Anomaly Detection in MODIS Land Products via Time Series Analysis

ZHANG Jingxiong  David Roy  Sadashiva Devadiga  ZHENG Min

Abstract  With remote sensing information products becoming increasingly varied and arguably improved, scientific applications of such products rely on their quality assessment. In an operational context such as the NASA (National Aeronautics and Space Administration) information production based on the MODIS (Moderate Resolution Imaging Spectroradiometer) instrument on board Earth Observing System (EOS) Terra and Aqua satellites, efficient ways of detecting product anomaly, i.e., to discriminate between product artifacts and real changes in Earth processes being monitored, are extremely important to assist and inform the user communities about potential unreliability in the products. A technique for anomaly detection, known as MAD (the median of absolute deviate from the median) in MODIS land products via time series analysis is described, which can handle intra- and inter-annual variation in the data by using MAD statistics of the original data and their first-order difference. This method is shown to be robust and work across major land products, including NDVI, active fire, snow cover, and surface reflectance, and its applicability to multi-disciplinary products is anticipated.

Keywords  anomaly detection; MODIS land products; time series

CLC number  P237.4; TP753

Introduction

A variety of terrestrial information is operationally derived from satellite data including those from NASA’s MODIS instrument. There are increasing awareness and interests in the accuracy and reliability of satellite sensor products[1]. Scientific applications of MODIS land products rely on their quality assessment (QA), which is facilitated by monitoring and analysis of time series of summary statistics derived from all the gridded MODIS land products at a number of fixed globally distributed locations. The value of time series analysis to QA is due to the facts that it captures algorithmic sensitivity to surface, atmospheric, and remote sensing conditions that change temporally, and that it allows changes in the instrument characteristics and calibration to be examined[2].

Time series images or their derivatives, such as summary statistics stratified on the basis of biome or land cover types, are valuable for monitoring ecosystem because of their capabilities of tracking the intra- and inter-annual variation of the phenomena being mapped. However, the interaction of intra- and inter-annual variation and real change affects the accuracy of change detection. Further complication arises due to anomaly in satellite sensor-derived products, which may be signaled by either abrupt changes in data values or anomalous/extreme values with respect to the usual annual or seasonal patterns observed. Thus, a key requirement in automatic QA is to be able to differentiate anomaly (i.e., outliers) from real
changes driven by either natural or human-induced forcing, while intra- and inter-annual variation is well accounted for in the inferential process.

There is large literature on time series analysis and outlier detection therein\cite{3-5}. Assuming that observations are independently and identically normally distributed, Fox\cite{6} discussed two types of outliers and their detection using likelihood ratios. Type 1 outliers occur where a gross error of observation affects a single observation whilst Type 2 outliers will affect not only a particular observation but also subsequent observations. More recent work can be seen in Reference\cite{7}, in which the estimation and detection of outliers in a time series generated by a Gaussian autoregressive moving average process is considered. Such conventional outlier detection is performed on the assumption that all trend and seasonal factors have been removed or accommodated properly, leaving only possibly serially correlated random deviates in the process of model identification and statistical estimation. This may not be the case for real time series data with significant seasonality and annual variability, whose modeling is far from trivial. Thus, for MODLAND products QA, it is important to perform time series analyses using methods appropriate for the data set at hand, with solid understanding of the complexity and heterogeneity of such data.

An important characteristic of MODIS land time series data is that they exhibit both intra- and inter-annual variations, as hinted on previously, which behave differently with different types of MODLAND time series data. One type of time series data is that of NDVI (normalized difference vegetation index) data, which have been widely used in environmental modeling and geophysical studies, as they provide valuable information about vegetation states and abundance, which are important indicators in global change\cite{8}. Such data exhibit strong annual pattern, which may be effectively modeled by some parametric functions, such as Gaussian and logistic, as shown in Jonsson and Eklundh\cite{9} and Zhang, et al.\cite{10}, respectively. Another type of MODLAND time series data includes surface reflectance, snow, and fire pixel counts, which have different seasonality than the phenology observed in NDVI and other vegetation index (VI) data. With such data sets, there is relative lack of generally regular and cyclic pattern as seen in NDVI data. Thus, their seasonality may not be easily modeled using parametric functions. Clearly, these two types of MODLAND time series data should be treated differently with respect to their behaviors.

For NDVI or similar data, it is possible to apply parametric regression to check any suspiciously large residuals based on some statistics. One such statistic is standardized residuals approximately following \textit{t}-distributions, which are residuals normalized by the standard deviation (\(\hat{s}\)) estimated from regression (or least squares) residuals. At significance levels of 0.05 and 0.01, thresholds of 2 and 3 can be applied to the calculated \textit{t}-statistics to detect yellow and red alarms, respectively. Another one is studentized residuals, which are residuals normalized by the regression standard deviation multiplied by the square root of \((1 - h_{ii})\), where \(h_{ii}\) is the diagonal elements of the hat matrix \(H\), which is \(X(X'X)^{-1}X'\) for a linear model \(Y = X\beta + \epsilon\)\cite{11}. Clearly, \(H\) is the derivative of the prediction \(\hat{y}_i\) with respect to \(y_i\)\cite{12}. In comparison, the studentized residuals are more sensitive in detecting deviates from regressed functions, and are dependent on the leverage (influence) of data points as measured by \(h_{ii}\).

For those data other than VI or the like, it may be tempting to apply the regression-based methods for outlier detection. The lack of regularity in non-VI data sets precludes easy specification of parametric functions in regression-based outlier detection. Although high-order sinusoidal functions can in principle approximate many real data, function-fitting can become extremely difficult due to cross-product difference. Even for VI data, there is an issue of specifying suitable trend and seasonal functions, and mis-specification of the models adopted for regression analysis can lead to inaccurate detection of anomaly, because the residuals and standard errors estimated from regression based on erroneous models are nowhere reliable\cite{13}. Non-parametric approaches are, therefore, needed for robust outlier detection, which should be applied to any time series data, both VI and non-VI types. This paper describes a non-parametric approach to anomaly detection based on MAD which works for very different MODLAND products. Results with real MODLAND time series data are presented, followed by some discussion.
1 MAD-based anomaly detection

A key element in time series analysis is to adjust for seasonality. For time series running many years, it is possible to take advantage of the seasonal signatures built up through accumulation of data in time. This allows for a non-parametric approach to seasonal adjustment and, hence, outlier detection, as discussed below.

Suppose a given time series data set is temporally split into bins (windows) of equal lengths, e.g., every 32 d to match the multiples of typical data periods of MODLAND products, which may be daily, every 8 or 16 d. These temporal windows can also be built from allocating the whole length of time series data into particular months or their groupings, such as every month or every three months. Let $y(x, m, yr)$ be the monthly or tri-monthly data value (e.g., NDVI) of pixel-patches $x$ (usually grouped by biome or land-cover types) in month $m$ (or the month-sequel centered at month $m$) and year $yr$. Let $\bar{y}(x, m)$ be the long-term monthly or seasonal average of $y$ in month $m$ (or the season centered at $m$). The mean $\bar{y}(x, m)$ and standard deviation ($s$) are calculated:

$$\bar{y}(x, m) = \frac{1}{n_{yr}} \sum_{yr=1}^{n_{yr}} y(x, m, yr)$$

$$s(x, m) = \sqrt{\frac{1}{(n_{yr}-1)} \sum_{yr=1}^{n_{yr}} (y(x, m, yr) - \bar{y}(x, m))^2}$$

where $n_{yr}$ is the number of years for which $y(x, m, yr)$ is available, as also described in Reference [14].

It is possible to treat a temporal window’s data points of one particular year, usually the current year, as the data set to be tested, while the rest data points in the same temporal window are used as reference. Statistical methods exist for detecting any anomalous data periods in the current year in comparison with the usual seasonal pattern manifested by the majority of the time series data obtained so far. There is $t$-test commonly applied for significance testing of differences in sample means. Assume there are two samples of data: $y_1 \in \{ y_{11}, y_{12}, \ldots, y_{1n_1} \}$ and $y_2 \in \{ y_{21}, y_{22}, \ldots, y_{2n_2} \}$. On the assumption of equal variance, the $t$-statistic is defined as:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_{\bar{y}_1 - \bar{y}_2}}$$

where the standard deviation in the difference is:

$$s_{\bar{y}_1 - \bar{y}_2}^2 = s_{\bar{y}_1}^2 + s_{\bar{y}_2}^2 = s_{\bar{y}_1}^2 \text{ or } s_{\bar{y}_2}^2$$

where the pooled sample variance is estimated as:

$$s_{\text{pooled}}^2 = \frac{1}{n_1 + n_2 - 1} \left( (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 \right)$$

where $\bar{y}_1$, $\bar{y}_1$, $\bar{y}_2$, and $\bar{y}_2$ are sample mean and variance for sample $y_1$ and $y_2$, respectively.

Thus, the degrees of freedom $v$ for the $t$-statistic is $v = n_1 + n_2 - 2$. The significance testing is stated as checking null versus alternative hypotheses:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

Rejection region for $H_0$: either $t \geq t_{v/2}$ or $t \leq -t_{v/2}$.

Depending on the longevity of time series data, there is an issue of trade-off between temporal window length and precision in locating anomalous points or their patches. If there is sufficient accumulation of time series data, which is the case for AVHRR VI time series (running more than 20 a), it is possible to perform point-specific detection of outliers. This can be done by $z$-score test, as employed in the evaluation of global NDVI anomaly by Myneni, et al[14]. In a $z$-score test, the mean ($\bar{y}$) and standard deviation ($s$) of the entire data set are used to obtain a $z$-score for each data point, according to the following formula:

$$z_i = \frac{y_i - \bar{y}}{s}$$

However, the performance of either $t$-test or $z$-score test relies on the sample size. There are only about 5 years’ data for MODLAND products, which is far from adequate. Some other more robust methods that are less sensitive to the length of time series data are needed to equip automatic QA in MODLAND products. One way is to use outlier resistant estimators of $z$-score. MAD is such an estimator. By MAD, it is possible to perform anomaly detection based on simple calculation of median and MAD over temporal/seasonal windows for a particular data point against the rest of data points in the bin (1 month or 3 months set). Let the set of data points
in the seasonal window be \( y \in \{ y_1, y_2, \cdots, y_n \} \). To test point \( y_i \), it is straightforward to calculate the its deviate (denoted by \( \text{dev}_i \)) from the median of the seasonal data set:

\[
\text{dev}_i = y_i - \text{median}(y_{1:i-1})
\]

where \( y_{1:i-1} \) represents the seasonal data set excluding point \( i \), which will not be affected by unknown nature of the point being tested. This is a novel way of stabilizing median and MAD estimation for improved accuracy in outlier detection.

With the set of deviate being calculated by Eq.(5), MAD is calculated as the median of the absolute values of deviates of all points in the set of \( y \), that is:

\[
\text{MAD} = \text{median}(|\text{dev}_i|)
\]

The \( z \)-score for MAD-based testing of \( \text{dev}_i \) is:

\[
z_i = 0.6745 \times \frac{|\text{dev}_i|}{\text{MAD}}
\]

It is easy to see that the detection rule is: if \( z_i \geq 3 \) then declare point \( i \) as possible alarm.

Due to the complication of inter- versus intra-annual variations, it is important to desensitize the MAD detector against error of commission whereby data points are wrongly identified as outliers simply because of an anomalous year, i.e., extreme inter-annual variations. One way of reinforcing the performance of anomaly detection is to apply the MAD detector not only to the original \( y \) data but also their first-order difference, i.e., \( \Delta y \), which is calculated as the difference of adjacent data points in the seasonal window:

\[
\Delta y_i = y_i - y_{i+1}
\]

whose function is such that only “outliers” identified by both the \( y \)- and \( \Delta y \)-MAD detectors are marked as true outliers.

Anomaly detection has to be adapted to the properties of the time series under study here. The first peculiarity concerns the inhomogeneity of the actual population from which individual data points (summary statistics) were computed. To aid interpretation of the detection results, extra information was provided along with the \( y \) data, i.e., numbers of pixels for individual data points from the database. A look-up table for the numbers of pixels falling into individual biome or land-cover types is stored, which are the maximum numbers of pixels belonging to respective biome or land-cover types on actual MODIS imagery and products due to varying cloud cover. Thus, a minimum percentage of these maximum numbers of pixels is used to filter out cloudy data points.

Another point yet about this study is that sound analysis comes from good knowledge about the data at hand and, at a deeper level, the underlying processes that have “created” the data. An intuitive rule for residual examination is that any residuals whose absolute values are less than some thresholds of values (100 for snow data and 25 for fire data, which are figured out by error and trial) should be ignored even if the calculated MAD-statistics exceed the thresholds of alarms. This will of course reduce error of commission.

2 Results with MODIS land products

Real MODLAND time series data sets were loaded up from LDOPE (the MODIS Land Data Operational Product Evaluation) database (collection 4:C4), with data dates ranging from August 2000 to end of 2004. Some examples are shown below to demonstrate the effectiveness of combining MAD of original data and their first-order difference in outlier detection.

Various data sets, such as NDVI, surface reflectance, NBAR (nadir BRDF-adjusted reflectance), snow cover, and detected active fire time series data, were employed. Fig.1 shows the working of MAD detection, where the MAD test based on original data, their first-order difference, and combination of detected alarms clearly demonstrate the complementarity of the tests based on \( y \) and \( \Delta y \), which emphasize inter-annual variation and local abrupt changes, respectively. The detected anomalies are shown as solid squares in Fig.1 and other diagrams to follow, where each data point is also labeled with its corresponding month. Clearly, the combination of MAD detections based on the original time series data and their first-order help to reduce error of commission.

Some more examples are shown in Fig.2 (NDVI) and Fig.3 (surface reflectance), where the effects of temporal window lengths (3 months vs. 1 month) can be inspected. As expected, narrower temporal windows tend to pick up more alarms, while wider temporal windows smooth out some local abrupt changes.
As narrowing temporal windows reduce the number of data points for seasonal statistics and detection, a data point under examination may not have sufficient (minimum 2) data to test against. If this occurs, a solid vertical line will be drawn across such data points. In Fig.2, the dashed vertical line is drawn to mark that the data point with extremely low value is not reliable due to insufficient number of cloud-free pixels available.

The proposed MAD-based anomaly detection works not only for NDVI data but also for other major data. Detection results across different products are shown in Fig.4, including those of surface reflectance, NBAR, active fire, and snow extent. The peculiarity about fire and snow data is that they are pixel counts of binomial events, implying that any “anomalous” counts relative to seasonal localities are significant only if they are sufficiently large in numbers. As hinted in the previous section, some thresholds of values (100 for snow data and 25 for fire data, as marked by the horizontal lines in Fig.4(c) and 4(d),
respectively) were set to filter out “outliers” even if the calculated MAD-statistics exceed the thresholds of alarms. The net result is that false alarms are rightly disregarded as anomalies.

![Fig.4 Anomaly detection with different data](image)

### 3 Conclusions

This paper has documented a robust approach to anomaly detection based on the MAD statistic, which is a major step toward automated quality assessment in operational remote sensing products evaluation. Results with real MODLAND QA data sets confirmed the effectiveness of the proposed method, which is enhanced with the ability to deal with the complication of intra- and inter-annual variation in time series data concerning the underlying Earth processes, flexibility of setting seasonal smoothing windows, and smart thresholds to desensitize the MAD detector against false alarms. The MAD-based technique was also successfully applied to NDVI and LST time series collected over Wuhan city, stratified to different IGBP land cover types, which were confirmed to be non-changing based on MODIS land cover products. Results from these on-going tests will be reported later via other avenues.

It is anticipated that the MAD-based approach will be able to be applied to large-scale QA activities in the remote sensing of land, ocean, and atmosphere alike. Further research should be directed towards analyses exploring and accounting for spatio-temporal dependence inherent in long-term global remote sensing products, as points along time series curves represent statistical quantities that are spatially aggregated and temporally composited.

### Acknowledgements

Our thanks to Mr. Masuoka and Dr Robert Wolfe (NASA official at Goddard Space Flight Center) for their advice and helps. The paper presented here is solely the leading author’s responsibility, who was with Land Data Operational Product Evaluation (LDOPE) Facility in the capacity of a senior scientist between 2003 and 2005. Any opinion expressed in this paper is his own, and not that of NASA.
References

[1] Justice C O, Belward A, Morisette J, et al. (2000) Development in the “validation” of satellite sensor products for the study of the land surface [J]. *International Journal of Remote Sensing*, 21: 3383-3390

[2] Roy D, Borak J, Devadiga S, et al. (2002) The MODIS land product quality assessment approach [J]. *Remote Sensing of Environment*, 83:62-76

[3] Barnett V, Lewis T (1984) Outliers in statistical data [M]. New York: John Wiley & Sons

[4] Tsay R S (1988) Outliers, level shifts, and variance change in time series [J]. *Journal of Forecasting*, 7:1-20

[5] Peña D (1990) Influential observations in time series [J]. *Journal of Business and Economic Statistics*, 8: 235-241

[6] Fox A J (1972) Outliers in time series [J]. *Journal of the Royal Statistical Society, B*, 34: 250-263

[7] Ljung G M (1993) On outlier detection in time series [J]. *Journal of the Royal Statistical Society, B*, 55(2): 559-567

[8] Kaufmann R K, Zhou Liming, Myneni R B, et al. (2003) The effect of vegetation on surface temperature: a statistical analysis of NDVI and climate data [J]. *Geophysical Research Letters*, 30(22):2 147-2 152

[9] Jonsson P, Eklundh L (2002) Seasonality extraction by function fitting to time-series of satellite sensor data [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 40(8): 1824-1832

[10] Zhang Xiaoyang, Friedl M A, Schaaf C B, et al. (2002) Monitoring vegetation phenology using MODIS [J]. *Remote Sensing of Environment*, 84: 471-575

[11] Rice J A (1995) Mathematical statistics and data analysis [M]. Belmont: Wadsworth Inc.

[12] Snedecor G W, Cochran W G (1989) Statistical methods [M]. Ames, IA: Iowa State University Press

[13] Rousseeuw P J, Leroy A M (1987) Robust regression and outlier detection [M]. New York: John Wiley & Sons

[14] Myneni R B, Tucker C J, Asrar G, et al. (1998) Interannual variations in satellite-sensed vegetation index data from 1981 to 1991 [J]. *Journal of Geophysical Research*, 103(D6): 6145-6160