MAPPING BROADACRE CROPPING PRACTICES USING MODIS TIME SERIES: HARNESSING THE DATA EXPLOSION

Peter Tan, Leo Lymburner, Medhavy Thankappan and Adam Lewis

National Earth Observation Group, Geoscience Australia, Canberra ACT 2601 Australia, peter.tan@ga.gov.au

KEYWORDS:

ABSTRACT:

The MODIS (or Moderate Resolution Imaging Spectroradiometer) 250 m EVI dataset provides a valuable ongoing means of characterising and monitoring changes in land use and resource condition. However the multiple factors that influence a time series of greenness data make the data difficult to analyse and interpret. Without prior knowledge, underlying models for time series in a given remote sensing image are often heterogeneous. So while conventional time series analysis methods such as wavelet transform and Fourier analysis may work well for part of the image, these models are either invalid or require to be substantially re-parameterised for other parts of the image. To overcome these challenges we propose a new approach to distil information from earth observation time series. The characteristic of a remote sensing time series are represented by a set of statistics (which we call coefficients) selected to correspond to the dynamics of a natural system. To ensure the coefficients are robust and generic, statistics are calculated independently by applying statistical models with less complexity on shorter segments within the time series. An International Standards Organization (ISO) Land Cover classification was generated for cropping regions in the Gwydir and Namoi catchments, in Australia. Areas identified in the classification as irrigated and rain fed cropping were analysed using a tailored time series analysis tool. The crop analysis tool identifies time series features such as the number and duration of fallow periods, crop timing, presence/absence of a crop during a year and the area under the curve (cumulative green biomass) for a specific growing season. This information is combined with paddock boundaries derived from Landsat imagery to provide detailed year-by-year insight into cropping practices in the Gwydir and Namoi catchments.

0. INTRODUCTION

1.1 Analysing Time Series of Remotely Sensed Imagery

Many current time series analysis methods used in remote sensing imagery analysis reconstruct a time series with one or a set of functions from a particular function class. Such methods often come with strong model assumptions and arbitrary parameters which must be manually specified. For example, autoregressive-moving average (ARMA) model assumes targeted time series are stationary (Emanuel 1982; Hamilton 1994), i.e., the behaviour (estimated parameters) of the time series do not shift dramatically along the time line. In many cases, such assumptions do not hold. In order to apply analysis methods such as wavelet transform (Percival and Walden 2006) and harmonic (Fourier) analysis (Roerink et al. 2000), a noisy and non-stationary remote sensing time series must be divided into a series of sub-time series in which model pre-requisites are satisfied. However, even with correctly pre-specified parameters, the temporal resolution of these sub-time series is often too limited for most conventional time series analysis.

The proposed remote sensing time series analysis method adopts a new strategy, which is inspired by following observations: a time series can be divided into a sequence of sub-time series with shorter lengths. Characteristics of a time series can be represented by a set of generic statistics extracted from these sub-time series. More sophisticated statistical methods with these generic statistics as input variables are more appropriate and less restricted to solve target problems. The method does not attempt to solve specific problems directly through tailored time series analysis algorithms. Instead, solutions for specific remote sensing applications are obtained in two stages. In the first stage, a set of statistics are extracted from many shorter sub-time series within the original time series. These statistics are generic, i.e., they are independent of model assumptions. No parameters need to be specified, therefore no prior knowledge has been assumed in the process. Then in the second stage, remote sensing scientists pursuing more specific targets are able to use these statistics as input features for sophisticated statistical analysis.

Advantages of the proposed method are: new types of statistics can be added to the record when new demands arise; statistics can be added to the record when new remote sensing imagery arrives; records can be stored in standard relational databases and relevant statistics can be retrieved using queries by various end users via web interfaces; relevant features for user specified targets can be obtained directly from the pre-calculated statistics or derived from them, re-usability and flexibility of the statistics are high; and statistics can be used as input features for sophisticated machine learning and statistical modelling of specific targets. The algorithm can be implemented in multi-threaded frameworks and be executed on high performance super computers or clustered servers. When the feature extraction procedure finishes, real-world problems can be solved by the proposed method, for example: clustering pixels in remote sensing imagery into homogeneous land cover classes, using various subsets of the coefficients; identifying bush fire events and the associated recovery period.

1.2 Time Series Analysis of Cropping Behaviour

MODIS Vegetation Index (VI) products are designed to provide consistent spatial temporal comparisons of vegetation conditions that can be used to monitor photosynthetic activity (Heute et al, 2002). Wardlow and Egbert (2005) demonstrated a scheme that uses MODIS 250m time series data to generate regional-scale crop mapping in the U.S. Central Great Plains. They concluded that the MODIS time series based approach was a cost and time-efficient means for large scale mapping. Jakubauskas et al (2001) applied harmonic analysis to VI time-series data to characterise seasonal changes to agricultural land use in southwest Kansas. They demonstrated the benefits of applying harmonic analysis to time-series remote sensing data for identification of crop types and reducing data volumes. Potgeiter et al (2007) investigated multivariate methods to estimate crop area for wheat, barley, chickpea, and total winter cropped area for a cropping region in northeast Australia. They reported that all multi-temporal methods showed significant overall capability to estimate total winter crop area. Thankappan et al (2008) demonstrated the feasibility of using time-series MODIS VI data for determining winter crop area in north-western Victoria, Australia, and highlighted the broader applicability of harmonic analysis to monitor landscape change. Xiao et al (2005) developed a paddy rice mapping algorithm that uses a time series of three vegetation indices derived from MODIS data to map paddy rice fields in 13 provinces of southern China. Their results showed that the MODIS-based paddy rice mapping algorithm could be applied at large spatial scales to monitor paddy rice. MODIS VI time series data was used for this work based on results from studies reported above.

0. METHODOLOGY

MODIS Enhanced Vegetation Index (EVI) time-series data from 2000 to 2007 was used for our methodology. The time series analysis scheme is proposed as a generic toolkit for remote sensing time series analysis. The aim is to provide a quantitative assessment of various aspects of ground phenomenon through a robust and generic modelling process, which in turn captures statistics related to characters of corresponding ground phenomenon. The proposed scheme consists of two stages. In the first stage, the time series data are passed through two filters to remove noisy elements in the MODIS time series. In the second stage, a set of 12 coefficients are calculated.

2.1 Noise Removal

The noise removal process consists of two stages. In the first step, time series data pass through a spectral filter which removes data points with abnormal values. Such points are defined as points which satisfy both the following conditions.

- Have a very high or very low value($\frac{x-\mu}{\sigma} > 2.1$ or $\frac{x-\mu}{\sigma} < -2.1$), where u is the mean of the time series and σ is the standard deviation of the time series.
- In the middle of a sudden rise (rate of rise above 95th percentile) and a sudden drop (rate of rise below 5th percentile) in the time series or vice versa

After the first step, most of the noisy data points are filtered out. However, consecutive noisy data points presented in some time series could not be detected by the spectral filter. Studies (Green et al 1988) have found that the distribution of noise in remote sensing imagery display strong local patterns. Hence, a spatial filter is designed to detect noisy elements missed by the spectral filter. The spatial filter detect points satisfying one of following conditions

- A large amount of noisy points (>75%) present among the neighbours
- Have exceptional high (or low) values ($\frac{x-\mu}{\sigma} > 3.97$) compared to those of neighbours.

The values of thresholds are based on experimental results on training samples provided by remote sensing scientists.

2.2 Time Series Coefficients

Twelve time series coefficients were developed in collaboration with remote sensing scientists. The goal was to capture different aspects of the characteristics of a remote sensing time series in the coefficient set. Therefore, as shown in subsequent sections of this paper, the set provided sufficient information to distinguish most land cover and land use features in earth observation imagery.

2.2.1 Mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

It is defined as the statistical mean of the time series, where x_i are the values of time series and N is the size of the time series. It measures the average level of the time series signals over the long term.

2.2.2 Standard Deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N-1}}$$

It is defined as the statistical standard deviation of the time series. It measures the standard deviation of the time series signals over the long term.

2.2.3 Flatness:

Step 1: Sort the time series in ascending order such that $x_i \le x_j, \forall i < j$

Step 2: Conduct a one-dimensional nearest neighbour clustering on the sorted time series, i.e., find the index C to separate two clusters, such that

$$\arg\min_{C}\left(\sum_{i=1}^{C}(x_i-\mu_1)^2+\sum_{i=C+1}^{N}(x_i-\mu_2)^2\right)$$

,where $\mu_1 = \frac{1}{C}\sum_{i=1}^{C} x_i$ and $\mu_2 = \frac{1}{N-C}\sum_{i=C+1}^{N} x_i$

Step 3: Define the threshold $\delta = u_1 - k * \sigma_1$ where

$$\mu_1 = \frac{1}{C}\sum_{i=1}^{C} x_i \,,\ \sigma_1 = \sqrt{\frac{\sum_{i=1}^{C}(x_i-\mu_1)^2}{C-1}}$$

and k is a predefined constant.

Step 4: Find the set of sub-time series $\{S_i\}$ satisfying

$$|S_i| > L_m \text{ and } x > \delta, \forall\, x \in S_i$$

Where L_m is a predefined minimum length.

Step 5: Calculate the coefficients defined as the ratio of the sum of the lengths of such sub-series against the length of the whole time series

$$\frac{1}{N}\sum |S_i|$$

2.2.4 Rate of Rise:

Step 1: Define a set of sub time series

$$S = \{S_1, S_2, \ldots, S_M \mid |S_i| \in [L_{min}, L_{max}], i \in [1, M]\}$$

Step 2: Calculate the rate of change r_i for S_i. Let $L_i = |S_i|$, x_b and x_e are the first and the last elements in the sub time series respectively, then

$$r_i = e^{\frac{1}{L_i}\log\frac{x_e}{x_b}} - 1$$

Step 3: Find the non-overlapping subset $P \subset S$ with the maximum sum of r_i

$$\arg\max_P \sum_{r_i \in P} r_i \;\Big|\; x \neq y, \text{ given } x \in P_i\,,\, y \in P_j\ \forall\, i, j$$

Step 4: Calculate the coefficient by averaging the rate of change r_i in set P

$$\frac{1}{|P|}\sum_{r_i \in P} r_i$$

2.2.5 Rate of Drop:

The procedure to calculate the coefficient rate of drop is similar to the procedure described in the above section. However, this time we are interested in the low end of the distribution of r_i.

Step 1: Define a set of sub time series

$$S = \{S_1, S_2, \ldots, S_M \mid |S_i| \in [L_{min}, L_{max}], i \in [1, M]\}$$

Step 2: Calculate the rate of change r_i for S_i. Let $L_i = |S_i|$, x_b and x_e are the first and the last elements in the sub time series respectably, then

$$r_i = e^{\frac{1}{L_i}\log\frac{x_e}{x_b}} - 1$$

Step 3: Find the non-overlapping subset $P \subset S$ with the minimum sum of (negative) r_i

$$\arg\min_P \sum_{r_i \in P} r_i \;\Big|\; x \neq y, \text{ given } x \in P_i\,,\, y \in P_j \forall\, i, j$$

Step 4: Calculate the coefficient by averaging the rate of change r_i in set P

$$\frac{1}{|P|}\sum_{r_i \in P} r_i$$

2.2.6 Global minimum: Step 1: Sort the time series in ascending order so that $x_i \leq x_j, \forall i < j$

Step 2: Calculate the coefficient by averaging the first M elements of the sorted time series, M is a predefined constant (in our implementation, M takes the value of the number of calendar year in the time series)

$$\frac{1}{M}\sum_{i=1}^{M} x_i$$

2.2.7 Average length of cycle: Step 1: Define a set of sub time series

$$S = \{S_1, S_2, \ldots, S_M \mid |S_i| \in [L_{min}, L_{max}], i \in [1, M]\}$$

that satisfy the following conditions

$$x_m > \mu + \sigma, x_b < \delta \text{ and } x_e < \delta,\ \forall S_i$$

where x_m is the maximum, x_b is the first and x_e is the last element in sub time series S_i. δ is the threshold obtained from step 3 for calculating Flatness.

Step 2: Calculate the coefficient by averaging the length of the sub time series S_i in the set P

$$\frac{1}{|P|}\sum_{S_i \in P} |S_i|$$

2.2.8 Global maximum: Step 1: Sort the time series in ascending order so that $x_i \leq x_j, \forall i < j$

Step 2: Calculate the coefficient by averaging the last M elements of the sorted time series, M is a predefined constant (in our implementation, M is the value of the number of calendar year in the time series)

$$\frac{1}{M}\sum_{i=1}^{M} x_i$$

2.2.9 Ratio of the Global Maximum to the Annual Maximum

The Annual maximum is calculated as the mean of the maximum in each calendar year of the time series. Then the coefficient is the ratio of annual maximum against the Global maximum.

2.2.10 Mean timing of the maximum: Assuming that time series is observed in a regular base, calculate the coefficient by averaging the timing (index) of the maximum element in each calendar year of the time series

2.2.11 Standard deviation in the timing of the maximum: Assuming that time series is observed in a regular base, calculate the coefficient as the standard deviation of the timing (index) of the maximum element in each calendar year of the time series

2.2.12 Annual minimum: The Annual minimum is calculated as the mean of the minimum in each calendar year of the time series.

2.3 Study area: The study area covers the Gwydir and Namoi catchments in north western New South Wales, Australia. There is a mix of land uses including irrigated agriculture (predominantly cotton), broadacre agriculture (mixture of oil-seed, hard wheat, durum wheat, sorghum and pulses), grazing of native and improved pastures (beef, wool and lambs), and reserves of native vegetation (Scott *et al.* 2004). The native vegetation ranges for open woodland towards the western edge of study area, and becomes increasingly dense towards the eastern edge, with pockets of closed forest on the slopes of Mt Kaputar. Irrigation flows are supplied from dams on the Gwydir and Namoi rivers. There is also a strong rainfall gradient from east to west, with areas in the east receiving more, and more consistent rainfall (~700mm per annum), whereas areas in the west receive less rainfall and the rainfall is less reliable (~450mm per annum) (Scott *et al.* 2004).

2.4 Generating a Classification using the Time Series Coefficients

A combination of aerial survey data and field survey data were used to identify different land cover types within the Gwydir catchment. Over 1000 polygons were identified and assigned a land cover type. This dataset was then divided at random on a 2/3rd 1/3rd basis into separate datasets. The 2/3rd portion was used to seed the classification algorithm and the 1/3rd was used as an independent dataset to evaluate the classification accuracy. The time series coefficients were classified using Definiens Developer 7 ™. The error assessment matrix for the classification is detailed in **Error! Reference source not found.**. The critical feature of **Error! Reference source not found.** is that the classification is effective in separating the cropping from the non cropping regions, with some limited confusion with improved pasture. The overall accuracy of classification was 69%. The irrigated and dry land crop classes were used to identify pixels that were dominated by these land cover types, and the time series of these pixels were interrogated using a tailored time series analysis module as described below.

	Irrigated Crop	**Dryland Crop**	**Improved Pasture**	**Native Pasture**	**Closed Forest**	**Open Forest**	**Wood land**	**Open Wood land**	**Dam**	**Wet land**	**Total**	**Users Accuracy**
Irrigated Crop	138	2	0	0	0	0	0	0	0	0	140	99%
Dryland Crop	18	73	0	4	0	0	0	0	0	0	95	77%
Improved Pasture	0	4	81	30	0	0	0	0	0	0	115	70%
Native Pasture	0	0	0	65	0	0	0	45	0	0	110	59%
Closed Forest	0	0	0	0	3	2	0	0	0	0	5	60%
Open Forest	0	0	0	0	1	4	5	2	0	0	12	33%
Woodland	0	3	0	10	0	1	5	20	0	0	39	13%
Open Woodland	0	1	3	25	0	0	3	20	0	2	54	37%
Dam	0	0	0	0	0	0	0	0	15	0	15	100%
Wetland	0	0	1	1	0	0	0	0	0	5	7	
Total	156	83	85	135	4	7	13	87	15	7	592	
Producers Accuracy	88%	88%	95%	48%	75%	0%	38%	23%	100 %	71%		
	Overall Accuracy			69%								

Table 1: Error Assessment Matrix for the MODIS Time Series Classification

2.5 Tailored Time Series Analysis Tools

2.5.1 Number of peaks: We use an approach similar to the one proposed in section 2.2.3 to find the number of peaks in the time series. The time series is sorted and then a one-dimensional nearest neighbour clustering is conducted. We define a 'growth period' as a sub-time series containing points from the cluster with higher mean and the start and the end point below the flatness threshold. The number of such growth periods is the number of peaks.

2.5.2 Length of fallow periods: The length of the fallow periods was calculated by applying the same algorithm to obtain the flatness coefficient (section 2.2.3).

2.5.3 Number of fallow periods: We count the number of fallow periods using a modified version of the algorithm described in section 2.2.3. Instead of summing up the length of the identified sub-time series, the number of such sub-time series was counted.

2.5.4 Area under the curve: The area under the curve is approximated by the sum of the time series. The start point of growth period is corresponding to the end point of a previous fallow period. The end point of a growth period is the start point of the next fallow period.

0. RESULTS

Figure 1 illustrates how the tailored time-series analysis parameters can be used to visualise changes across the landscape. There are a number of key features shown in Figure 1, the most obvious is the difference in cumulative greenness between the irrigated (predominantly red) and non irrigated (greens and purples) portions of the two catchments. The other feature is the gradient in cumulative greenness from east (right hand side) to west (left hand side) this gradient reflects the rainfall gradient across this region.

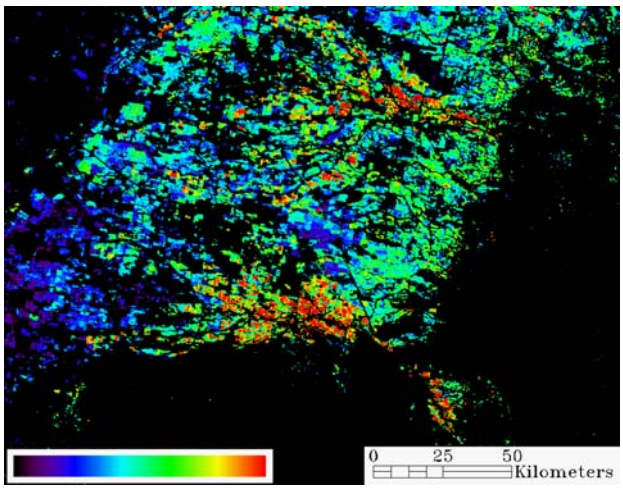

Figure 1. Cumulative area under the Curve

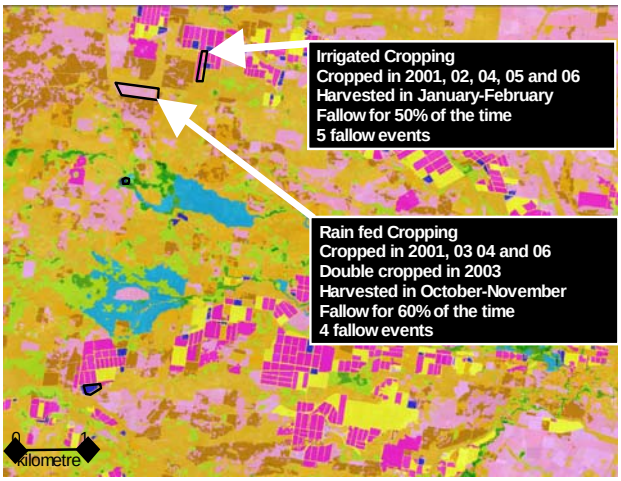

Figure 2. Cropping Practices Identified Using the Time Series

Similarly other features such as the number of crop cycles, crop cycles have been double cropped or the percentage of fallow can be represented. Comparisons between these features provide an understanding of long-term changes in the cropping practices.

0. DISCUSSION

One of the main advantages of this approach is that it provides consistent nationwide data that can be used to characterise cropping practices. This enables within-catchment and between region comparisons of cropping practices, and, when combined with appropriate ancillary data could be used to compare productivity and water use patterns within and between regions. The data generated by the crop analysis module also provides a framework for extrapolating existing crop monitoring programs. The capacity to apply this module to archival data means that it can be used to provide a range of input parameters for assessing relationship between large scale shifts in climate and the response of cropping regions to existing and emerging climate variability. The cropping practices information generated by the crop analysis module can be combined with rainfall surfaces and evapotranspiration models to identify and characterise regions that are being subject to irrigation, and therefore be used to assess changes in irrigation practices as a response to altered water availability. This information could potentially be combined with the right ancillary data to monitor land use practices such as the use of green manure crops. The data can also be combined with object oriented analysis of 25 metre data to provide a field-by-field assessment of crop practices in areas where the fields are large enough to contain multiple MODIS pixels as shown in Figure 2.

The 250 m resolution limits the application to large scale dry land cropping, not suitable in areas where fields are 300 m x 300 m or smaller, although there is some research into applications at this resolution (Xiao et al. 2007). An alternate solution is to characterise areas with small paddock sizes using Landsat. Another limitation is that because EVI only characterises greenness it does not capture non-green-fraction dynamics i.e. it does not measure and is not sensitive to the Non-Photosynthetic Vegetation (NPV) and bare soil fractions. To address this issue it is necessary to use multi-spectral data that includes short wave infra red bands to characterise the bare soil-NPV fractions, both MODIS (500 m resolution) and Landsat (25 m resolution) can be used for this purpose.

CONCLUSION

We have proposed a novel statistical method for remote sensing time series analysis. Compared to conventional approaches, the proposed methods contain no assumptions about the nature of the time series. Therefore they provide a more robust basis for modelling remote sensing time series at national scale. The time series coefficients generated using this new technique proved suitable for generating a land cover classification of acceptable accuracy (69% over all accuracy). To show case the effectiveness of the new method, a toolkit tailored for mapping broadacre cropping practices was developed. This technique provided results that are consistent with rainfall gradients and irrigation regimes that are known to exist within the study area, and provide valuable insight into the cropping practices within the study area on a year by year basis. These results can also be combined with higher resolution GIS data to provide a field-by-field assessment of crop cycles, provided that the fields are of an appropriate size.

REFERENCES

Emanuel, P. (1982). ARARMA models for time series analysis and forecasting. *Journal of Forecasting **1**(1): 67-82.*

Green, A. A., Berman, M.; Switzer, P., Craig, M. D. (1988). A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing **26**(1): 65-74.*

Hamilton, J. D. (1994). Time series analysis, Princeton University Press.

Heute, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X. and Ferreira, L.G. (2002) Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment 83 195-213*

Jakubauskas, M. E., Legates, D. R. and Kastens, J. H. (2001). Harmonic analysis of time-series AVHRR NDVI data. *Photogrammetric Engineering and Remote Sensing*, *67, pp. 461-470*

Potgeiter, A.B., Apan, A., Dunn, P. and Hammer, G. (2007) Estimating crop area using seasonal time series of Enhanced Vegetation Index from MODIS satellite imagery *Australian Journal of Agricultural Research* 58(4) 316–325.

Roerink, G. J., Menenti, M. and Verhoef, W. (2000) Reconstructing cloud free NDVI composites using Fourier analysis of time series *International Journal of Remote Sensing **21**(9): 1911-1917.*

Scott, J. F., Farquharson, R. J. and Mullen, J.D. (2004) Farming Systems in the Northern Cropping Regions of NSW: *An Economic Analysis Economic Research Report No. 20 NSW Department of Primary Industries, Tamworth.*

Thankappan, M., Lawson K., Reddy, R., and Kokic, P. (2008) Harmonic analysis of timeseries MODIS vegetation index data for monitoring winter crops in the Wimmera-Mallee region of Victoria. *Proceedings of the 14th Australasian Remote Sensing and Photogrammetry Conference (ARSPC), Darwin, NT, Australia.*

Percival, D. B. and Walden, A. T. (2006) Wavelet Methods for Time Series Analysis, Cambridge University Press.

Wardlow, B. D. and Egbert, S. L. (2005) State-level crop mapping in the U.S. Central Great Plains agroecosystem using MODIS 250-metre NDVI data. *Pecora 16 "Global priorities in Land Remote Sensing", Sioux Falls, South Dakota, October 23-27*

Xiao, X., Boles, S., Liu, J., Zhuang, D., Frolking, S., Li, C., Salas, W. and Moore, B. III (2005). Mapping paddy rice agriculture in southern China using multi-temporal MODIS images, *Remote Sensing of Environment, 95(4):480-492.*