Temporal Modeling and Missing Data Estimation for MODIS Vegetation data

Rie Honda*

1 Introduction

The Moderate Resolution Imaging Spectroradiometer (MODIS) is the primary instrument on board NASA's Earth Observing Satellite Terra and Aqua. It produces observations covering the whole Earth's surface, acquiring data in 36 spectral bands with wavelengths ranging from 0.4 μm to 14.4 μm. MODIS data are used to compute vegetation indices (VI) that reflect the level of activity of vegetation on the ground.

Understanding spatio-temporal VI patterns helps constructing accurate dynamical models of vegetation, from regional to global scales. In turn this is a stepping-stone in understanding phenological dynamics of the terrestrial ecosystem. Zhang et al.(2003) [1] presented a framework for modeling VI time series using piecewise logistic functions. However, their method depends on the heuristic process, because the raw measurement data is quite noisy, and a relatively large fraction of the observations are in effect"missing" during the colder months, due to the fact that accurate EVI values cannot be computed in many locations because of snow cover.

The primary objective of this study is to propose a general statistical framework for estimating the parameters for the type of model proposed by Zhang et al. (2003). Missing data and noises are treated in a more sound manner in our method via maximum a posteriori (MAP) approach.

The secondary objective is to apply the above method to spatio-temporal pattern mining. We propose the method of missing value estimation that combines temporal estimates by MAP and other information such as spatially neighbouring data by random forests regression.

The performance of both methods are examined by the experiments using the annual MODIS VI data of Northeastern United States.

2 Temporal Estimation based on the Probabilistic Model

Figure 1 shows the typical pattern of annual change of MODIS EVI data in the northeastern United States. Each data point corresponds to a set of EVI values averaged over a 16-day period, at this particular location. Land cover type (so-called IGBP index) determined by the spectrum measurement are also provided.

The vegetation phenology follows a temporal pattern consisting of four phases: a stable period with a low level of EVI (Dormancy), a period of rapid increase in EVI, a stable period of high levels of EVI (Senescence), and a period of rapidly decreasing EVI. The general shapes of rapid increase and decrease in the observed time series seem to be well approximated by two piecewise logistic functions, as proposed by Zhang et al. (2003).

*Kochi University, Akebono-cyo 2-5-1, Kochi, Japan.

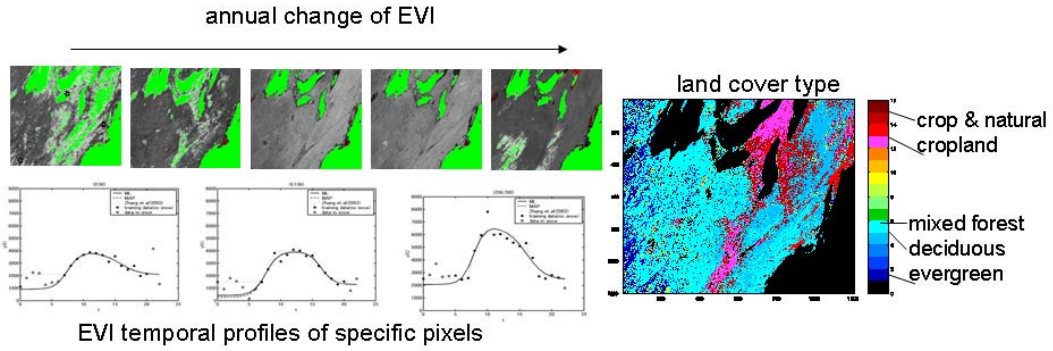

Figure 1: Example of typical time series data and fitted models for MODIS EVI together with land cover type indices. Green pixels in the upper left images indicate missing data.

$$F(t_i|\theta) = \begin{cases} f_1(t_i|\theta), & t < t_c, \\ f_2(t_i|\theta), & t \geq t_c, \end{cases} \tag{2.1}$$

$$f_j(t_i) = \frac{c_j}{1 + \exp(a_j + b_j t_i)} + d_i \ , \ j = 1, 2, \tag{2.2}$$

where t_i is the observation time of i-th data, θ is the model parameter set $\{\theta_j | i = 1, 2, \cdots, k\}$ (i.e. $\{a_j, b_j, c_j, d_j | j = 1, 2\}$), and t_c is the "crossover point" of $f_1(t|\theta)$ and $f_2(t|\theta)$.

Suppose the time series $D = \{y(t_1), y(t_2), \cdots, y(t_n)\}$ is obtained from observation (e.g., at a single pixel) and we wish to fit equations in the form of Equations 2.1 and 2.2 to this data. We also assume that the set of times $t_1, \ldots, t_n$ are known.

A common approach in regression modeling is to assume that the observations were generated from a deterministic functional form, but with additive measurement noise. Gaussian measurement noise with fixed (unknown) variance is assumed. In this manner we can write down the probability to observe D for a given set of θ as:

$$P(D|\theta) = \prod_{i=1}^{n} N\left(F\left(t_i|\theta\right), \sigma\right), \tag{2.3}$$

where N denotes a Normal (Gaussian) density function, with mean value $F\left(t|\theta\right)$ and where σ is the standard deviation.

In what follows, we adopt Maximum a Posteriori (MAP) methods for parameter estimation in stead of commonly used Maximum likelihood estimation (ML) [2][3]. In the Bayesian approach, θ is viewed as a random variable, and we are uncertain about the distribution of θ. Our prior uncertainty is reflected in a so-called prior distribution or density, $p(\theta)$. Our posterior uncertainty after seeing some observed data D is reflected in the posterior $p(\theta|D)$. We can connect the two expressions via Bayes rule, i.e., $p(\theta|D) = p(D|\theta)p(\theta)/p(D)$, where the first term in the numerator on the right hand side is the likelihood $p(D|\theta)$.

Here we assume that each parameter θ_j in the parameter vector has a prior probability that is a Gaussian density function, and that these priors are independent of each parameter θ_j:

$$P(\theta) = \prod_{j=1}^{k} N(\mu_{\theta_j}, \tau_{\theta_j}), \tag{2.4}$$

where μ_{θ_j} is the mean of the prior for parameter θ_j, and τ_{θ_j} is the standard deviation of θ_j.

"Fully Bayesian" estimation consists of computing $p(\theta|D)$, but from a practical viewpoint it is often sufficient to have "point estimates" of the θ values, in particular the most likely value of θ given the data is obtained by:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(D|\theta)p(\theta).$$

This is known as *maximum a posteriori* (or MAP) estimation. When a large fraction of data are missing, MAP tries to find the solution based on the prior. Thus MAP is expected to work robustly even for the data set including numerous missing data. This equation is numerically solved by a Newton-Raphson method modified to solve a model containing two functions.

3 Combination of Spatial and Temporal Estimation by Random Forests

Random Forests [4] is a supervised learning method which learns the classifier or predictor composed of multiple trees from the set of target values x (could be a label or a value) of data cases and their attributes. Once the trees learn the relation between the attributes and the target value, they can predict a target value for a particular unknown data case from its attributes. In what follows. we adopt this method to combine temporal model and spatial neighboring data and the other information such as land cover type.

We assume the target value is the EVI at the pixel coordinate (k, l) and time t_i, denoted by $y_{k,l}(t_i)$. The attributes of the input data consist of EVIs of 8 spatial neighboring pixels, the land cover type (called as IGBP index) of both the center and the 8 neighbor pixels, temporal prediction value of $F(t_i|\theta_{k,l})$ calculated in MAP approach.

4 Experimental Results

We applied the above method to the real MODIS data of Northeastern United States to examine the performance of missing data estimation by random forests and MAP estimation.

The data set is composed of both NBAR EVI, "snow flag", and land cover type (IGBP indices) for pixels in the Northeastern US. The data set ranges from January to December in 2001 with a temporal resolution of 16 days (thus, 23 time values for an annual cycle) and a spatial resolution of 1km per pixel. There is a time-series for each of 1200 $\times$ 1200 pixels. We used the snow flag index as an indicator for missing data.

Evaluations are conducted for the six data sets, each of which consists of the data of evergreen forests, deciduous forests, mixed forests, cropland, cropland and natural vegetation mosaics, and the mixture of five groups, respectively. About 1000 samples are obtained for each land cover type. The half of the data set was used for training of random forests and the rest was used for the test of prediction accuracy.

To compare the prediction accuracy objectively, we conducted the experiments on four prediction methods: (1) logistic model fitting via MAP (temporal prediction), (2) simple four-

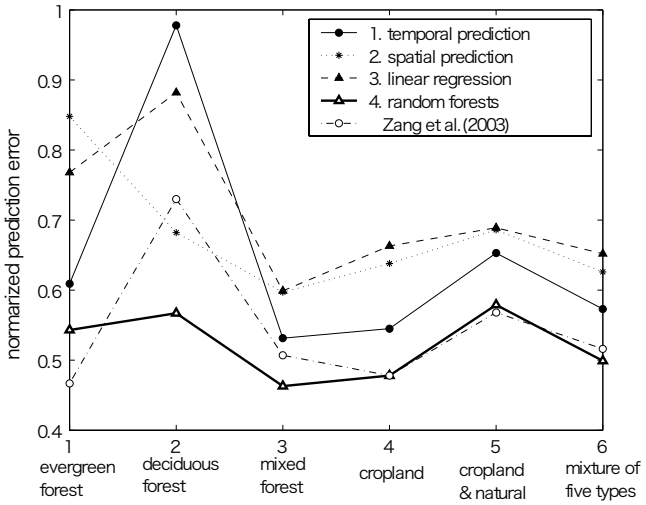

Figure 2: The normalized prediction errors for each land cover type and each method.

neighbor average (spatial prediction),and (3) linear regression by using all the attributes, and (4) spatio-temporal prediction by random forests.

We defined $\bar{e}_{rms}$ as the root mean square error normalized with the root mean square error against the mean predictor, and compared $\bar{e}_{rms}$ of each data set and each prediction method. Figure 2 shows $\bar{e}_{rms}$ for each land cover type group and each estimation method. The fitting error of Zhang et al. (2003) for the same data set is also plotted for reference (please note Zhang et al. (2003)'s result is not a prediction error, so just for reference). Temporal prediction by MAP is more accurate than any of spatial average and linear regression, although it creates less accurate estimates for deciduous forest group.

The result also indicates random forest creates the most accurate predictor among all methods and for all land cover groups. The improvement of the prediction errors ranges from about 4 % up to 32 %. Once the random forests predictor is created, missing values are estimated in quite a short time. This implies random forests regression is easily included in the improvement of missing data estimates in a spatio-temporal data set, and it would be utilized in the process such as renewal of spatio-temporal model via iteration.

5 Conclusion and Future Directions

The effectiveness of the statistical approach (MAP) for temporal modelling and the improvement of the temporal estimates with the other spatial information through random forests was presented. From this framework, we anticipate to generate more general and complex models, e.g., estimation for data spanning multiple years. models that can estimate parameters at multiple pixels simultaneously and that can leverage spatial correlation.

Acknowledgements

The author deeply appreciates Prof. Padhraic Smyth, Prof. Mark Friedl and Dr. Xiaoyang Zhang for for their suggestion from the very early stage of this study.

References

[1] X. Zhang, M. A. Friedl, C. B. Schaaf, A. L. Strahler, J. C. F. Hodges, F. Gao, B. C. Reed, A. Heute, *Monitoring vegetation phenology using MODIS*, Remote Sensing of Environment, 84 (2003), pp. 471–475.

[2] L. Wasserman, *All of Statistics - A Concise course in Statistical Inference*, Springer-verlag, New York, 2004.

[3] R. O. Duda and P. E. Hart, D. G. Stork *Pattern Classification* , Willey-interscience (2000).

[4] L. Breiman, *Random Forests* , Machine Learning, vol. 45 no. 1 (2001) pp. 5-32.