A simplified data assimilation method for reconstructing time-series MODIS NDVI data

Juan Gu, Xin Li, Chunlin Huang, Gregory S. Okin

Cold and Arid Region Environmental and Engineering Research Institute, CAS, 322 Rd. Donggang West, Lanzhou, Gansu 730000, China

UCLA Department of Geography, Los Angeles, California, 90095, USA

Received 3 December 2007; received in revised form 4 May 2009; accepted 10 May 2009

Abstract

The Normalized Difference Vegetation Index (NDVI) is an important vegetation index, widely applied in research on global environmental and climatic change. However, noise induced by cloud contamination and atmospheric variability impedes the analysis and application of NDVI data. In this work, a simplified data assimilation method is proposed to reconstruct high-quality time-series MODIS NDVI data. We extracted 16-Day L3 Global 1 km SIN Grid NDVI data sets for western China from MODIS vegetation index (VI) products (MOD13A2) for the period 2003–2006. NDVI data in the first three years (2003–2005) were used to generate the background field of NDVI based on a simple three-point smoothing technique, which captures annual features of vegetation change. NDVI data for 2006 were used to test our method. For every time step, the quality assurance (QA) flags of the MODIS VI products were adopted to empirically determine the weight between the background field and NDVI observations. Ultimately, more reliable NDVI data can be produced. The results indicate that the newly developed method is robust and effective in reconstructing high-quality MODIS NDVI time-series.

Keywords: Data assimilation; Reconstruction; MODIS NDVI; Time-series data

1. Introduction

TheNormalized Difference Vegetation Index (NDVI) is widely used vegetation index due to its simplicity, ease of application, and wide-spread familiarity. Time-series NDVI products, derived from multiple satellite sensors such as the NOAA/AVHRR (Advanced Very High Resolution Radiometer), MODIS (Moderate Resolution Imaging Spectroradiometer), Landsat TM (Thematic Mapper), ETM+ (Enhanced Thematic Mapper Plus), and SPOT/VEGETATION, have proven powerful tools for learning from past events, monitoring current natural-resource conditions and long-term land-use/cover changes, and extracting canopy biophysical parameters and phenological information. Forecasting the trends of terrestrial ecosystems on global, continental, and regional scales (Tucker, 1979; Sellers et al., 1994; Field et al., 1998; Van Leeuwen et al., 1999; Fang et al., 2001) is also performed with these tools. However, the current NDVI product is still spatio-temporally discontinuous due to cloud cover, seasonal snow, atmospheric variability, bi-directional effects and instrument problems (Gutman, 1991; Huete and Liu, 1994; Xiao et al., 2003; Moody et al., 2005). These biases limit the application of NDVI in vegetation dynamics monitoring and global change research.

The most common compositing criterion used to produce NDVI composite data is the Maximum Value Composite (MVC) algorithm, which is applied to obtain a higher percentage of clear-sky data. Nevertheless, significant residual effects remain (Holben, 1986). Moreover, in recent years a number of mathematical filters have been applied to reduce noise and to reconstruct high-quality time-series of NDVI data for further analysis and application. These filters are generally grouped into two types:
(1) noise reduction in the frequency domain such as Fourier-based fitting methods (Sellers et al., 1994; Verhoef et al., 1996; Roerink et al., 2000);

(2) noise reduction in the temporal domain with approaches such as the best index slope extraction algorithm (BISE) (Viovy et al., 1992); weighed least-squares linear regression (Swets et al., 1999); modified BISE filtering (Lovell and Graetz, 2001); asymmetric Gaussian function-fitting approach (Jonsson and Eklundh, 2002); polynomial least squares operation (PoLeS) approach (Jose et al., 2002); Savitzky-Golay filtering (Chen et al., 2004); and the mean-value iteration filter (Ma and Veroustraete, 2006).

These methods are commonly used to restore NDVI multi-temporal profiles, but they suffer from several drawbacks that limit their use (Jonsson and Eklundh, 2002; Chen et al., 2004). For instance, the BISE algorithm requires the definition of a sliding period and a threshold for acceptable percentage increase in NDVI for re-growth during a sliding period based on an empirical strategy. That strategy is usually subjective and depends on the skills and experience of the analyst (Viovy et al., 1992; Lovell and Graetz, 2001). The remaining noise after application of the BISE algorithm may make the extracted temporal information unreliable (Ma and Veroustraete, 2006). Fourier-based fitting methods may generate spurious oscillations when applied to asymmetric NDVI time-series since they depend critically on symmetric sine and cosine functions (Roerink et al., 2000; Chen et al., 2004). The asymmetric Gaussian function-fitting approach is more flexible and effective in obtaining high-quality NDVI time-series, but it fails to identify a reasonable and consistent set of maxima and minima to which the local functions can be fitted, especially for noisy data or for data from areas where there is no clear seasonality (Jonsson and Eklundh, 2002). With respect to the Savitzky-Golay filter approach, it can reduce the noise of NDVI time-series effectively, but it also requires empirical analysis to determine the width of the smoothing window and the degree of the smoothing polynomial (Chen et al., 2004). Many rank order-based filters, such as the mean-value iteration filter method (Ma and Veroustraete, 2006), are efficient at removing additive impulsive noise, while linear filters succeed in suppressing Gaussian noise. However, most suffer from trade-offs, e.g., removing noise yet preserving details of the NDVI temporal dynamics.

In fact, smoothing filter design in signal processing basically depends on different types of noise. Moreover, as the type and the amount of noise are mixed differently, noise removal continues to provide a challenge for smoothing filter designers (He, 2007). Due to this issue, more information is required to improve the quality of multi-temporal NDVI data. This process is data assimilation, which has been widely used to initialize numerical weather prediction models (Daley, 1991) and to improve estimations of soil moisture or soil temperature profiles in the vertical direction by assimilating remote sensing data or in situ observations of recent years (Li et al., 2004; Huang et al., 2008a, b). During the last 2 years, data assimilation methods have been used to dynamically reconstruct remote sensing data. Gu et al. (2006) adopted the optimal interpolation method to generate multi-temporal MODIS LAI data. He (2007) performed an experiment based on the gradient inverse weighted filter and performed an objective analysis to improve the estimation of multi-temporal series of MODIS LAI data products. However, until now, only limited research has focused on multi-temporal NDVI time-series.

In this paper, a simplified data assimilation method is developed using NDVI quality assurance (QA) data sets to reconstruct multi-temporal time-series of MODIS NDVI products. The method was tested and evaluated by MODIS 16-Day L3 Global 1 km SIN Grid VI data sets (MOD13A2). The remainder of this paper is organized as follows. In Section 2 the data analysis strategy for reconstructing MODIS NDVI is introduced. The results are analyzed in Section 3, and conclusions are presented in the last section.

2. Data analysis strategy

The data assimilation method is used to integrate all information available to improve state estimation, which is widely applied to atmospheric and oceanic numerical predictions. In this study, a simplified data assimilation scheme is used to improve the quality of time-series MODIS NDVI data. A flowchart for reconstructing MODIS NDVI multi-temporal data is shown in Fig. 1.

(1) Historical multi-year NDVI data, extracted from MODIS VI products (i.e., MOD13A2), are used to produce a background field of NDVI based on a three-point smoothing method. The smoothing results describe the general features of the mean NDVI for every pixel occurring during the past years.
(2) Subsequently, 1 × 1 km 16-day composited NDVI are extracted with their corresponding QA data sets from MODIS VI products. These NDVI data are defined as observations and used to reconstruct high-quality data. The parameters of VI usefulness index in the QA data sets are used to determine the relative weight of the background NDVI field and observation NDVIs.

(3) Finally, the proposed data assimilation method is applied to calculate the optimal value of the NDVI at the current time according to the background field, the observed NDVI and the relative weight between them.

2.1. Data description

MODIS vegetation index products have a wide range of applications, including global biogeochemical and hydrologic modeling, agricultural monitoring and forecasting, land-use planning, land cover characterization, and land cover change detection. MOD13A2 products are 16-Day L3 Global 1 km SIN Grid VI data sets, which include two vegetation index types.

(1) The normalized difference vegetation index (NDVI).
(2) The enhanced vegetation index (EVI).

These grid vegetation indices include QA data sets with statistical data indicating the quality of the VI product. The MODIS VI algorithm makes use of a filter for data dependent on cloud and viewing geometry (Huete et al., 1999). The objective of this compositing methodology is to determine a single value per pixel from all the data retained by the filter. This value is assumed to be representative for each pixel over the 16-day period of interest. This VI compositing algorithm includes (1) the maximum value composite (MVC) and (2) a constraint on view angle – maximum value composite (CV-MVC). As a result, the quality of MODIS NDVI data is significantly enhanced (Huete et al., 2002).

In this paper, MOD13A2 products covering western China (73–112°E, 26–50°N), were acquired for the period of interest (POI) 2003–2006. These products are used to evaluate the performance of our method. For 16-day composited datasets, each pixel holds 23 observations per year. As described earlier, the first three years of NDVI data are used to generate a background NDVI field and the last year of NDVI data is applied as test data enabling the evaluation of method performance. During the process of NDVI data reconstruction, NDVI QA data are used to determine the weight required by the data assimilation method for every time step.

2.2. The data assimilation scheme

Typically, the data assimilation scheme is composed of a model operator, an observation operator, error estimators, and a minimization algorithm. A cost function is defined to enable to integrate the above components by fitting both the background field and time-dependent observations. It is expressed as (Daley, 1991):

\[ J(x_a) = (H(x_a) - y_o)^T R^{-1} (H(x_a) - y_o) + (x_a - x_b)^T B^{-1} (x_a - x_b) \]  \hspace{1cm} (1)

where \( J(.) \) is the cost function. \( H(.) \) is the observation operator, used to convert state variables into observations, \( y_o \) is an observation and \( R \) is the observation error covariance. \( x_a \) is the analyzed state variables. \( x_b \) is the background field. \( B \) is background field error covariance. The main objective of this data assimilation scheme is to minimize the cost function \( J(.) \) to obtain an optimal estimate of the state variables. In this paper, both the state variables and observations are NDVI values, hence the cost function can be simplified as

\[ J(NDVI_a) = \frac{(NDVI_a - NDVI_b)^2}{B_{NDVI}} + \frac{(NDVI_a - NDVI_o)^2}{R_{NDVI}} \]  \hspace{1cm} (2)

Where \( NDVI_a \), \( NDVI_b \), and \( NDVI_o \) are the analyzed, background field and observed NDVI, respectively. \( B_{NDVI} \) and \( R_{NDVI} \) are background and observation error variances of the NDVI. The analyzed value of NDVI is found by setting the gradient of the cost function to zero:

\[ \nabla J(NDVI_a) = 2 \frac{(NDVI_a - NDVI_b)}{B_{NDVI}} + 2 \frac{(NDVI_a - NDVI_o)}{R_{NDVI}} = 0 \] \hspace{1cm} (3)

From the above formula, the analyzed value of NDVI \( (NDVI_a) \) is given as:

\[ NDVI_a = NDVI_b + K \times (NDVI_a - NDVI_b) \] \hspace{1cm} (4)

\[ K = \frac{B_{NDVI}}{B_{NDVI} + R_{NDVI}} \] \hspace{1cm} (5)

\( K \) is a weight coefficient. The rationale of the data assimilation method used in this study is the determination of optimum values for the \( K \) elements to minimize total analysis error. If errors associated with the background field are large compared to the observations, then \( K \) will be large and a large correction will be made to the background to attain the optimal value. Conversely, if the background errors are small compared to the observation errors, \( K \) will be small, and the observations will only slightly influence the final analyzed value.

2.3. Generation of the background field

Because the noise on the NDVI signal induced by clouds and poor atmospheric conditions is negatively biased, the “upper NDVI envelope” smoothing strategy is usually applied to reconstruct NDVI data (Van Dijk et al., 1987; Chen et al., 2004; Jonsson and Eklundh, 2006). Temporal averaging techniques is a practical method for finding smooth traces for NDVI time series confounded with occasionally spikey noise. Temporal averaging techniques is
quite close to the ideal low-pass filter, yet they are robust and resistant to brief spikes (Velleman, 1980).

In this paper, a simple three-point smoothing technique is combined with a selection of maximum NDVI values to provide the background field for NDVI at every time step (Gu et al., 2006):

$$NDVI_b(t) = \text{MAX}[NDVI_o(t), (0.5 \ast NDVI_o(t) + 0.25 \ast (NDVI_o(t-1) + NDVI_o(t+1)))]$$

Where \((t-1)\) represents the data at the previous time-step, and \((t+1)\) represents the data at the following time-step. To remove noise more effectively, the smoothing formula is repeated three times. The final step for the temporal smoothing background field of NDVI is the averaging for the 3 years \((N = 3)\) available in the context of this study (i.e., 2003–2005):

$$VI_{\text{mean}}(t) = \frac{1}{N} \sum_{\text{year}}^{} NDVI_b(t, \text{year})$$

$$NDVI_o(t) = 0.5 \ast NDVI_{\text{mean}}(t) + 0.25 \ast (NDVI_{\text{mean}}(t-1) + NDVI_{\text{mean}}(t+1))$$

This method ensures that the NDVI background field is smooth and of high quality, which represents the gradual process dynamics of an annual vegetation cycle and captures the general dynamical features that occurred in past years.

### 2.4. Determination of weight

Since cloudiness and viewing geometry are variable for every time-step, it is difficult to directly estimate the background and observation errors for NDVI data. Fortunately, the MODIS VI products provide QA data sets, which reflect the quality of the NDVI data. In this study, we applied the QA data sets to empirically determine \(K\) for the corresponding pixel locations without having to calculate background and observation errors.

The QA data sets are coded with 16 bit values. Bits 2–5 represent the VI usefulness index. Its value for a specific pixel is determined by several boundary conditions, including aerosol quantity, atmospheric correction conditions, cloud cover, shadow, and sun-target-viewing geometry. Thus it actually represents a high resolution quality indicator (Huete et al., 1999). As illustrated in Table 1, a specific score is assigned to each boundary condition, and the sum of all the scores represents the usefulness index value for each pixel. An index value of 0000 corresponds with the highest quality, while the lowest quality is equal to a value of 1111. According to the VI usefulness index, we empirically assign a value to \(K\) for each QA flag, also illustrated in Table 1. When the QA flag is larger than 5, we assume the existence of a large observation error. Hence, the corresponding \(K\) value is set to zero. If \(K\) equals 1, the analyzed value is assumed equal to the observed NDVI. If \(K\) equals zero, the optimal value is given by the background field of NDVI.

### 3. Results and analysis

The simplified assimilation of MODIS NDVI data has been performed for 2006 according to the strategy described earlier in this paper. For a detailed assessment of the new method, comparisons are presented in the next chapters related to spatial patterns and temporal evolution. The dominant vegetation types in west China include needle leaf forest, broadleaf forest, shrub vegetation, grassland, woody savannah, meadows and swamps, cropland, sparse vegetation desert, barren land and water bodies (Fig. 2).

#### 3.1. Analysis of the spatial patterns of the NDVI

Figs. 3 and 4 illustrate NDVI maps with results analysis, for MODIS data. The difference between these results for days 1, 97, 193 and 298 are also presented. The figures indicate that some isolated anomalous pixels with nonsensical low values in the original NDVI image are effectively replaced by the previously mentioned algorithm. Regions with a high enhancement are distributed mainly in the northwest and southeast of our study region. They have been outlined and identified with A–C in Fig. 2, respectively.

On days 1 and 97 (Fig. 3), MODIS NDVI values are very low for the northern and southern regions of western China. This is mainly attributed to the strong snow effect on an NDVI increase. The analyzed NDVI shows that regions of increase are mainly located in the northwest and south of western China which seems both realistic and consistent with the land-cover databases for these regions, consisting mainly of broadleaf forest, shrub land, and crop land.

<table>
<thead>
<tr>
<th>Bit comb.</th>
<th>QA flag</th>
<th>Quality description</th>
<th>(K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000</td>
<td>0</td>
<td>Perfect quality</td>
<td>1</td>
</tr>
<tr>
<td>0001</td>
<td>1</td>
<td>High quality</td>
<td>0.9</td>
</tr>
<tr>
<td>0010</td>
<td>2</td>
<td>Good quality</td>
<td>0.8</td>
</tr>
<tr>
<td>0011</td>
<td>3</td>
<td>Acceptable quality</td>
<td>0.7</td>
</tr>
<tr>
<td>0100</td>
<td>4</td>
<td>Fair quality</td>
<td>0.6</td>
</tr>
<tr>
<td>0101</td>
<td>5</td>
<td>Intermediate quality</td>
<td>0.5</td>
</tr>
<tr>
<td>0110</td>
<td>6</td>
<td>Below intermediate quality</td>
<td>0.0</td>
</tr>
<tr>
<td>0111</td>
<td>7</td>
<td>Average quality</td>
<td>0.0</td>
</tr>
<tr>
<td>1000</td>
<td>8</td>
<td>Below average quality</td>
<td>0.0</td>
</tr>
<tr>
<td>1001</td>
<td>9</td>
<td>Questionable quality</td>
<td>0.0</td>
</tr>
<tr>
<td>1010</td>
<td>10</td>
<td>Above marginal quality</td>
<td>0.0</td>
</tr>
<tr>
<td>1011</td>
<td>11</td>
<td>Marginal quality</td>
<td>0.0</td>
</tr>
<tr>
<td>1100</td>
<td>12</td>
<td>Low quality</td>
<td>0.0</td>
</tr>
<tr>
<td>1101</td>
<td>13</td>
<td>No atmospheric correction performed</td>
<td>0.0</td>
</tr>
<tr>
<td>1110</td>
<td>14</td>
<td>Quality too low to be useful</td>
<td>0.0</td>
</tr>
<tr>
<td>1111</td>
<td>15</td>
<td>Not useful for other reasons</td>
<td>0.0</td>
</tr>
</tbody>
</table>

*(Huete et al., 1999).*
On day 193 (left panel of Fig. 4), the observed and analyzed NDVI data show better agreement than expected. Only the NDVI analysis in the southern part of western China increased from approximately 0.1–0.3. The main causes are heavy precipitation events and cloud contamination.

On day 298 (right panel of Fig. 4), differences between the analysis and observations mainly appear in the south of western China. These are attributed to contamination by clouds and other atmospheric processes. For these regions, most of the pixels from the MODIS NDVI maps have values close to zero. This is unexpected since the region should be covered by vegetation at that time of the year. The NDVI analysis based on the simplified assimilation method seems to yield more realistic values. In conclusion, this method is capable of removing the effects of snow and cloud contamination.

3.2. Temporal evolution analysis of NDVI

To investigate the de-noising performance of this method more fully, we undertook a further analysis of the NDVI time-series. 12 test points for different vegetation types were chosen and indicated over the selected three rectangle regions (A–C) in Fig. 2.

Fig. 5 shows the NDVI time series from background field, observation, analysis and the corresponding QA flag at the selected test points for 2006. As illustrated in Fig. 5, the NDVI evolution obtained by using the three-point “upper envelope” smoothing method illustrates the smoothing method works well when perturbations (errors) occur only for one data point but fail to effectively remove noise for continuously contaminated data. These results also give insight into the influence of QA flag on the background field and observation. Most, NDVI observations take lower values than the NDVI background under the boundary condition that the QA flag value is less than 5, but close to the NDVI background value when QA flag value is higher than 5. Hence, to some extent the QA flag explains the uncertainty of the NDVI observation. Additionally, most of the noisy points in the NDVI time-series – caused primarily by cloud contamination and atmospheric variability – are successfully identified and corrected. They strongly suggest that the final value of the NDVI, (e.g., the analyzed NDVI) exhibits a more realistic temporal evolution.

For annual double cropped land cover (A3 point), the highest yearly variations of the NDVI occur during the months April and August. Since the background of NDVI is generated by a three points smoothing method, the two NDVI peaks phenomenon is not easily explained. When coupling the background, observation of NDVI and the corresponding QA flags, the resultant analyzed NDVI reflects a good agreement with its actual annual dynamics. Additionally, as can be observed from the points B1, B4, C2 and C3, continuous contamination for the observed NDVI is apparent. Nonetheless, noisy points are still accurately identified and removed. Hence, the analyzed NDVI time-series is much smoother and more realistic compared to the observed NDVI. This indicates that the simplified assimilation method is insensitive to neighbor effect values and is able to deal with continuous noises as well in multi-temporal NDVI profile processing.

4. Conclusions

With the increasing use of the NDVI in vegetation monitoring research, the quality of the NDVI products requires greater attention. Much debate exists with respect to NDVI reconstruction methods, but few studies focus on the
MODIS NDVI product and the use of ancillary QA data sets. A simplified data assimilation method is therefore proposed in this paper to reconstruct multi-temporal time-series MODIS NDVI data. In this work, we use a multi-temporal time-series data smoothing method to produce NDVI background fields, and we take advantage of ancillary QA data sets of NDVI product to determine the weight. Then the analyzed NDVI is calculated by the MODIS NDVI observation, background field, and empirical weight factor. The performance of this method has been successfully tested and validated for MODIS NDVI data covering western China.

The background NDVI field has certain limitations in capturing detailed variations, but it can offer valuable information with respect to identifying and correcting noisy points in a multi-year average profile of vegetation. The ancillary QA flags are an effective means for determining the weight of the NDVI background, and for...
avoiding information loss during the analysis process. Finally, the final NDVI analyses multi-temporal profiles based on the presented methods are of a high quality, show a smoother temporal evolution, and elicit more consistency with respect to land-cover databases. Additionally, this method reduces continuous contamination of the multi-temporal NDVI evolution without the effects of adjacent points in the multi-temporal temporal NDVI profile. An important reason for this is the use of the synchronous QA flag to estimate the NDVI and to enable realistic corrections.

However, the current reconstructing scheme can not always perform well at the onset of spring green due to the average background. The effects of outliers (e.g. strong spurious lows) will be considered as well for reducing the extension from neighboring data points. So a more effective scheme to determine the weight coefficient $K$ will be developed.
Fig. 5. Comparison of the multi-temporal time-series of MODIS NDVI data for background, observation, and analysis NDVI data for 2006.

In summary, the simplified assimilation method is theoretically simple and easy to implement for the reconstruction of multi-temporal MODIS NDVI time-series for different time intervals. The analyzed NDVI – which is a final output result – is a more realistic and appropriate variable for global vegetation dynamics research.
Acknowledgements

This research was supported by the CAS (Chinese Academy of Sciences) Action Plan for West Development Project (Grant No.: KZCX2-XB2-09), the NSFC (National Science Foundation of China) project (Grant Nos.: 40771036, 40871190, 40801126) and Chinese Academy of Sciences Knowledge Innovation Project (Grant No.: KZCX2-YW-Q10-2-4). The authors thank the anonymous reviewers for their constructive comments and suggestions.

References


