A Weighted Least-Squares Approach to Temporal NDVI Smoothing

Daniel L. Swets, Computer Science Department, Augustana College, 2001 S. Summit Ave, Sioux Falls, SD 57197, swets@inst.augie.edu

Bradley C. Reed and James D. Rowland, Raytheon ITSS, EROS Data Center, Sioux Falls, SD 57198, reed@edcmail.cr.usgs.gov, rowland@edcmail.cr.usgs.gov

Shaun E. Marko, Computer Science Department, Augustana College, 2001 S. Summit Ave, Sioux Falls, SD 57197, semarko@inst.augie.edu

Abstract

Satellite imagery provides a unique vantage point for observing seasonal dynamics of the landscape that have implications for global change issues. An objective evaluation of surface conditions may be performed using the normalized difference vegetation index (NDVI) derived from National Oceanic and Atmospheric Administration advanced very high resolution radiometer data. NDVI data are typically very noisy, affected by a number of phenomena including cloud contamination, atmospheric perturbations, and variable illumination and viewing geometry, each of which usually reduces the NDVI. This work describes a weighted least-squares linear regression approach to temporal NDVI smoothing to more efficiently reduce contamination in the NDVI signal. This approach uses a moving window operating on temporal NDVI to calculate a regression line. The window is moved one period at a time, resulting in a family of regression lines associated with each point; this family of lines is then averaged at each point and interpolated between points to provide a continuous temporal NDVI signal. Also, since the factors that cause contamination usually serve to reduce NDVI values, the system applies a weighting factor that favors peak points over sloping or valley points. A final operation assures that all peak NDVI values are retained. The resulting relationship between the smoothed curve and the original data is statistically based. The smoothed data may be used to improve applications involving time-series NDVI data, such as land cover classification, seasonal vegetation characterization, and vegetation monitoring.

INTRODUCTION

Vegetation indices, computed from combinations of visible red and near-infrared spectral measurements, have been commonly used over the past 20 years for studying vegetation characteristics using images collected by remote sensing satellites. The advantages of using these numerical transformations rather than strictly spectral observations include minimizing soil and other background effects, reducing data dimensionality, providing a degree of standardization for comparison, and enhancing the vegetation signal (Curran 1981) (Malingreau 1989) (Goward 1989). One of the more commonly used vegetation indices, the normalized difference vegetation index (NDVI), takes advantage of the reflective and absorptive characteristics of plants in the red and near-infrared portions of the electromagnetic spectrum and has been used in research on vegetation productivity. The combination of the NDVI with the frequent temporal coverage and moderate spatial resolution of the advanced very high resolution radiometer (AVHRR) make this sensor well suited for regional- to global-scale studies on ecosystem dynamics.

The AVHRR sensors are carried on weather satellites operated by the National Oceanic and Atmospheric Administration (NOAA). The AVHRR records data with 1-km nominal spatial resolution and has daily global coverage. The primary advantage of the AVHRR is its frequent temporal coverage over large geographic areas, which allows an opportunity for obtaining cloud-free coverage during important phenological stages of the land cover. A number of studies directed toward analyzing global or continental-scale vegetation patterns using AVHRR data were initiated in the 1980's (Goward 1985, Justice 1985, Tucker 1985, Townshend 1987). Many of these investigations incorporated multitemporal profiles (NDVI plotted against time) (Figure 1), frequently using these

graphs as indicators of vegetation phenology. A growing body of research indicates that the phenological behavior of different broad vegetation types can be observed, analyzed, and mapped using the NDVI profiles (Defries 1994) (Reed 1994).

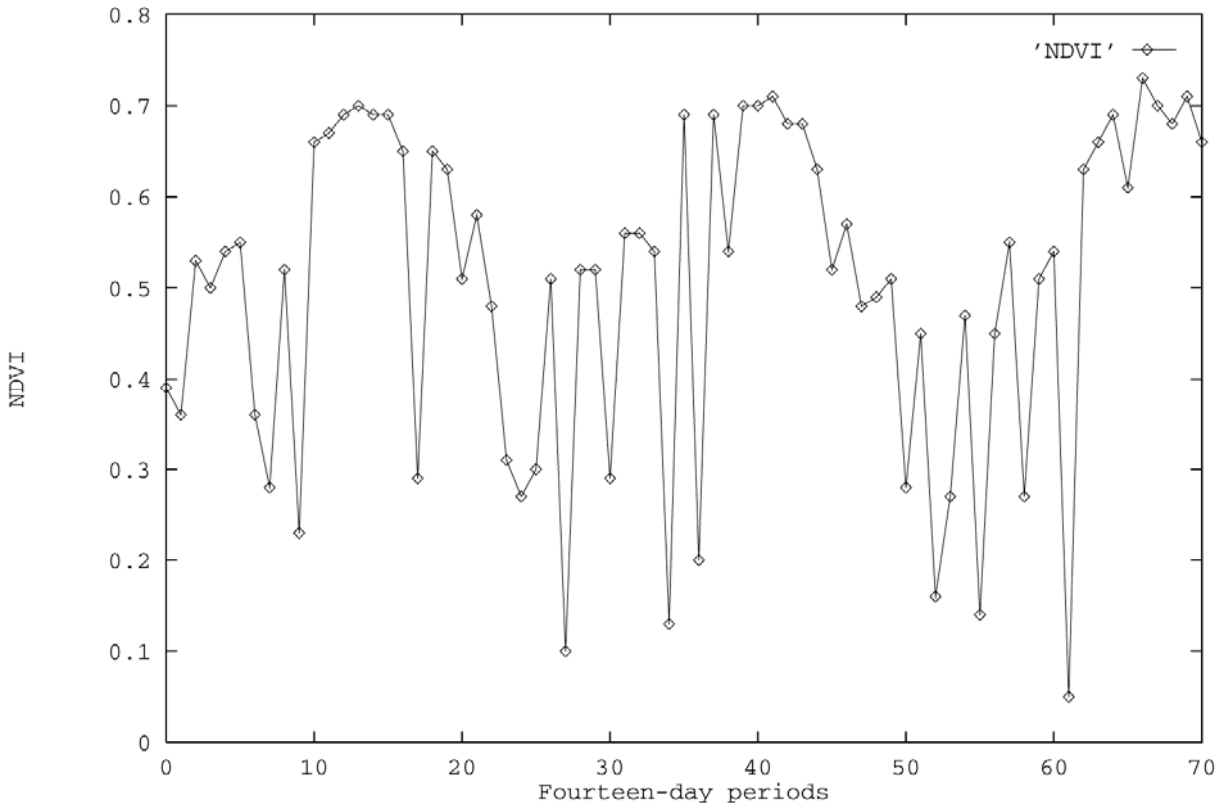


Figure 1. An example NDVI time series from AVHRR data.

Several standard NDVI data sets have been produced to enable land cover studies using different temporal intervals. For example, selecting one NDVI value to represent the set of values obtained in a 2-week period produces a biweekly data set. These standard data sets have both positive and negative characteristics for temporal vegetation studies. Positive characteristics include (1) a relatively dense temporal coverage, (2) a relatively low data volume covering large areas, and (3) reduced cloud contamination. Negative aspects of the biweekly data set are that (1) atmospheric corrections may be incomplete and (2) the biweekly composite period may be too long to determine phenological events in sufficient detail for some applications.

Noise Reduction

The NDVI data are often affected by a number of phenomena, including cloud contamination, atmospheric perturbations, and variable illumination and viewing geometry, all of which tend to reduce the NDVI value (Los 1994). Figure 1 illustrates the characteristics of some of these effects with the sharp, but temporary, reductions in NDVI values. Although radiometric corrections of AVHRR data are routinely performed, there are still problems due to unreliable preflight calibration, no onboard calibration, and difficulty with in-flight calibration. Atmospheric correction is performed only for Rayleigh scattering and for ozone. Atmospheric aerosols and water vapor corrections are not performed because there is a lack of community agreement on feasibility stemming from difficulties presented by spatial and temporal variability. Due to these negative effects, further processing---in effect, a smoothing---of the temporal NDVI signal is required. Furthermore, the requirement to smooth large

datasets (e.g., 17 years of continent-wide 10-day composites), as well as the desire to smooth the NDVI signal in near real-time, requires that the smoothing algorithm be computationally efficient.

NDVI compositing

Maximum value compositing is often performed on NDVI data sets to reduce both cloud contamination and data volume. The first factor to consider in the compositing process is the length of the compositing period. The choice of the period is usually based on the length of time necessary to obtain a composite with minimal cloud contamination and/or the amount of time necessary to observe meaningful changes in surface characteristics. The compositing period for AVHRR data has generally been 10 or 14 days. The data used in this study are biweekly or 14-day composites.

Maximum value NDVI compositing is performed by examining the NDVI pixel by pixel for each observation during the compositing period to determine the maximum value. The retention of the highest NDVI value reduces the number of cloud-contaminated pixels and favors near-nadir pixels (Holben 1986). Although maximum value compositing increases data quality, data analysis remains affected by the residual effects of sub-pixel clouds, prolonged cloudiness, persistent haze, and other negative effects.

BISE (Best Index Slope Extraction). Viovy and colleagues (Viovy 1992) proposed a method of temporal profile extraction based on the slope of increasing or decreasing data values referred to as the "best index slope extraction". The method accepts a point if it has a higher value than the previous observation. Where the NDVI value decreases, the decrease is only accepted if there is no point within the next *n* periods with a value greater than 20 percent of the difference between the first low value and the previous high value. This technique is dependent on both the 20 percent threshold and the predefined period of time (i.e., the *n*). The resulting profiles tend to lose some of the nuances of the NDVI profile and, in some cases, appear to be insensitive to the timing of NDVI increases.

Temporal Smoothing. VanDijk and others (VanDijk 1987) applied six different smoothers to vegetation index data: polynomial, Fourier, mean, median, "4253H, twice," and "3RSSH, twice." The shortcoming of the polynomial and Fourier smoothers is that they only determine the general shape of the curve, rather than pinpointing particular cycles. The Fourier series smoother must also be rerun over the entire time series each time a new data point is added. The running mean and median filters have the disadvantage (even when weighted) of altering the timing of local maxima and minima. Combinations, or compound, filters can also be used for smoothing. Velleman and Hoaglin (Velleman 1981) developed the "4253H, twice," and "3RSSH, twice" filters. The numbers of the 4253H refer to the span of the running median filter being run in each iteration, and the H refers to a hanning or weighting system for subsequent weighted mean filters. Smoothed residuals are then added back in a re-roughing process, hence the term "twice." The "3RSSH, twice" smoother uses a running median of span 3, then re-smoothes repeatedly until the data no longer change. A re-roughing process is also used in this compound smoother. The authors conclude that the "4253H, twice" smoother was the best of the above examples. However, it appears that local maxima values are not always retained in the resulting smoothed profile. The authors also declare that the inability of smoothing the endpoints of a profile limit the utility of the smoother for real-time monitoring applications. Finally, the inherent computational inefficiency of median filters does not allow the implementation of real-time monitoring on large datasets.

The work described herein addresses some of the drawbacks of the other approaches. The system uses a weighted least-squares windowed linear regression approach to temporal NDVI smoothing. The linear regression model is a well-understood statistical method that can be efficiently implemented to run on large sets of NDVI data. The resulting smoothed curve has the further favorable characteristic that it is statistically bound to the original raw data points.

WEIGHTED WINDOWED LINEAR REGRESSION

To extract vegetation information from satellite data, the noise embedded in the signal must be removed. This work reports on a technique for producing a smooth vegetation index curve using stable, robust, statistical methods as described below.

Weighted linear regression

Given a set of noisy observations, such as the temporal NDVI data for a particular pixel, we want to find a model that more succinctly describes the dataset (Press 1988) in order to more easily use it for further analysis. Regression is a well-known, statistically sound technique for obtaining the parameters for such a model and has been studied extensively (e.g., (Larsen 1986), (Press 1988), (Rousseeuw 1987), (Walpole 1985), (Wilks 1963)).

If we have a single independent variable x and a single dependent random variable Y, the data may then be represented by the pairs of observations $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$. If we assume that all of the means $\mu_{Y|x_i}$ lie on a straight line, then each y_i can be described by the simple linear regression model

$$Y_i = \mu_{Y|x_i} + E_i = \alpha + \beta x_i + E_i,$$

where the random error E_i has a mean of zero. Thus, each sample instance satisfies the equation

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

where ε_i is the value assumed by E_i when Y_i takes on the value y_i. If we estimate α by a and β by b, we get an estimated line

$$\hat{y} = a + bx.$$

Now each observation pair satisfies the relation

$$y_i = a + bx_i + e_i$$

where the "residual" $e_i = y_i - \hat{y}_i$ describes the model error at the i th data point.

For weights $w_i, i = 1,2,\ldots,n$ defined for each data point, the method of the "least sum of weighted squares" finds estimates a and b so as to minimize the sum of the weighted squared errors (SWSE), namely

$$SWSE = \sum_{i-1}^{n} w_i e_i^2 \qquad (1)$$

$$= \sum_{i-1}^{n} w_i (y_i - \hat{y}_i)^2 \qquad (2)$$

$$= \sum_{i-1}^{n} w_i (y_i - a - bx_i)^2. \qquad (3)$$

To minimize SWSE, we find the partial derivatives with respect to both a and b, and set them to 0:

$$\frac{\partial SWSE}{\partial a} = -2\sum_{i=1}^{n} w_i (y_i - a - bx_i)$$

$$= -2\sum_{i=1}^{n} (w_i y_i - aw_i - bx_i w_i)$$

$$\frac{\partial SWSE}{\partial b} = -2\sum_{i=1}^{n} x_i w_i (y_i - a - bx_i)$$

$$= -2\sum_{i=1}^{n} (w_i x_i y_i - aw_i x_i - bw_i x_i^2)$$

Setting to 0 yields:

$$\sum_{i=1}^{n} w_i y_i = a\sum_{i=1}^{n} w_i + b\sum_{i=1}^{n} x_i w_i \tag{4}$$

$$\sum_{i=1}^{n} w_i x_i y_i = a\sum_{i=1}^{n} w_i x_i + b\sum_{i=1}^{n} w_i x_i^2 \tag{5}$$

From equation 4 we have:

$$a = \frac{\sum w_i y_i - b\sum w_i x_i}{\sum w_i} \tag{6}$$

Substituting equation 6 into equation 5 and rearranging terms, we get

$$b = \frac{\left(\sum w_i\right)\left(\sum w_i x_i y_i\right) - \left(\sum w_i y_i\right)\left(\sum w_i x_i\right)}{\left(\sum w_i\right)\left(\sum w_i x_i^2\right) - \left(\sum w_i x_i\right)^2} \tag{7}$$

Notice that if we use a uniform weighting of 1 for each point, the expressions simplify to the more familiar

$$a = \frac{\sum y_i - b\sum x_i}{n} = \overline{Y} - b\overline{x}$$

$$b = \frac{n\sum x_i y_i - \sum y_i \sum x_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

The weighted linear regression formulae are more appropriate for the purposes of the work presented here; therefore equations 6 and 7 are used.

Thus we have simple and efficient expressions for solving the weighted linear regression problem. In practice, we can efficiently compute

$$S_w = \sum w_i$$

$$S_y = \sum w_i y_i$$

$$S_x = \sum w_i x_i$$

$$S_{xy} = \sum w_i x_i y_i$$

$$S_{x^2} = \sum w_i x_i^2$$

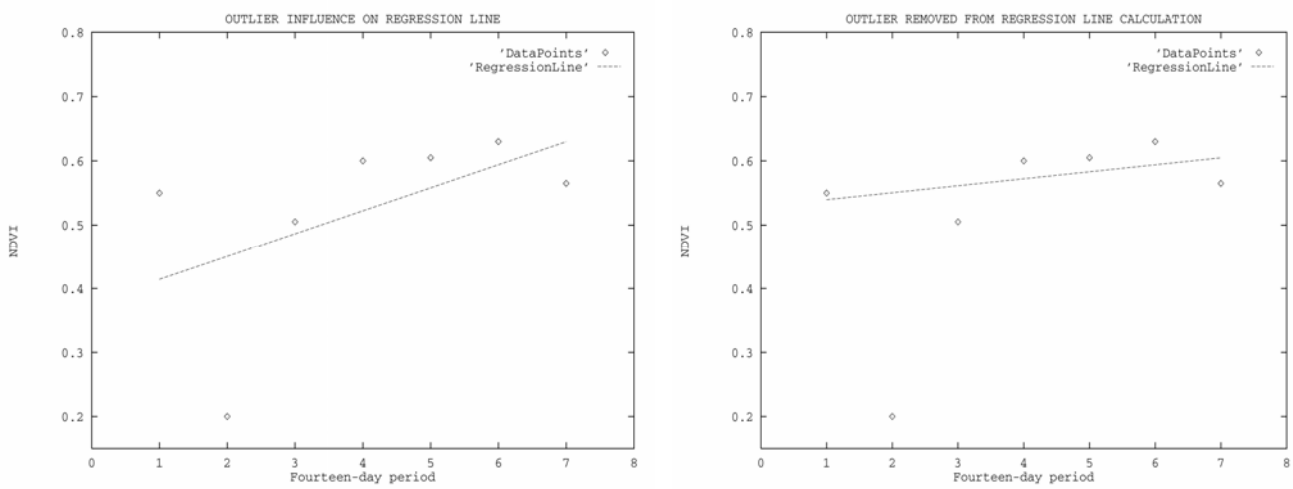
Then

$$b = \frac{S_w S_{xy} - S_x S_y}{S_w S_{x^2} - (S_x)^2} \tag{8}$$

$$a = \frac{S_y - bS_x}{S_w} \tag{9}$$

Outlier Detection

As with any data collection and analysis task, occasionally the collected data will be faulty in some way, usually through noise in the collection system. An *outlier* is an erroneous point that has an undue influence on the regression line. An example series of points containing an outlier is shown in Figure 2.



(a) (b)

Figure 2. (a) A sample series of simulated data points containing an outlier (point 2). The influence of the outlier point has a great effect on the calculated regression line. (b) The same series with the influential point removed before the regression line calculation.

The regression line obtained when the outlier is included in the calculation does not adequately reflect the real trend of the data points. When the regression line is calculated without the outlier point contributing to the calculation, the regression line more closely reflects the actual trend of the data.

Determining when a point is an outlier has received considerable attention (Rousseeuw 1987). The contaminating factors of the NDVI signal tend to lower the index's value in the y direction (Defries 1995); this particular application has no points that will be unduly influential in the x direction.

The linear regression described above has a statistical basis in that the expected value of the sample data lies along the regression line $y = a + bx$. If the nature of the noise corrupting the signal along the regression line is normally distributed, then the sum of the weighted squared errors given in equation 1 is χ^2 (Larsen 1986). Therefore, we can use a simple outlier detection technique such as the χ^2 test. In this test, we determine the existence of an

outlier using the calculated SWSE, comparing it to the χ^2 with $n-2$ degrees of freedom (for n points contributing to the linear approximation) to see whether any of the temporal NDVI points are outliers in the y direction. If it is determined that an outlier exists, we can examine each residual error to determine which point(s) are likely candidates for outliers. We can then mask these outliers from a recalculation of the regression curve.

Windowed regression

The preceding analysis assumes, of course, that a linear relationship exists between the independent variable x and the dependent observations y_i. Many relationships are not linear, however, such as the case with vegetation index derived from remotely sensed satellite imagery. However, many relationships that are not entirely linear are still approximately linear over some interval of interest. In fact any arbitrarily complex curve can be approximated by a piecewise linear curve through some window of a particular size. Using non-overlapping windows will produce a piecewise linear curve.

In this work, however, a smooth curve is required rather than a piecewise linear one, as shown in Figure 3.

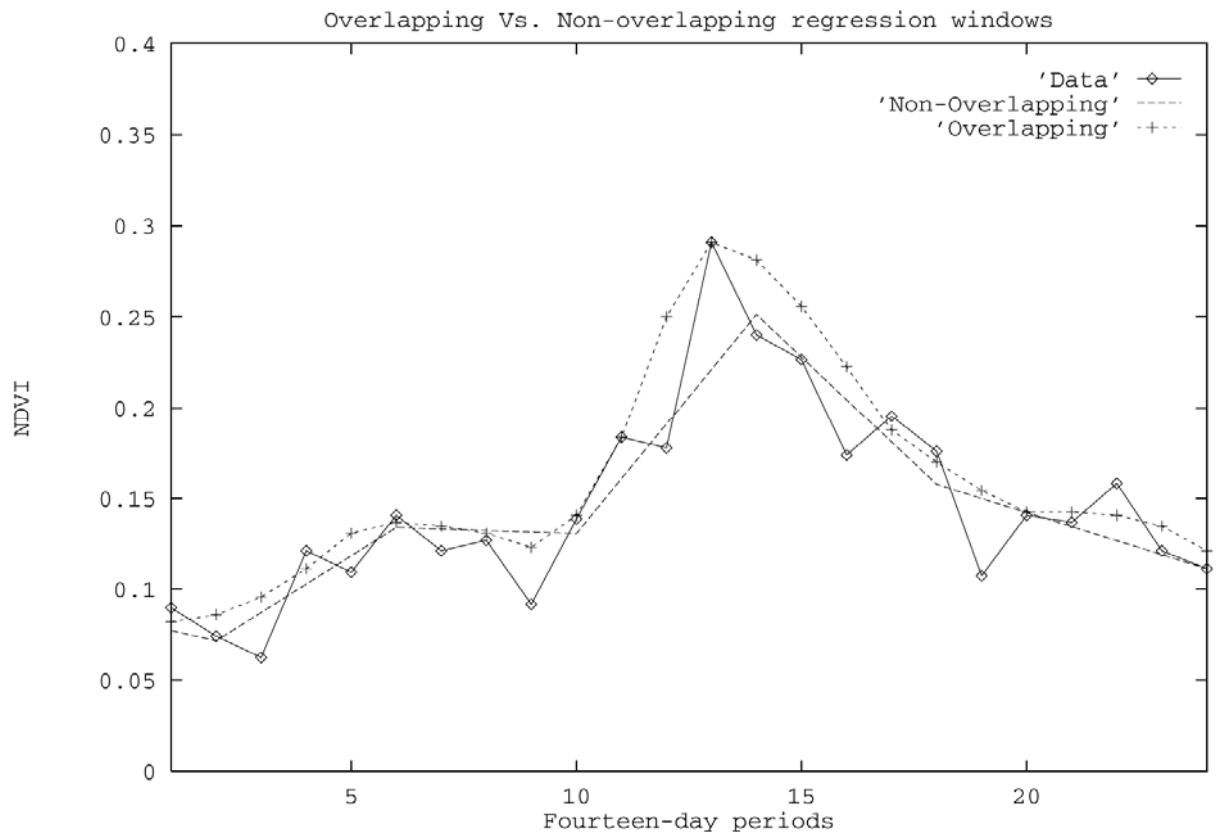


Figure 3. The difference between the piecewise linear curve obtained using non-overlapping windows for the linear regression and the smooth curve obtained through overlapping windows.

Therefore, the noisy signal is divided into a collection of overlapping windows. By making each of these overlapping windows small enough, our assumption that the signal in each of these overlapping windows is approximately linear is valid. By combining overlapping windows, we generate a smooth curve rather than a piecewise linear curve.

Using overlapping windows, however, poses the problem of how the resulting regression lines should be combined. There are a number of options that could be used for the combination algorithm. Since the contaminating factors of the NDVI signal tend to lower the index's value in the y direction (Defries 1995), for each point we could use the regression line that produces the maximum y value for a given x position. We found that this approach tended to yield widely variable results. The y values produced in this manner often were significantly higher than even the season peak point from the actual data. While we want to bias the smoothed curve upward in order to counteract the contaminating agents, we do not want to overcompensate and produce NDVI results that exceed the season's peak point.

Alternatively, we could compute a different regression line for each point, saving the line only where the point under consideration corresponds to the midpoint of the regression window. This approach produced curves that closely adhered to the original data sample but were still noisy enough to require further smoothing.

To facilitate the smoothing required by keeping each point's regression line, we took a combination of the lines for each point in a window. We defined a *combination window* in addition to the regression window and produced a new set of linear coefficients as a combination of the coefficients associated with each line in the combination window. For this combination, we tried using a median, a simple average, and an average weighted by a Gaussian kernel. Of each of these techniques, we found that the simple average produced a curve that most closely matched the original data while producing a nicely smoothed curve.

Weighting

Since the contaminating factors (noise) in the NDVI signal tend to lower rather than raise the y-value, we favored the high y-values over their lower-valued counterparts. In a three-point neighborhood, we can view each point as a local peak, a local valley, or a local sloping point, as shown in Figure 4.

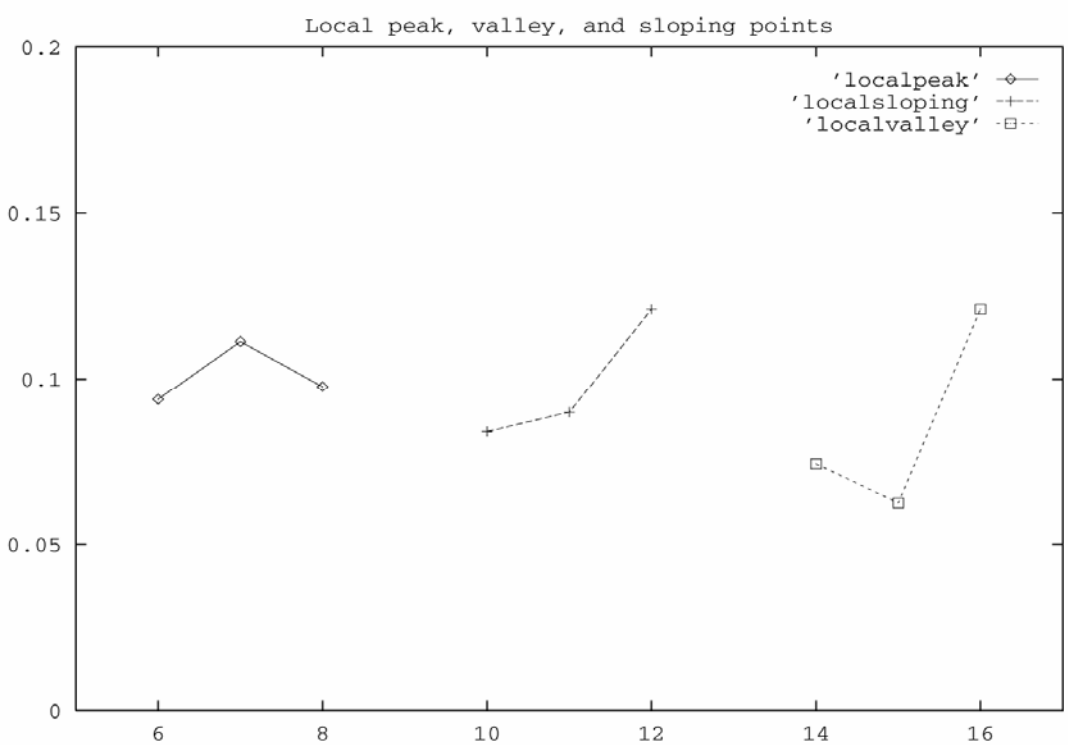


Figure 4. Local peak, sloping points, and local valley points in a three-point neighborhood.

The local peak points are likely closer to the actual data values than are the valley points, and for our application, we would like to interpolate rather than approximate these local peak points. Therefore, we give them a relative weighting much higher than either the sloping points or the valley points. The local valley points, on the other hand, are most likely the result of some corrupting factor in the data rather than an actual physical phenomenon. Therefore, we give these local valley points a relative weighting much lower than the other cases. The sloping points are not as important as the local peak points, but are much more important than the local valley points. Our analysis has led us to give the local peak points a weighting of 1.5, the local sloping points 0.5, and the local valley points 0.005. The effect of this weighting scheme makes us favor the local peak points while effectively discounting the contribution of the local valley points.

This manipulation of the weighting is done to fit our particular application's parameters. But we are only modifying the weights on the series of points in the signal; the underlying regression technique is still firmly planted in rigorous statistical methods. We can still indicate a confidence value for how well the smoothed curve represents the raw data; we can still locate outliers using our χ^2 test.

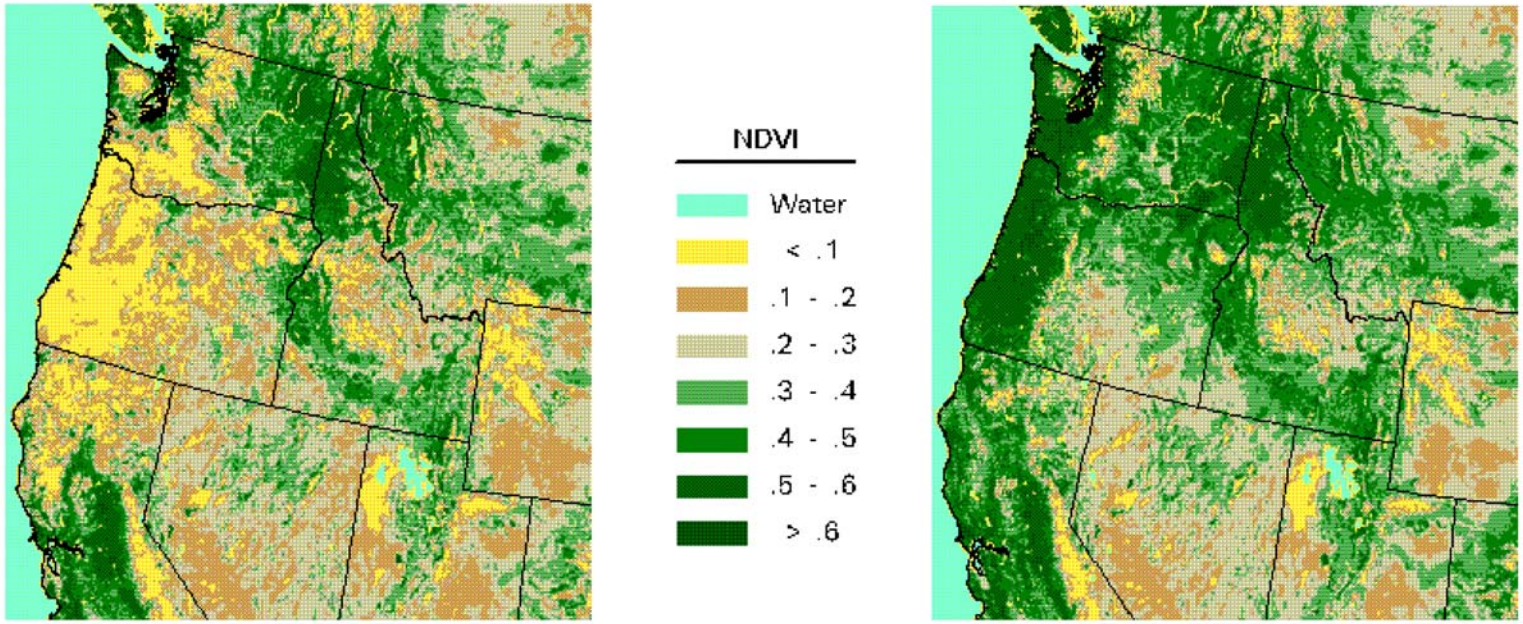
Figure 5 illustrates the results of implementing the smoother on an NDVI time series in May 1998. Note the large number of low NDVI values (depicted in yellow) that occur in western Washington, Oregon, and northern California due to atmospheric perturbations during this 14-day period. The smoother effectively replaces these low values with higher NDVI that is more typical of the evergreen coniferous forests found in this area.

APPLICATIONS

Once the NDVI time series has been smoothed, the data used in many applications are less affected by atmospheric and sensor perturbations. Such applications include land cover classification, seasonal characterization, and vegetation monitoring. Land cover classifications based on unsmoothed, multitemporal satellite data have been performed for some time (Loveland 1991). Operating on the smoothed data may reduce the number of pixels that are misclassified because of one or more noisy values. Seasonal characterization includes the derivation of seasonal metrics such as onset and duration of the growing season (Reed 1994), (Defries 1994). Such characterizations are usually based on identifying trend changes in NDVI values. Reducing the number of spuriously low NDVI values caused by atmospheric contamination reduces the probability of identifying spurious starts of season. Finally, since this technique allows smoothing data in near real-time, the NDVI data may be used in an improved monitoring system. Even if there are contaminated data, the smoother is able to estimate what the true value would be on the basis of previous observations.

SUMMARY

We have developed a weighted least-squares approach to smoothing multitemporal NDVI data. The result is an efficient, statistically based smoother that can be implemented in real-time, allowing refinement to particular smoothing applications by means of the regression and combination window parameters. The smoother greatly reduces the effects of atmospheric and sensor perturbations that generally tend to lower NDVI values. The resulting smoothed NDVI time series can be used to more effectively map land cover, identify phenological trends, and monitor vegetation, in particular through the development of seasonal metrics such as onset and duration of the growing season.



(a) (b)

Fig. 5. NDVI data from May, 1998 (a) and the results of data smoothing (b).

REFERENCES

Curran, P.J. (1981). Multispectral remote sensing for estimating biomass and productivity. In: *Plants and the Daylight Spectrum* (H. Smith, editor), Academic Press, London.

Defries, R., and J. Townshend (1994). Global land cover: comparison of ground-based data sets to classifications with AVHRR data. In: *Environmental Remote Sensing from Regional to Global Scales* (G. Foody and P. Curran, editors), Wiley, Chichester, pp. 84-110.

Defries, R., M. Hansen, and J. Townshend (1995). Global discrimination of land cover types from metrics derived from AVHRR pathfinder data. *Remote Sensing of the Environment*, *52*:209-222.

Goward, S.N., C. Tucker, and D. Dye (1985). North {American} vegetation patterns observed with the NOAA-7 advanced very high resolution radiometer. *Vegetation*, *64:*3-14.

Goward, S.N. (1989). Satellite bioclimatology. *Journal of Climate*, *2*:710-720.

Holben, B.N. (1986). Characteristics of maximum-value composite images from temporal AVHRR data. *The International Journal of Remote Sensing*, *7*:1417-1434.

Justice, C.O., J.R. Townshend, B. Holben, and C.J. Tucker (1985). Analysis of the phenology of global vegetation using meteorological satellite data. *International Journal of Remote Sensing*, *6*:1271-1318.

Larsen, R.J., and M.L. Marx (1986). *An Introduction to Mathematical Statistics and its Applications*, Prentice-Hall, Englewood Cliffs, New Jersey, second edition.

Los, S.O., C.O. Justice, and C.J. Tucker (1994). A global 1° by 1° NDVI data set for climate studies derived from the GIMMS continental NDVI data. *International Journal of Remote Sensing*, *15*:3493-3518.

Loveland, T.R., J.W. Merchant, D.O. Ohlen, and J.F. Brown (1991). Development of a land cover characteristics data base for the conterminous U.S. *Photogrammetric Engineering and Remote Sensing*, *57*(11):1453-1463.

Malingreau, J.P. (1989). The vegetation index and the study of vegetation dynamics. In: *Application of Remote Sensing to Agrometeorology*, (F. Toselli, editor), ECSC, Brussels and Luxembourg, pp. 285-303.

Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling (1988). *Numerical Recipes in C*, Cambridge University Press.

Reed, B.C., J.F. Brown, and D. VanderZee (1994). Measuring phenological variability from satellite imagery. *Journal of Vegetation Science*, *5*:703-714.

Rousseeuw, P.J., and A.M. Leroy (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, New York.

Townshend, J.R.G., C. Justice, and V. Kalb (1987). Characterization and classification of South American land cover types using satellite data. *International Journal of Remote Sensing*, *8*:127-150.

Tucker, C.J., J.R. Townshend, and T.E. Goff (1985). African land-cover classification using satellite data. *Science*, *227*:369-257.

VanDijk, A., S.L. Callis, C.M. Sakamoto, and W.L. Decker (1987). Smoothing vegetation index profiles: an alternative method for reducing radiometric disturbance in NOAA/AVHRR data. *Photogrammetric Engineering and Remote Sensing*, *53*(8):1059-1067.

Velleman, P., and D.C. Hoaglin (1981). *Applications, Basics and Computing of Exploratory Data Analyses.* Duxbury Press, Boston.

Viovy, N., O. Arino, and A.S. Belward (1992). The best index slope extraction (BISE): a method for reducing noise in NDVI time-series. *International Journal of Remote Sensing*, *13*(8):1585-1590.

Walpole, R.E., and R.H. Myers (1985). *Probability and Statistics for Engineers and Scientists*. MacMillan Publishing, New York, third edition.

Wilks, S.S. (1963). *Mathematical Statistics*, Wiley, New York.