects of the problem. If we have relatively large images and find it safe to assume that the parameter surfaces do not have many very fine details, then using an FCLM with partial wavelet bases will efficiently accelerate the search for λ_{opt}. Reference [12] discuss this and other methods for reducing computational time in the modeling.

In the examples and approach of [12], it was assumed that there were no missing data in the images. However, in many of the images for defoliation mapping or other change detection applications, there are missing or uninformative pixels due to various causes, including instrument errors, clouds, cloud shadows, snow, sun glint, or other image anomalies or the presence of nonforest such as fields, rivers and lakes, buildings, etc. More generally, in many applications missing data are common, and therefore, in this paper, we describe extensions of the FCLM to cover cases, in which there are missing data in part of a series of colocated images. Note that, while prediction of defoliation for missing data in nonforested pixels is not logical nor of interest, these "missing data" pixels must still be dealt with in a spatially explicit regression framework.

Manuscript received October 9, 2013; revised March 4, 2014 and June 6, 2014; accepted July 21, 2014. This work was supported by the National Aeronautics and Space Administration under Grant NNX06AD45G and Grant NNX08AN31G.

J. Zhang is with the Department of Financial and Institutional Research, Northern Illinois University, DeKalb, IL 60115 USA (e-mail: Zhang.Jun1@gmail.com).

M. K. Clayton is with the Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706 USA.

P. A. Townsend is with the Department of Forest and Wildlife Ecology, University of Wisconsin-Madison, Madison, WI 53706 USA.

Digital Object Identifier 10.1109/TGRS.2014.2345513

In a general sense, we might have missing pixels in the response image Y or in some of the covariate images X_i, or both. As in [12], we will focus on the single replicate case, in which the images $Y, X_1, X_2, \ldots,$ and X_p are the only images available.

Whenever there are missing pixels in the images, we can divide the pixels (locations) into two parts—an *incomplete part* and a *complete part*. In our image-based regression setup, the incomplete part includes all pixels that have missing or uninformative pixel values for at least one of the images under analysis, and the complete part includes all pixels that have valid pixel values for all of the images.

Depending on the goals, there are two main strategies for coping with missing data: the first, and most common, is to ignore regions where the data are incomplete and build the model over the complete part only. This is a sensible strategy when the regions with incomplete data are not of interest anyway. In a nonspatial regression problem, this amounts to simply deleting those observations which are incomplete.

The second strategy is to use some form of implicit or explicit imputation to fill in the missing pixels for the incomplete part of the data and then to fit the model as if there are no missing data [5]–[7]. This is particularly useful when interpolation in the missing regions of the response image Y is desired. However, an additional problem can arise: to use either of the aforementioned strategies, we first need to know the location of the missing or uninformative pixels. This may not be trivial. For example, in remotely sensed data, uninformative pixels caused by cloud cover, snow cover or sun glint usually require physical/optical models, segmentation software or manual identification. As we will see, one side-effect of our approach is that the FCLM appears to directly and appropriately handle unidentified cloud cover occurring in the covariate images.

In the following sections, we describe the two strategies for dealing with missing data and illustrate each with examples. In Section II, we use the same gypsy moth defoliation problem described in [12] to illustrate an approach for building an FCLM while ignoring missing data. In Section III, we use data from Landsat 7 imagery to discuss interpolation problems. In Section IV we make some concluding remarks. All computations in this paper were done with GNU Octave version 3.6.4 using a commodity PC with a 3 GHz Intel Core i7 CPU and 12 GB memory. We applied subsampling to accelerate the search of optimal penalizing term as described in [12]. Except where noted explicitly, we use Haar wavelet bases throughout [9]. In addition, an FCLM with partial wavelet bases was used in every numerical example in this paper. (Details about subsampling appear in the Appendix.)

II. Ignoring Missing Data

Our general approach is to handle missing data by creating appropriate covariates and then fitting an FCLM on the resulting data. To begin, we reconsider the gypsy moth data discussed in [12]. In this setting, missing and uninformative data are well-defined, and include such things as highways, parking lots, agricultural fields, and lakes. Defoliation rates in those areas are meaningless, and thus subregions corresponding to these objects should be excluded from the modeling process. This is equivalent to having missing pixels in some areas of the image for which there is no need or interest in interpolating the response image values Y for the missing regions.

For the moment, assume that there is only one covariate image X_1. To proceed, we construct a missing data "mask" matrix M with elements 1 or 0

$$M_{i,j} = \begin{cases} 1 & \text{if pixel value exists at row } i \text{ col } j \\ 0 & \text{if pixel value is missing at row } i \text{ col } j. \end{cases} \tag{3}$$

We then rewrite the FCLM as

$$Y \circ M = B_0 \circ M + B_1 \circ X_1 \circ M + E \tag{4}$$

and estimate $Y \circ M$ with

$$\hat{Y} \circ M = \hat{B}_0 \circ M + \hat{B}_1 \circ X_1 \circ M. \tag{5}$$

As in [12], we can expand B_0 and B_1 in (4) with a wavelet expansion

$$Y \circ M = \left\{ \sum_{j=1}^{H} v_j \phi_j \right\} \circ M + \left\{ \sum_{j=1}^{H} w_j \phi_j \right\} \{X_1 \circ M\} + E \tag{6}$$

and after rearranging (6) we get

$$Y \circ M = \sum_{j=1}^{H} v_j \{\phi_j \circ M\} + \sum_{j=1}^{H} w_j \{\phi_j \circ M \circ X_1\} + E. \tag{7}$$

With (7), we can use the same large scale l_1 constrained LSE to estimate v_j and w_j as was used in [12]. Since we are using l_1 constrained least squares, if there is no valid pixel value inside the support of ϕ_j, or if $\phi_j \circ M$ is equal to a zero matrix, then it is easy to see that the corresponding v_j and w_j will be zero. On the other hand, as long as there are some valid pixel values inside the support of ϕ_j, $\phi_j \circ M$ will not be a zero matrix and the values of v_j and w_j will be determined by those valid pixel values. The "new" wavelet base $\phi_j \circ M$ is quite flexible and can adapt well to different configurations of missing patterns and different missing proportions.

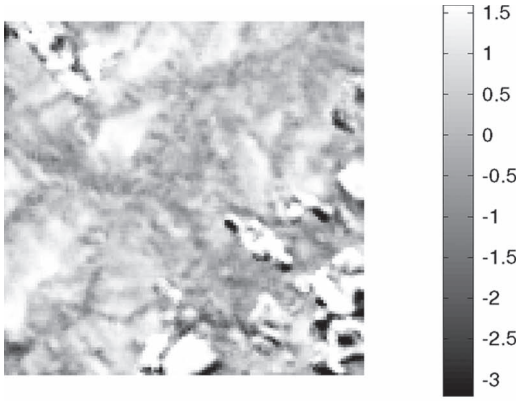
To illustrate this, in Fig. 1, we show 128 pixel $\times$ 128 pixel images for the defoliation rate, elevation and nonforest mask. In the original defoliation rate image Y there are some unusually bright areas of clouds in the northwest and southeast corners of the images which are not related to the gypsy moth defoliation study. We therefore mask these out as well.

In Fig. 2, we show the masked defoliation rate image $Y \circ M$, the masked elevation image $X_1 \circ M$, the estimated defoliation rate $\hat{Y} \circ M$, estimated constant surface $\hat{A} \circ M$, estimated slope surface $\hat{B} \circ M$ and residual surface from applying the aforementioned procedure.

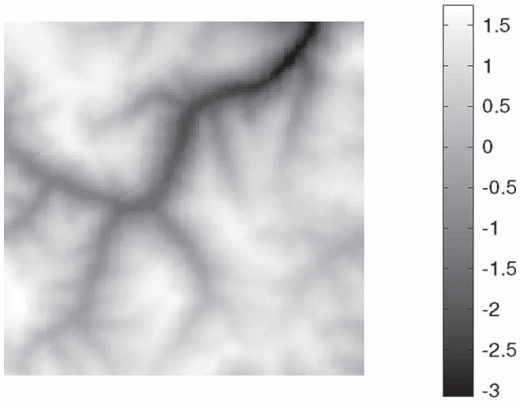
To assess the fit of our model, we use a form of R^2 [8]

$$R^2 = 1 - \frac{\|Y \circ M - \hat{Y} \circ M\|_F^2}{\|Y \circ M - \overline{Y \circ M}\|_F^2}$$

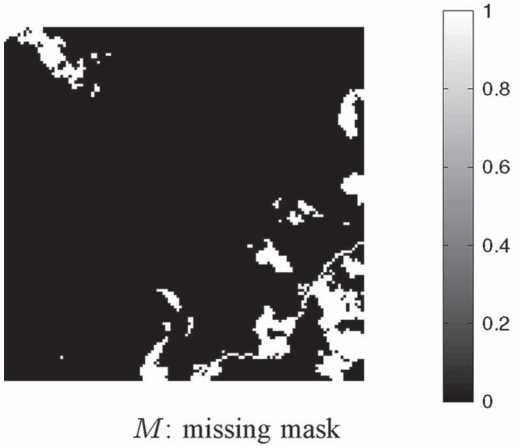
where $Y \circ M$ is the observed masked response image, $\hat{Y} \circ M$ is the estimated masked response image and $\overline{Y \circ M}$ is the mean image whose pixel values are equal to the mean of pixel values



Y: original defoliation rate image



X_1: original elevation image



M: missing mask

Fig. 1. Original Y defoliation rate image, original X_1 elevation image and missing data mask M for gypsy moth defoliation data. In the mask M, white areas stand for the missing subregions. In Fig. 2, we will show the masked defoliation rate image $Y \circ M$, the masked elevation image $X_1 \circ M$.

in $Y \circ M$. (In [12], the pointwise standard deviation of the estimated parameter surfaces was also used to diagnose the quality of fit. This measure can be of value particularly for cases with R^2 close to 1, a situation that does not arise in the examples in this paper.)

For this example, the smallest base support was 2×2, the largest wavelet base support was 128×128, and the overall R^2 was 0.586. Although we do not show the details here, we did investigate the use of an alternative wavelet base, namely Coif1 wavelets [9]. In that case the smallest base support we used

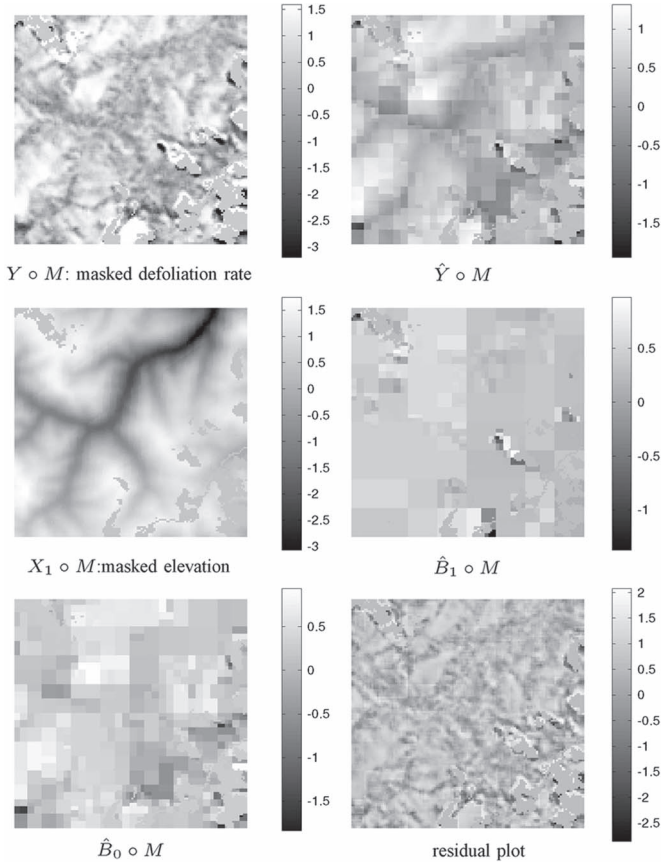


Fig. 2. Masking noninformative regions. Haar wavelets are used, with the smallest base support was 2×2, the largest wavelet base support was 128×128. The model is $Y \circ M = B_0 \circ M + B_1 \circ X_1 \circ M + E$ where Elevation is the explanatory variable and Defoliation rate is the response. Note that the area with missing data is masked off in the elevation and defoliation rate images. The overall R^2 was 0.586.

was 6×6 and the largest wavelet base support was 76×76. The overall R^2 was 0.688 and the resulting $\hat{Y} \circ M$ image (not shown) was smoother, but at a cost of computational time—implementing the Coif1 wavelets had a computational time approximately seven-fold more than using the Haar wavelet basis.

The aforementioned work illustrates how an FCLM can be constructed to utilize the complete part of the data and ignore the incomplete part, at least, if we are only interested in using one mask M for all of the images involved in the modeling process. A more challenging situation arises when the response image and covariate images each have different missing regions which may or may not overlap. We address this in the next section, and also demonstrate an approach for image imputation based on covariate images.

III. Filling in Gaps: Landsat 7 Images

A. Background

Landsat 7 was launched on April 15, 1999 and was equipped with the enhanced thematic mapper (ETM+) multispectral instrument. However, on May 31, 2003 the scan line corrector (SLC) in the ETM+instrument failed. The function of the SLC is to compensate for the forward motion of the satellite, and

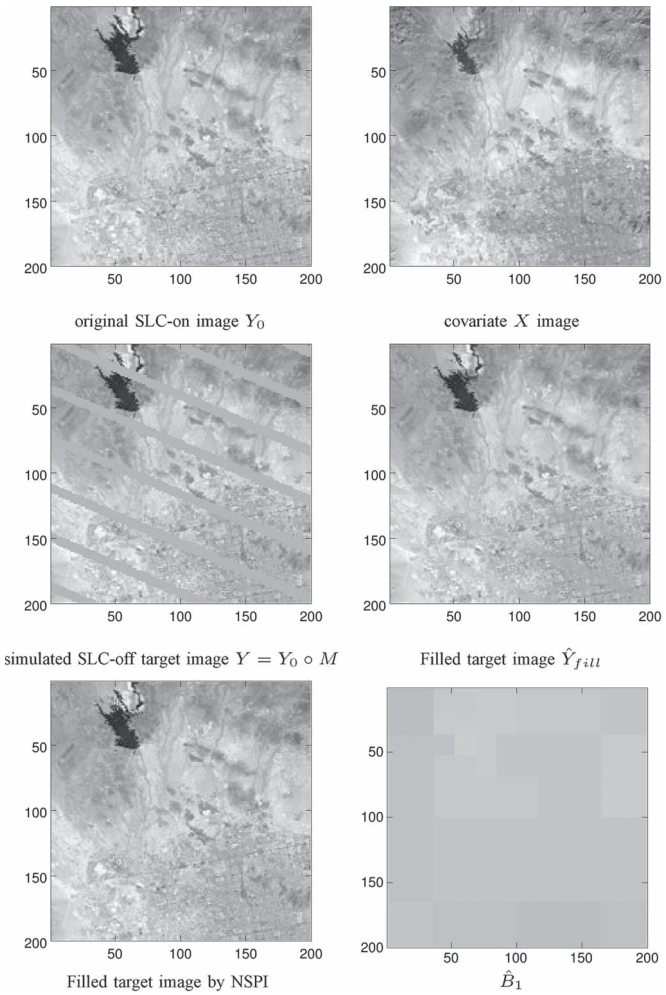


Fig. 3. Filling in the gaps. Missing Stripes in $\hat{Y}_{fill}$ are from $\hat{Y}_0 = \hat{B}_0 + \hat{B}_1 \circ X_1$ and nonmissing parts are from simulated SLC off target image Y. $\hat{Y}_{fill} = \hat{Y}_0 \circ \widetilde{M} + Y_0 \circ M$. For our model, the partial R_p^2 was 0.7145 and RMSE was 16.42.

the failure of the SLC creates stripes of missing data in images acquired by Landsat 7 (see Fig. 3). Such images are referred to as "SLC-off" images. Although Landsat 8 is now online, for many long term monitoring projects, the historical images from Landsat 7 are required, and our approach has significant utility for data-filling of archival Landsat 7 images prior to the Landsat 8 becoming operational (i.e., 2003–2013). Another important reason for the need to fill gaps in Landsat 7 images is that the Landsat 5 stopped working before the launch of Landsat 8, and therefore there is a data gap in the Landsat continuous archive. As a result, there is substantial interest in filling in the missing stripes in Landsat 7 images. Currently, NASA uses several methods to correct the SLC-off mode Landsat 7 images, including an early approach called a "localized linear histogram matching algorithm" which incorporates an SLC-on image into an SLC-off image, and a later version called an "Adaptive Local Linear Histogram Adjustment (ALLHA)" which utilizes multiple SLC-off and SLC-on images to fill in the missing stripes in an SLC-off image [10], [11]. In recent papers, authors present different approaches for filling in gaps of SLC-off images. In [2], a neighborhood similar pixel interpolator (NSPI) is developed to utilize SLC-on and SLC-off input images to fill the gaps in target SLC-off images. A geostatistical approach for filling gaps is introduced in [3]. Most recently, a multi-temporal regression and regularization method is developed in [4]. Since the NSPI in [2] is a simple and effective method used as a basis for comparison by many authors, we compare our results with the results obtained by the NSPI later in this paper.

Landsat 7 follows a near polar Sun-synchronous orbit. This means that the satellite passes over any given point approximately at the same local time. The plane of the orbit fully rotates around the axis of the Earth in a year. In the case of Landsat 7, this plane comes back to the same acquisition location approximately at the same local time every 16 days. Thus the acquisition dates for the same location are always 16 days apart. At any surface location, the incidence angle with respect to the satellite will be nearly the same at each pass, although the illumination angle from the sun will vary with the Earth's axial tilt. (Note that Landsat 5 operated in tandem with Landsat 7 through January 2013, but offset by 8 days. Although Landsat 5 data are complete, that satellite was launched in 1984, and the data quality is considerably degraded with respect to Landsat 7. Thus use of Landsat 5 data to fill Landsat 7 gaps for 2003–2012 is suboptimal.)

Our general goal is to use an FCLM to interpolate the missing data in an SLC-off image—to fill in the gaps in images resulting from the SLC failure. For a given image that we want to "repair," our general strategy is to use other images as covariates. Ideally, a covariate image will be as close as possible, in some sense, to the image to be repaired. We illustrate this in the following. We use real satellite images in all examples but we create all SLC-off images by masking off stripes from intact, SLC-on images. This makes it easy to check our results against the original SLC-on images. Landsat images are 185 km $\times$ 185 km with SLC-off missing stripes ranging from one to 15 pixels. In our simulated missing data the typical width of missing stripes is about 7 to 9 pixels. For these simulations, the image sizes are all 200 $\times$ 200 unless otherwise mentioned. Images in our examples are downloaded from the U.S. Geological Survey Earth Explorer (http://earthexplorer.usgs.gov). They correspond to regions around Phoenix, AZ and were captured on different dates from 2000 to 2001, noted more specifically in what follows.

B. Using FCLMs to Fill in Missing Stripes

1) Example 1: Using an SLC-on Image to Fill in an SLC-off Image: We first use a related SLC-on image to fill in the missing stripes in a given target SLC-off image. In Fig. 3, the SLC-off image Y with missing stripes is the image we want to repair. This target image Y was generated from a complete SLC-on image Y_0 acquired on May 21, 2000, with

$$Y = Y_0 \circ M$$

where M is the corresponding missing data matrix defined in (3). The covariate image that we use covers the same area as Y does, but was acquired on November 29, 2000.

We can write our model as

$$Y_0 \circ M = Y = B_0 \circ M + B_1 \circ X_1 \circ M \tag{8}$$

where M is the missing data mask matrix defined in (3). After we have calculated $\hat{B}_0$ and $\hat{B}_1$, and if we assume the concurrent linear relationship is the same in the missing regions, then the estimated *complete* target image $\hat{Y}_0$ is

$$\hat{Y}_0 = \hat{B}_0 + \hat{B}_1 \circ X_1. \tag{9}$$

By combining the estimated stripes in $\hat{Y}_0$ with the nonmissing parts in the target image Y, the final filled image $\hat{Y}_{fill}$ can be expressed as

$$\hat{Y}_{fill} = \hat{Y}_0 \circ \widetilde{M} + Y_0 \circ M \tag{10}$$

where $\widetilde{M}$ is equal to $J - M$ and J is a matrix whose elements are all equal to 1. To evaluate the filled stripes, we define the partial R_p^2 for the fitted pixel values in missing regions as

$$R_p^2 = 1 - \frac{\|Y_0 \circ \widetilde{M} - \hat{Y}_0 \circ \widetilde{M}\|_F^2}{\|Y_0 \circ \widetilde{M} - \overline{Y \circ \widetilde{M}}\|_F^2}$$

and root-mean-squared error (RMSE) as

$$\text{RMSE} = \frac{\|Y_0 \circ \widetilde{M} - \hat{Y}_0 \circ \widetilde{M}\|_F}{\sqrt{n}}$$

where is n is the number of missing pixels in all gaps of the target image.

For our model, the partial R_p^2 was 0.7145 and RMSE was 16.42 and for the Adaptive Local Linear Histogram Adjustment (ALLHA), the partial R_p^2 was 0.7287 and RMSE was 16.01. Since we used partial wavelet bases and an accelerated searching algorithm for optimal penalization terms, the computational time for the FCLM was 9.75 s. For the Neighborhood Similar Pixel Interpolator (NSPI), the partial R_p^2 was 0.6695 and RMSE was 17.67. The computational time for the NSPI was 9.59 s.

Although the seasonal effects (lake size) and possibly illumination conditions between the response image and the covariate image are quite strong, for these data, the FCLM seems to successfully interpolate the missing subregions between the two nonmissing subregions. The relatively high R_p^2s for the FCLM and the NSPI indicate the fit is good for both models. Although we do not show specific results here, we have observed in general that the interpolated values are comparable to the results from the adaptive linear histogram adjustment algorithm. By observation, we have a good $\hat{Y}_{fill}$ and do not see any large artifacts in $\hat{Y}_{fill}$ from Fig. 3. The quality of response image is important since our loss function is quadratic and sensitive to outliers. However, if there are clouds in the target SLC-off image Y, then the filled $\hat{Y}_{fill}$ will be greatly impacted.

2) Examples 2 and 3: Using Multiple SLC-Off Covariate Images to Fill Target SLC-Off Image: Using SLC-on images to fill the missing stripes in a target SLC-off image was discussed in [2]. However, SLC-on images were only acquired before May 31, 2003 and some scenes covered by SLC-on images may have changed substantially over time. This raises the priority of using SLC-off images or a combination of SLC-off images and SLC-on images to fill the stripes in a target SLC-off image.

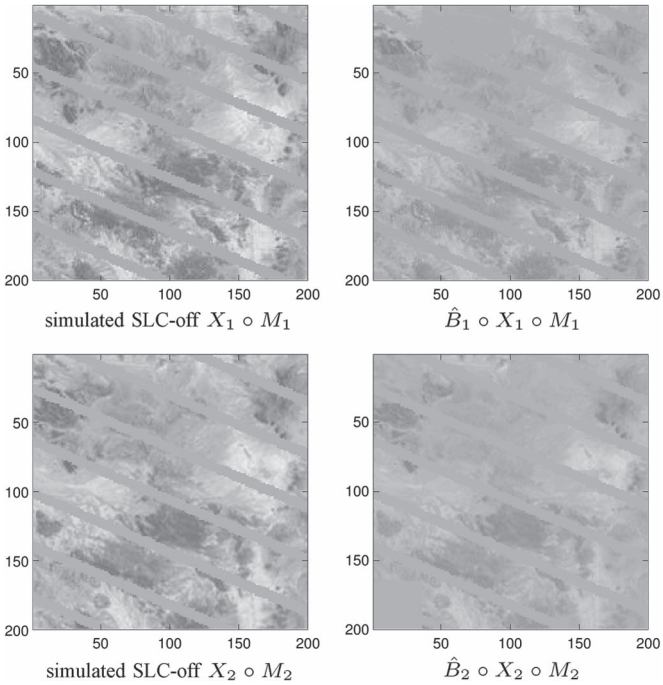


Fig. 4. Filling in the gaps with 2 SLC-off images and this figure demonstrates how the modeling approach leads to automatic pixel selection.

Although this requires dealing with extra missing stripes in the covariate images and potentially even more incomplete data, these concerns may be balanced by the advantage of having more current SLC-off images available. We next focus on using SLC-off images to fill in missing gaps.

To use SLC-off images as covariates, we face the situation where the response image and covariate images have different but identified missing regions. To handle such a situation, we introduce a missing data mask matrix for each SLC-off image in additional to the initial missing data mask M for the response image. We illustrate this in our next two examples, where we will use two SLC-off images to fill in the missing stripes in the target SLC-off image. The corresponding model is

$$Y_0 \circ M = Y = B_0 \circ M + B_1 \circ M \circ X_1 \circ M_1 + B_2 \circ M \circ X_2 \circ M_2 \tag{11}$$

where M, M_1 and M_2 are the missing data mask matrices for the response and covariate images, respectively. Here, we assume that M, M_1, and M_2 are not identical. After this model has been fitted, the estimated *complete* target image $\hat{Y}_0$ will be

$$\hat{Y}_0 = \hat{B}_0 + \hat{B}_1 \circ X_1 \circ M_1 + \hat{B}_2 \circ X_2 \circ M_2 \tag{12}$$

and (10) can be used to compute the filled image $\hat{Y}_{fill}$. We require that $\widetilde{M}_1 \circ \widetilde{M} = \mathbf{0}$ and $\widetilde{M}_2 \circ \widetilde{M} = \mathbf{0}$ where $\mathbf{0}$ is a zero matrix. In subsequent examples, we will impose similar conditions on the mask matrices. These conditions are the only strict requirements we need for the corresponding images, although we will discuss image choice a bit more in the conclusions.

In Figs. 4 and 5, we show two simulated SLC-off images. One covariate, $X_1 \circ M_1$, was acquired on May 5, 2000 and the other covariate, $X_2 \circ M_2$, was acquired on June 6, 2000. The acquisition date for the (simulated) SLC-off target image $Y \circ M$ was May 21, 2000. The missing mask in $X_1 \circ M_1$ was

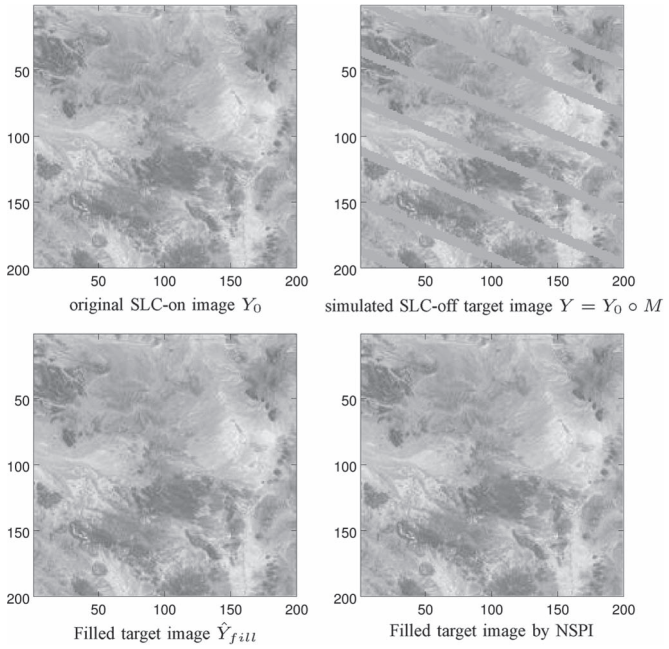


Fig. 5. Missing stripes in $\hat{Y}_{fill}$ are from $\hat{Y}_0 = \hat{B}_0 + \hat{B}_1 \circ X_1 \circ M_1 + \hat{B}_2 \circ X_2 \circ M_2$ and the nonmissing parts are from the simulated SLC off target image Y. $\hat{Y}_{fill} = \hat{Y}_0 \circ \widetilde{M} + Y_0 \circ M$ and the partial R_p^2 was 0.9575 and RMSE was 5.51.

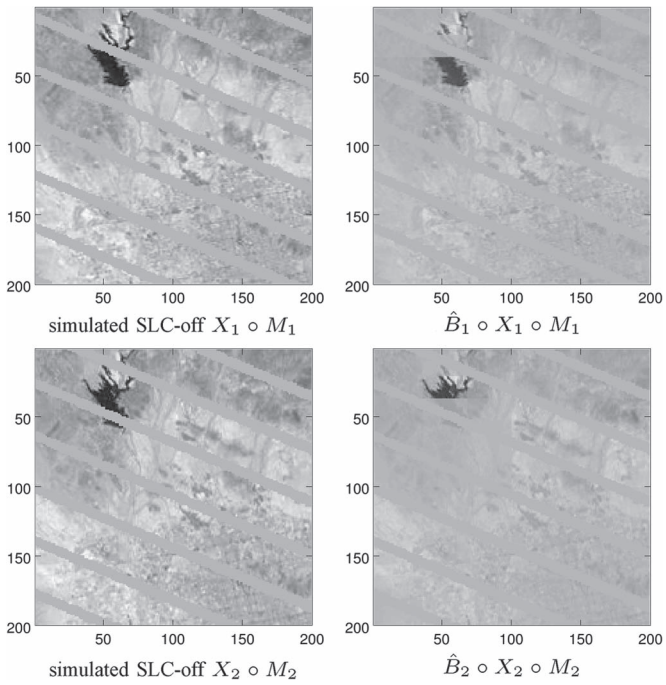


Fig. 6. Automatic, adaptively choosing the subregions of $X_1 \circ M_1$ and $X_2 \circ M_2$ to explain different parts of the response image.

generated by shifting the gaps in $Y_0 \circ M$ down 10 pixels. The missing mask in $X_2 \circ M_2$ was generated by shifting the gaps in $Y_0 \circ M$ up 10 pixels. The gaps in $Y_0 \circ M$ are captured from a real SLC-off image. The acquisition dates were chosen to be very close with the expectation that the temporal variations of the surface will be small. For our model, the partial R_p^2 was 0.9575 and RMSE was 5.51, indicating a good fit. The computational time for the FCLM was 10.97 s. We applied the Neighborhood Similar Pixel Interpolator (NSPI) on the same data set and for the NSPI R_p^2 was 0.9438 and RMSE was 6.34. The computational time for the NSPI was 7.36 s.

In Fig. 5, we can see that the original target image Y_0 is quite homogeneous. In the next example, we apply FCLM on a heterogeneous site to demonstrate that FCLM is able to handle heterogeneous images. We use two simulated SLC-off covariate images, which were also acquired separately on May 5, 2000, and June 6, 2000. The target SLC-off image was acquired on May 21, 2000. Each image has the same (simulated) missing stripes as those in Example 1. All covariate images and the target image are shown in Figs. 6 and 7. From Fig. 7, we can see that the site contains a lake, a town and mountains and is quite heterogeneous. For our model, the partial R_p^2 was 0.9347 and RMSE was 7.34. The computational time was 12.03 s. The results are shown in Figs. 6 and 7. We also applied the NSPI on this data set and for the NSPI R_p^2 was 0.9194 and RMSE was 8.16. The computational time for the NSPI was 7.60 s.

For either homogeneous sites or heterogeneous sites, it is interesting to see that our algorithm adaptively chose subregions of $X_1 \circ M_1$ and $X_2 \circ M_2$ to explain different parts of the response image. We note that the model appears to automatically select pixels across the covariate images with different missing subregions. Although our approach was not designed to have this feature, its existence is not surprising.

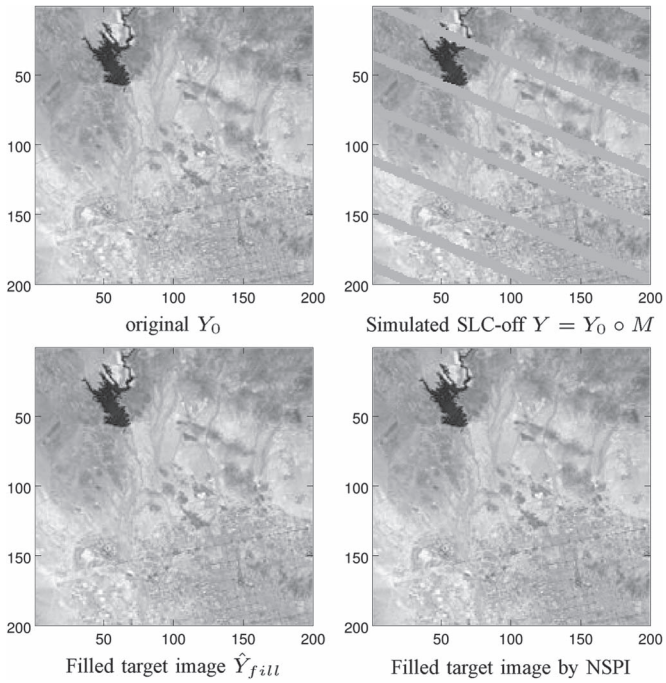


Fig. 7. Filling in the gaps with 2 SLC-off images. Missing Stripes in $\hat{Y}_{fill}$ are from $\hat{Y}_0 = \hat{B}_0 + \hat{B}_1 \circ X_1 \circ M_1 + \hat{B}_2 \circ X_2 \circ M_2$ and nonmissing parts are from simulated SLC off target image Y. $\hat{Y}_{fill} = \hat{Y}_0 \circ \widetilde{M} + Y_0 \circ M$. The partial R_p^2 was 0.9347 and RMSE was 7.34.

The use of LASSO means we are trying to efficiently use the information at hand, whereas the use of wavelets allows the modeling to select subregions spatially within the covariate images to provide such information.

3) Example 4: Using SLC-Off Images and a Cloudy SLC-on Image to Fill Target SLC-Off Image: In examples 2 and 3, although missing some stripes due to SLC problems, the

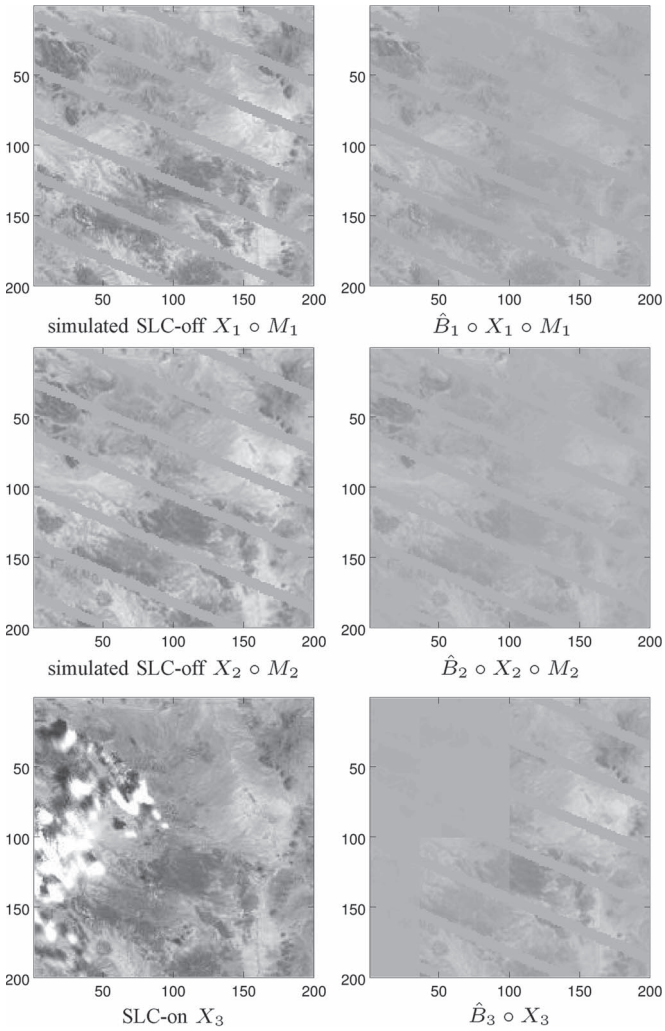


Fig. 8. Filling in the gaps with 2 SLC-off image and one cloudy SLC-on image. It is quite interesting to see that our model automatically selects X_3's cloud-free right half to fill in the stripes although the cloud cover in X_3 is not marked.

covariate images are otherwise of high quality. As we have mentioned before, the estimation result is sensitive to the quality of the response image due to the quadratic loss function in our model. What happens if we have unidentified random missing regions such as cloud covers in the covariate images? In our next example, we will use two simulated SLC-off images and one SLC-on image to fill in the stripes in the target SLC-off image. Central to this example is a new SLC-on X_3 in Fig. 8 which has some clouds in the left half. The acquisition date for the SLC-on image was September 26, 2000, and the two SLC-off images $X_1 \circ M_1$ and $X_2 \circ M_2$ in Fig. 8 are the same as those in Fig. 4 acquired separately on May 5, 2000 and June 6, 2000. The SLC-off target image in Fig. 9 is the same as the target image in Fig. 5 (with acquisition date May 21, 2000). Note in this example, the region of interest is from the same region as in Example 2.

For our model, the partial R_p^2 was 0.9550 and RMSE was 5.67. For the NSPI, the partial R_p^2 was 0.9438 and RMSE was 6.34. The results for the NSPI were the same as those in Example 2. This is because the NSPI uses covariate images according to their acquisition dates and the closest is given first

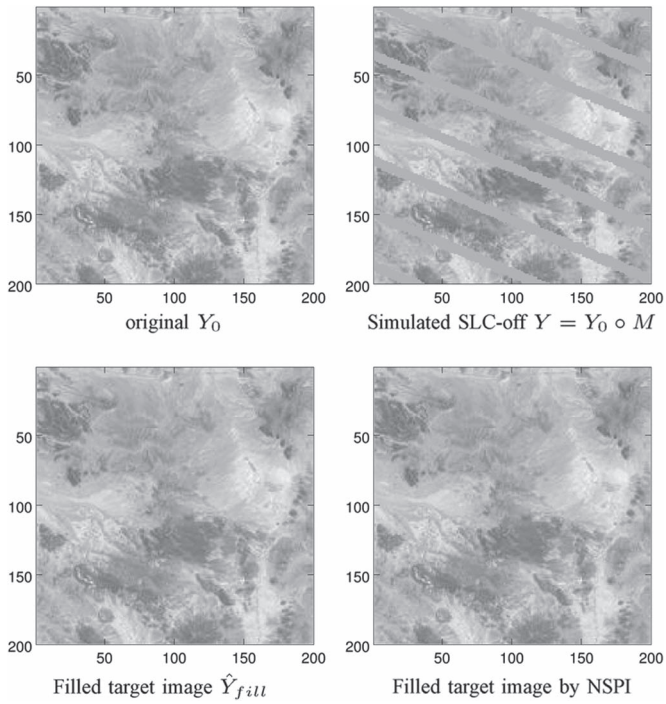


Fig. 9. Filling in the gaps with 2 SLC-off image and one cloudy SLC-on image. Missing Stripes in $\hat{Y}_{fill}$ are from $\hat{Y}_0 = \hat{B}_0 + \hat{B}_1 \circ X_1 \circ M_1 + \hat{B}_2 \circ X_2 \circ M_2 + \hat{B}_3 \circ X_3$ and nonmissing parts are from simulated SLC off target image Y. $\hat{Y}_{fill} = \hat{Y}_0 \circ \widetilde{M} + Y_0 \circ M$. The partial R_p^2 was 0.9550 and RMSE was 5.67.

priority. $X_1 \circ M_1$ and $X_2 \circ M_2$ are free of cloud and either $X_1 \circ M_1$ or $X_2 \circ M_2$ completely covers the gaps in $Y_0 \circ M$. If the acquisition date of X_3 had been the closest and given first priority, then the cloud cover in X_3 would have impacted the performance of the NSPI significantly. The results of the FCLM are shown in Figs. 8 and 9. It is interesting to see that our model automatically selects the cloud-free right half of X_3 to fill the stripes although the cloud cover in X_3 is not marked. This example shows that even with very different missing mechanisms and patterns in different covariate images, our algorithm adapts well to the situation and automatically selects the best useful region in each covariate to fill in the stripes in the target image.

IV. Conclusion

We have shown that an FCLM is capable of handling missing data, and in particular, filling in the stripes for SLC-off Landsat 7 images. In comparisons with another established method, NSPI, we find that the two generally perform similarly; for the examples seen here, the FCLM is a little more accurate, but the NSPI is somewhat faster in terms of computational time. For straightforward applications, we think that either method can be reasonably applied. However, we think the FCLM has some attributes that are of particular value when images have uncertain quality. We have demonstrated that an FCLM can interpolate unobserved subregions for the response image and that use of an FCLM can be robust when the quality of each covariate image is in question. These two attributes are particularly valuable in real world applications. First, interpolating unobserved regions in spatial images is very important since acquisition conditions cannot be controlled. When filling in the

unobserved regions, an appealing feature of an FCLM is that it utilizes spatial information at multiple resolution levels via wavelet bases. It is commonly known that images may contain different information at different resolution levels, but it is not easy to capture the information at multiple resolutions with simple schemes such as moving windows. An FCLM implemented with wavelet bases provides a systemic way to interpolate at multiple resolutions. The NSPI does not include multiple resolution information although it is also a local method.

Second, adaptation to the quality of each covariate image is actually a variable selection function based on the quality of pixel values. In some cases, like SLC-off images, the missing pixels can be clearly marked in a missing mask matrix. However, even if we have the locations of the missing pixels, different missing patterns in different images still complicate pixel selection for interpolation. More difficult situations arise when missing pixels are unmarked. For example, clouds represent missing subregions that cannot always be easily marked. Putting this together, we may have marked missing pixels, unmarked missing pixels, low-quality pixels, and high-quality pixels all mingled together in a single satellite image. A useful regression model for spatial images should be able to select the best possible combination of pixel values across all covariate images to predict the pixel values in the response image. We note that when the nonmissing part of each covariate image completely covers the gaps in the target image, the NSPI actually uses only one of the covariate images to fill in the gaps and the FCLM uses different parts from different covariate images to produce an improved filled image. The FCLM is thus able to more flexibly use information from covariate images to best perform imputations.

Our formal assumptions regarding covariate images focus on where the missing data are located. These formal conditions are illustrated, for example, following (12). Apart from these requirements, if possible, it is preferable to choose an image that best matches the target image. This might be an image from an adjacent date (to eliminate major land-cover changes), or an image from a similar date but a different year (to eliminate illumination or seasonal differences). It is not a requirement that the covariate comes from a different year; at the same time, if another year's image is particularly useful, it could be valuable to include it.

When we use multiple covariate images, we can relax even these guidelines somewhat, because we have found that the method tends to select useful sections of covariate images and disregard sections that are not useful. Overall, then, it makes sense to try to choose images that have useful information. The choice may not be too critical because the method will compensate if necessary, although perhaps at the expense of having to add additional images, and thus requiring additional computational time.

As we have seen, an FCLM implemented with a variable selection procedure can provide robust interpolation and pixel selection capabilities despite being a simple looking model. We expect that there are circumstances in which the alignment of clouds, missing data or other anomalies are such that a good prediction cannot be made. This situation could be remedied by including a larger selection of covariate images, at the cost of greater computation time. In this paper, we manually selected several covariate images from a set of images to fill the gaps in the response image. If there are many images available and we choose a subset of these, then the total number of sets of possible choices is potentially very large. In future work, we will address the automatic selection of covariate images. If these can be chosen in a near-optimal fashion, this may further improve the results and allow for a more efficient use of available satellite images.

APPENDIX A
ACCELERATING THE SEARCH OF OPTIMAL PENALIZATION TERM USING SUBSAMPLING

From [12], we know that the computational cost for an FCLM is about proportional to $O(n^{1.3})$. For a relatively big image, like a 512×512 image, this will require several hours to finish the computation. Most of computational time is spent on finding the optimal penalty term, because, for example, we need to go through a grid of lambda values sequentially to compute the BIC score for each λ value and find the λ_{opt}.

To improve the computational performance, we discuss here a subsampling method to accelerate the search for an optimal λ. The key idea is that we will randomly sample several sets of subimages from the original response and covariate images and use these subimages to estimate $\hat{\lambda}_{\text{opt}}$. Suppose we randomly sample n sets of subimages (y_i, x_i) and we obtain the optimal $\lambda_{\text{opt},i}$ for each set of subimages with

$$\hat{\lambda}_{\text{opt},i} = \arg\min_{\lambda_i} BIC(\lambda_i, y_i, x_i). \tag{13}$$

The final estimated $\hat{\lambda}_{\text{opt}}$ can be obtained as the mean of $\hat{\lambda}_{\text{opt},i}$'s

$$\hat{\lambda}_{\text{opt}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\lambda}_{\text{opt},i}. \tag{14}$$

Subsampling can substantially accelerate the search for the optimal penalization term. Based on empirical work, the $\hat{\lambda}_{\text{opt}}$ obtained with subsampling is consistent with $\hat{\lambda}_{\text{opt}}$ obtained with complete images. Of course the conditions of the images such as signal to noise ratio, size and total number of subimages will affect the estimation accuracy.

ACKNOWLEDGMENT

The authors would like to thank J. Foster, C. Kingdon, and A. Singh for assistance with data preparation. The comments by two anonymous reviews also greatly improved the manuscript. The authors also thank Dr. J. E. Vogelmann and Dr. D. Liu for sharing their IDL source code with us.

REFERENCES

[1] P. A. Townsend *et al.*, "A general Landsat model to predict canopy defoliation in broadleaf deciduous forests," *Remote Sens. Environ.*, vol. 119, pp. 255–265, Apr. 2012.

[2] J. Chen, X. Zhu, J. E. Vogelmann, F. Gao, and S. M. Jin, "A simple and effective method for filling gaps in Landsat ETM plus SLC-off images," *Remote Sens. Environ.*, vol. 115, no. 4, pp. 1053–1064, Jan. 2011.

[3] X. Zhu, D. Liu, and J. Chen, "A new geostatiscial approach for filling gaps in Landsat ETM+ SLC-off images," *Remote Sens. Environ.*, vol. 124, pp. 49–60, Sep. 2012.
[4] C. Zeng, H. Shen, and L. Zhang, "Recovering missing pixels for Landsat ETM+ SLC-ff imagery using multi-temporal regression analysis and a regularization method," *Remote Sens. Environ.*, vol. 131, pp. 182–194, Apr. 2013.
[5] B. Marcelo, S. Guillermo, C. Vincent, and B. Coloma, "Image inpainting," in *Proc. SIGGRAPH Proc. 27th Annu. Conf. Comput. Graph. Interactive Tech.*, 2000, pp. 417–424.
[6] A. Levin, A. Zomet, and Y. Weiss, "Learning how to inpaint from global image statistics," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, vol. 1, pp. 305–312.
[7] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.
[8] J. Ramsay and B. Silverman, *Functional Data Analysis*, 2nd ed. New York, NY, USA: Springer-Verlag, 2005.
[9] J. S. Walker, *A Primer on Wavelets and Their Scientific Applications*. Boca Raton, FL, USA: CRC Press, 1999.
[10] P. Scaramuzza, E. Micijevic, and G. Chander, *SLC Gap-Filled Products Phase One Methodology*, U.S. Geological Survey Earth Resources Observation and Science (EROS) Center, Mar. 2004. [Online]. Available: http://landsat.usgs.gov/documents/SLC_Gap_Fill_Methodology.pdf
[11] P. Scaramuzza, E. Micijevic, and G. Chander, *SLC-off Gap-Filled Products Gap-Fill Algorithm Methodology Phase 2 Gap-Fill Algorithm*, U.S. Geological Survey Earth Resources Observation and Science (EROS) Center, Oct. 2004. [Online]. Available: http://landsat.usgs.gov/documents/L7SLCGapFilledMethod.pdf
[12] J. Zhang, M. K. Clayton, and P. A. Townsend, "Functional concurrent linear regression model for spatial images," *J. Agricultural, Biol., Environ. Stat.*, vol. 16, no. 1, pp. 105–130, Mar. 2011.

Jun Zhang received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 1993, the M.E. degree from Southeast University, Nanjing, China, in 1996, and the M.S. degree from Iowa State University, Ames, IA, USA, in 1999. In December 2008, he received the Ph.D. degree in statistics from University of Wisconsin-Madison, Madison, WI, USA.

His research interests include spatial statistics and statistical methods for high-dimensional data.

Murray K. Clayton received the B.Math. degree from the University of Waterloo, Waterloo, ON, Canada, in 1979 and the Ph.D. degree in statistics from the University of Minnesota, Minneapolis, MN, USA, in 1983.

Since 1984, he has been on the faculty with the University of Wisconsin-Madison, Madison, WI, USA. His research interests include applications of statistics to the agricultural, environmental, and biological sciences, with a focus on statistical methods for spatial data.

Philip A. Townsend received the Ph.D. degree in geography from the University of North Carolina, Chapel Hill, NC, USA, in 1997.

He is currently Professor of forest ecology with the University of Wisconsin-Madison, Madison, WI, USA. He specializes in optical remote sensing, imaging spectroscopy ecosystem ecology, and the development of quantitative methods to link biochemical, biophysical, and remote sensing data. His research interests focus on applications of spectroscopy to the measurement of the physiology of vegetation, characterizing ecosystem dynamics in response to environmental change, and quantifying the effects of disturbance and insects on nutrient cycling and water quality.