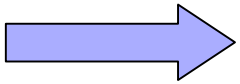
Temporal Modeling and Missing Data Estimation from MODIS Vegetation Data

Rie Honda
Kochi University, Japan
Second NASA Data Mining Workshop
May 23-24, 2006, Pasadena, CA

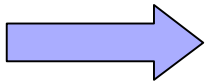
Backgrounds

- Spatio-temporal data mining from Earth observation satellite data is a powerful approach for Phenology study, and the other application field such as detection of famine, deforestation.
- The common obstacles are, noises, missing data, and the sparseness of the data
- How we can build a temporal model accurately for such dataset?



Model fitting by Maximum a posterior (MAP) approach

- How we can utilize spatial information to improve the accuracy of temporal modeling?



Integration of MAP estimate and spatial information by Random Forests

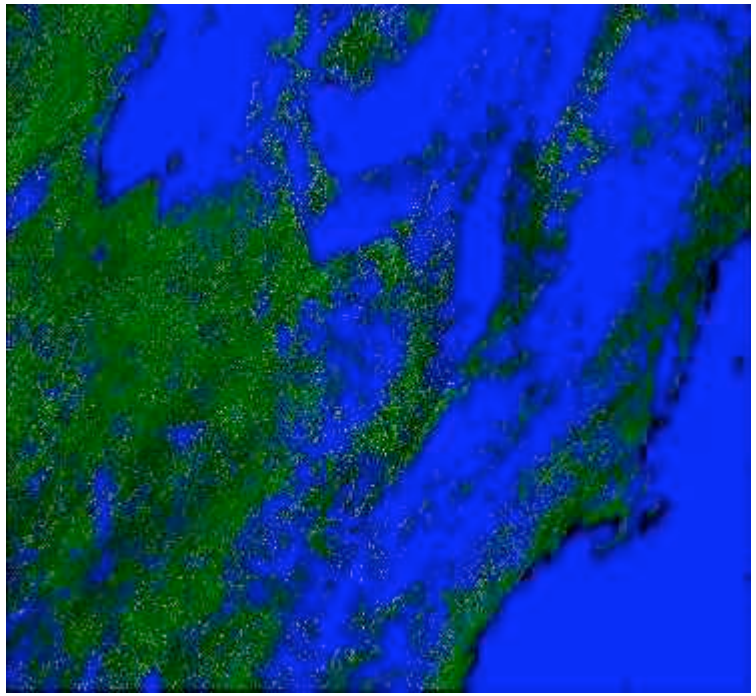
Outline

- Dataset.
- Temporal modeling of MODIS EVI by MAP.
- Missing value estimation by Random Forests by using both MAP estimate and the other and spatial information as the input.
- Conclusion.

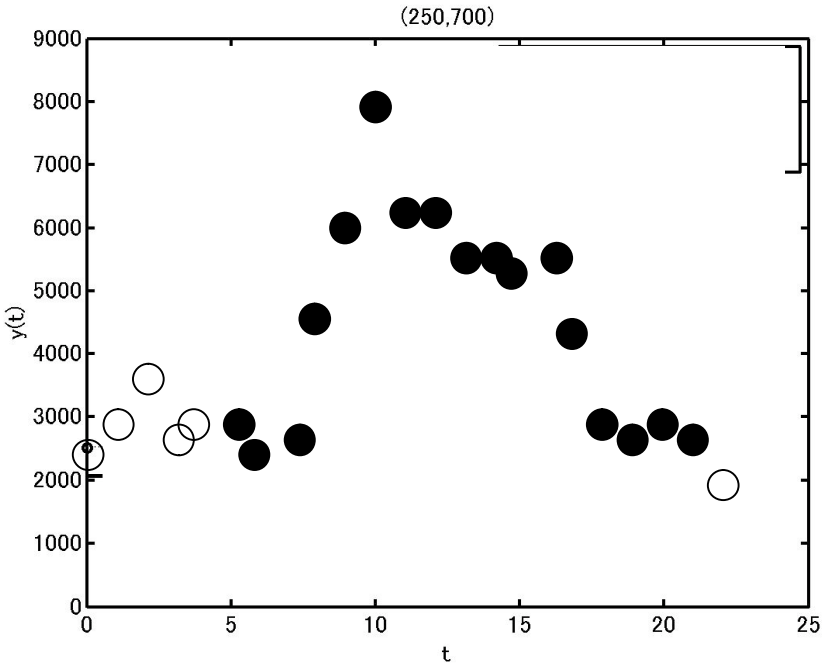
Dataset

- The enhanced vegetation index (EVI) of Moderate Resolution Imaging Spectroradiometer (MODIS)
  - Produced globally over land at 500m/1km resolutions and 16-day compositing periods.
  - Optimized index with accurate calibration of atmosphere influences and effect of canopy, etc.
  - Useful for study of global phenology, ecosystem, etc.

Example of MODIS EVI Data



N



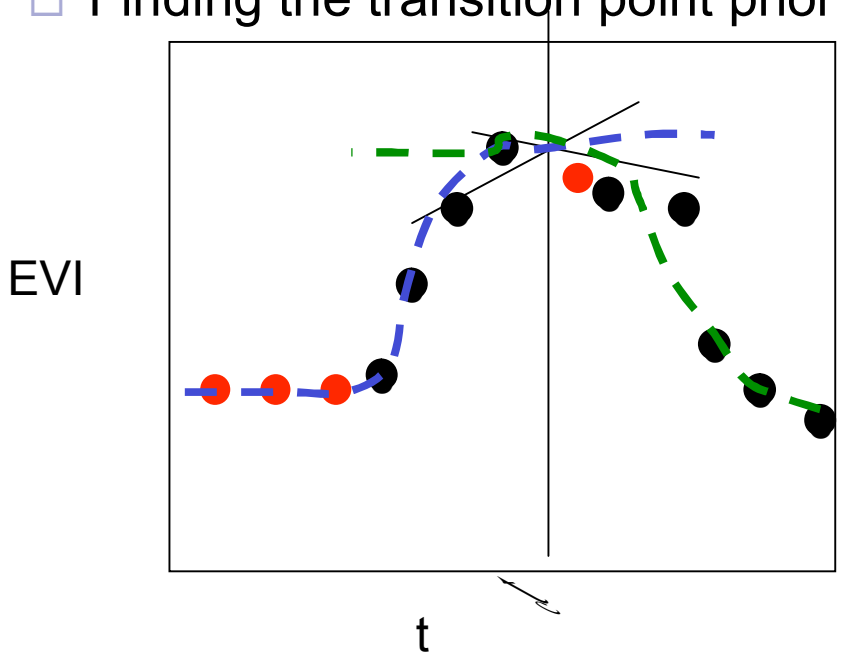
valid data
data in snow

- Problems
  - Numerous data are missing due to cloud or snow coverage (particularly in the winter time).
  - Data are noisy and data points exist temporally sparsely.

Method used in the previous study

- Zhang et al. (2003) adopted the piecewise logistic function model and fitted the data to the model by least-square-error method.
- Preprocessing was required prior to the fitting
  - Filling missing values by the temporal neighbors
  - Finding the transition point prior to the fitting

These may cause a biased solution.



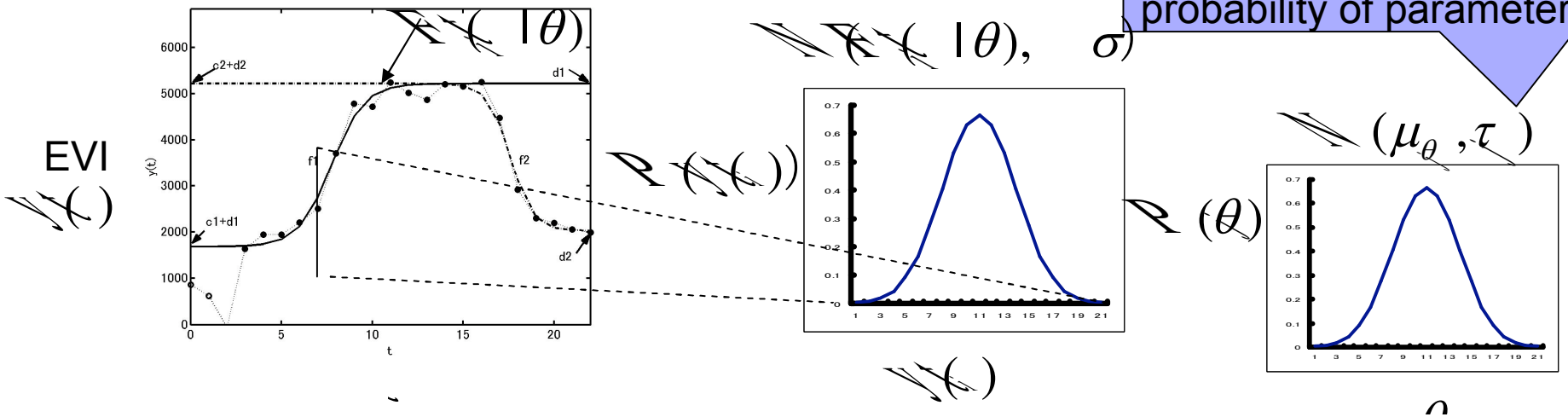
$$f(\ |\theta) = \begin{cases} f_1(\ |\theta_1) & \\ f_2(\ |\theta_2) & \end{cases}$$

$$f(\ |\theta) = \frac{}{1+}$$

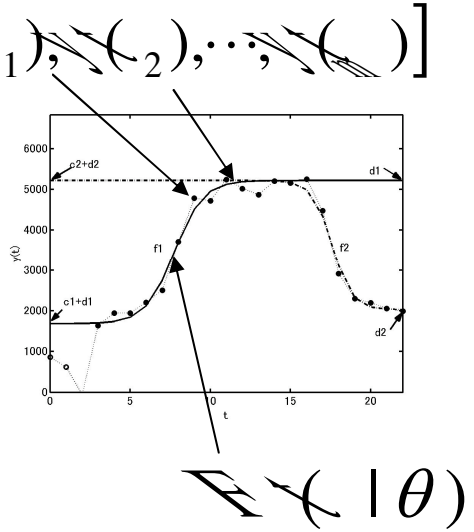
Features of temporal modeling in this study

- Following the Zhang et al's piecewise logistic function model, but statistically-sound method is used.
  - Maximum a Posterior (MAP) approach
  - Both probability distribution of observation including noise (σ), and the probability distribution of model parameter ((μ_θ, τ)) are incorporated in the model.

When the many of data are missing, MAP relies on the probability of parameters.



Method - Maximum a Posterior (MAP) Estimation -



- Equations.

$$P(\quad|\theta) = \prod \quad(\quad|\theta), \qquad \sigma),$$

$$(\theta) = \prod_{=1} (\mu_\theta, \tau)$$

Prior

$$(\theta) = \log \quad(\quad|\theta) \quad(\theta) \leftarrow \propto \quad(\theta|\quad)$$

Posterior Probability

$$\hat{\theta} = \arg\max_{\theta} \quad(\theta)$$

- Initial prior parameters are given explicitly.

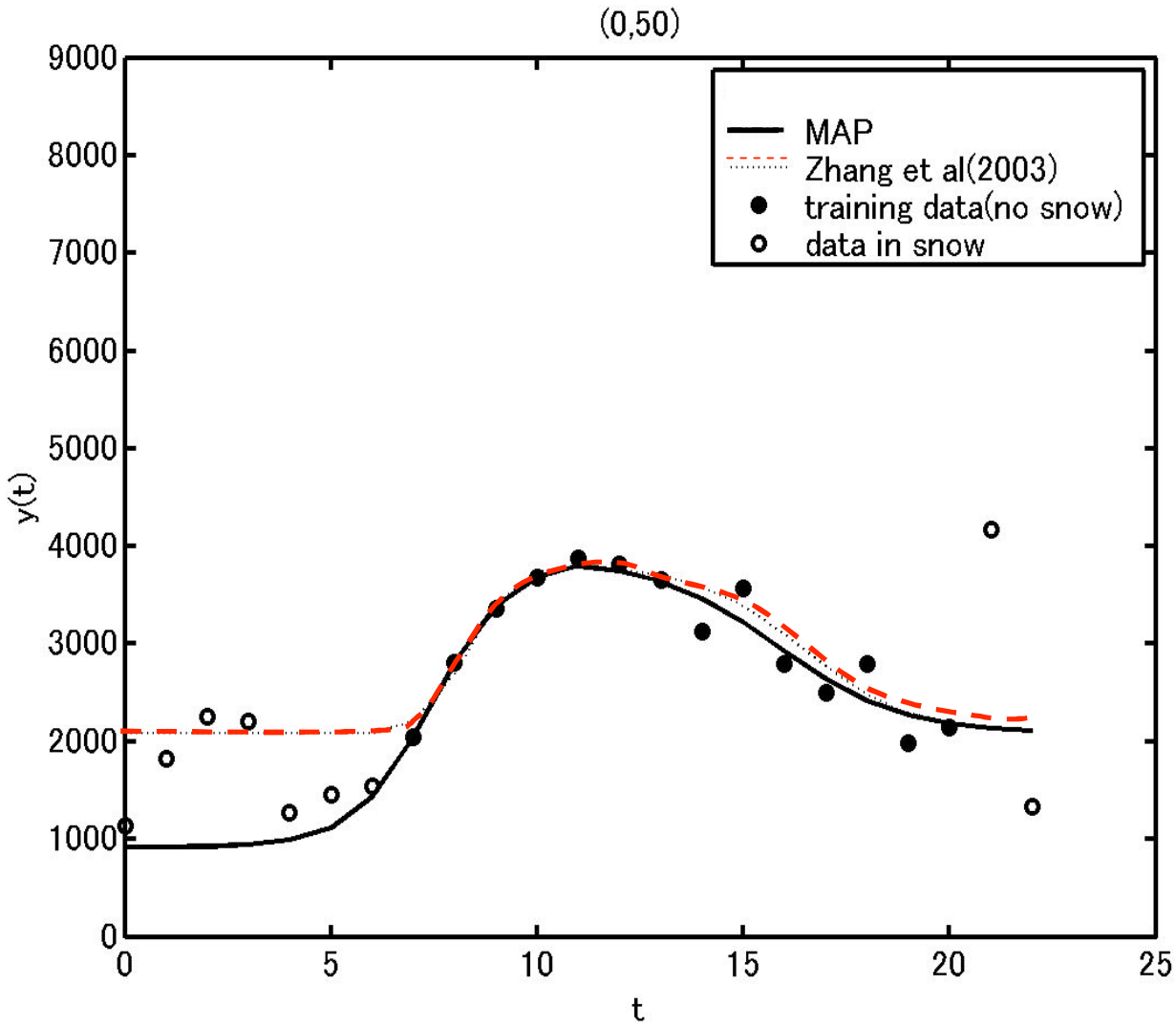
μ_θ : Mean of ML solutions for many pixels.

τ : Standard deviation of ML solutions for many pixels.

Positive effects

- Preprocessing of the data such as smoothing , guessing of missing data and segmentation are not required.
  - Still solutions are obtained accurately and robustly.
- Uncertainties of parameter estimates are assessed via confidence intervals
- Relevant prior information is introduced if it is available.

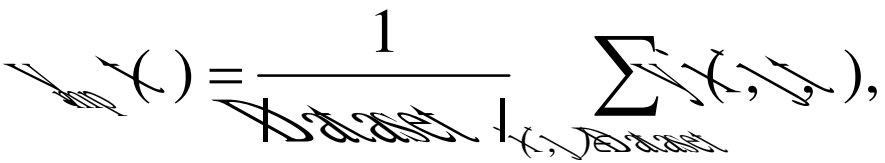
Example of fitting

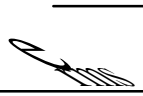


Performance of modeling by MAP

- Take one time point from each time series and guess that value from the model built from the rest of the data.
- Prediction accuracy measure
  - Normalized prediction error

  - Mean predictor

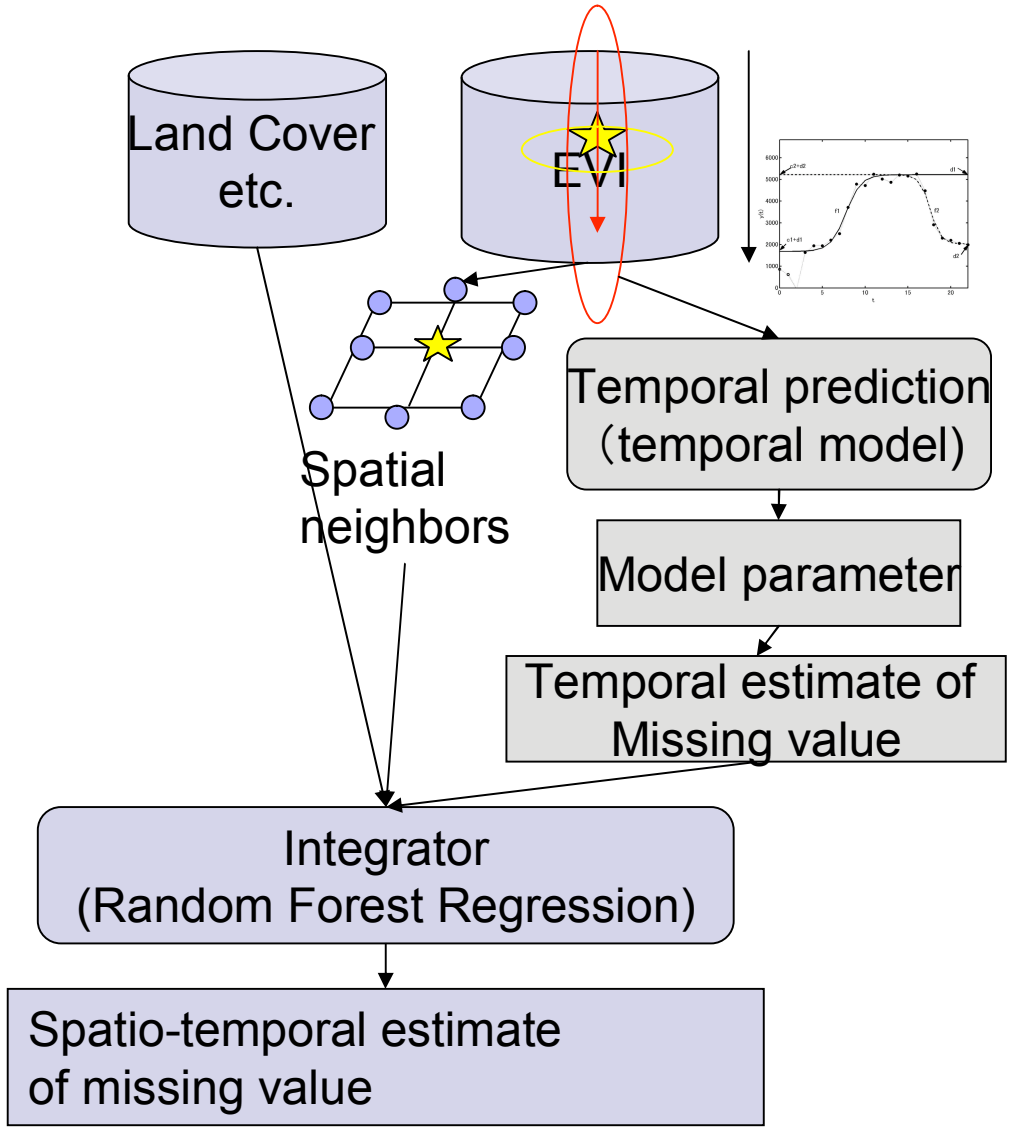


	MAP	Zhang et al (2003)
	0.135	0.209*

* Fitting error (not a prediction error)

Spatio-temporal estimation

- Missing value estimation by use of both temporal model, spatial neighbors and other information

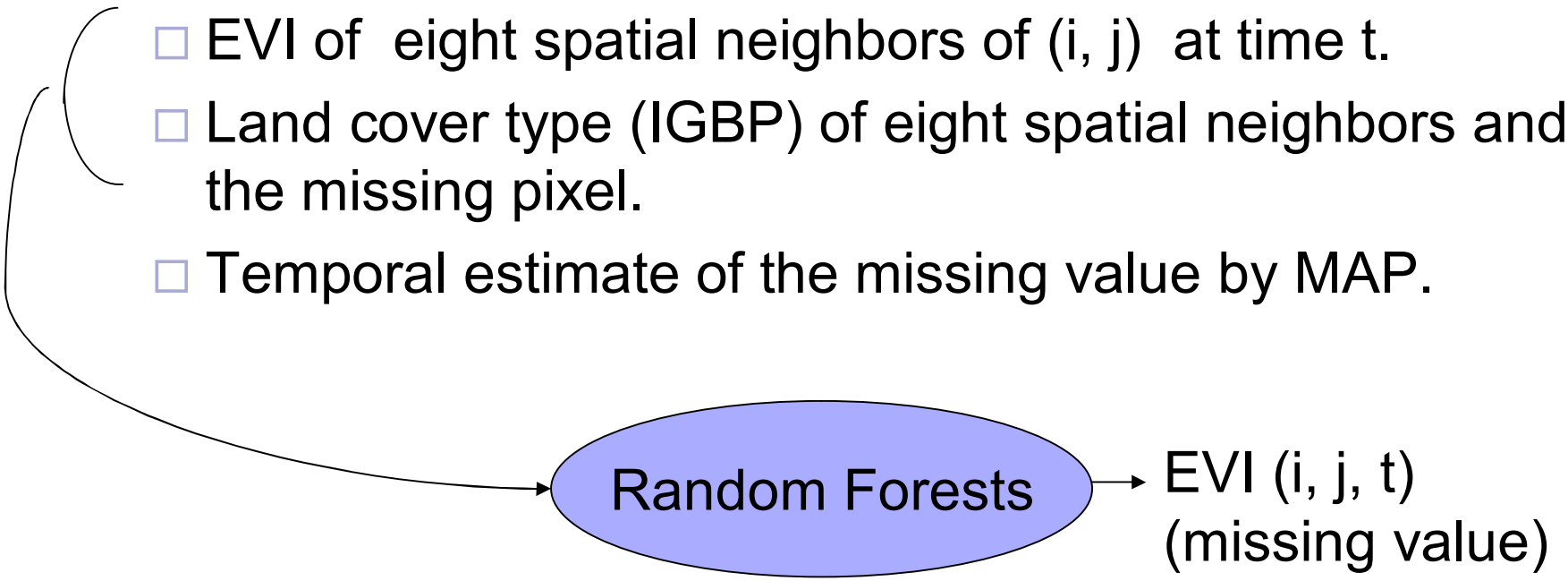


Random Forests (Breiman, 2001)

- An ensemble of unpruned classification or regression trees induced from bootstrap samples of the training data, using random feature selection in the tree induction process.
- Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble.

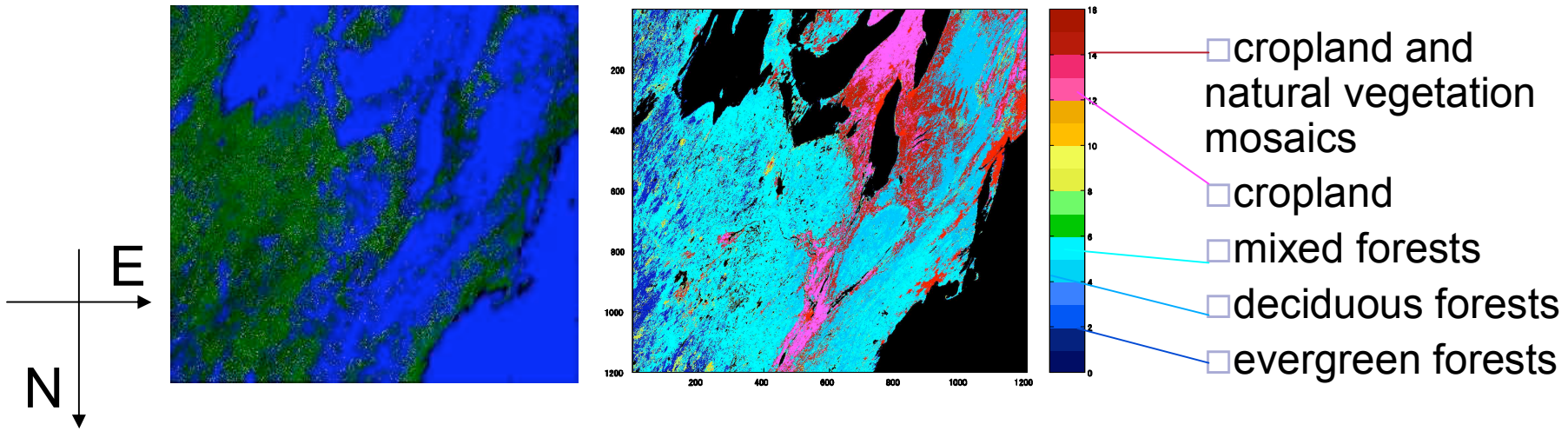
Input and target attributes of Random Forests

- Target
  - EVI of the (i, j) pixel at time t. (the missing value)
- Input
  - EVI of eight spatial neighbors of (i, j) at time t.
  - Land cover type (IGBP) of eight spatial neighbors and the missing pixel.
  - Temporal estimate of the missing value by MAP.



The dataset used for evaluation

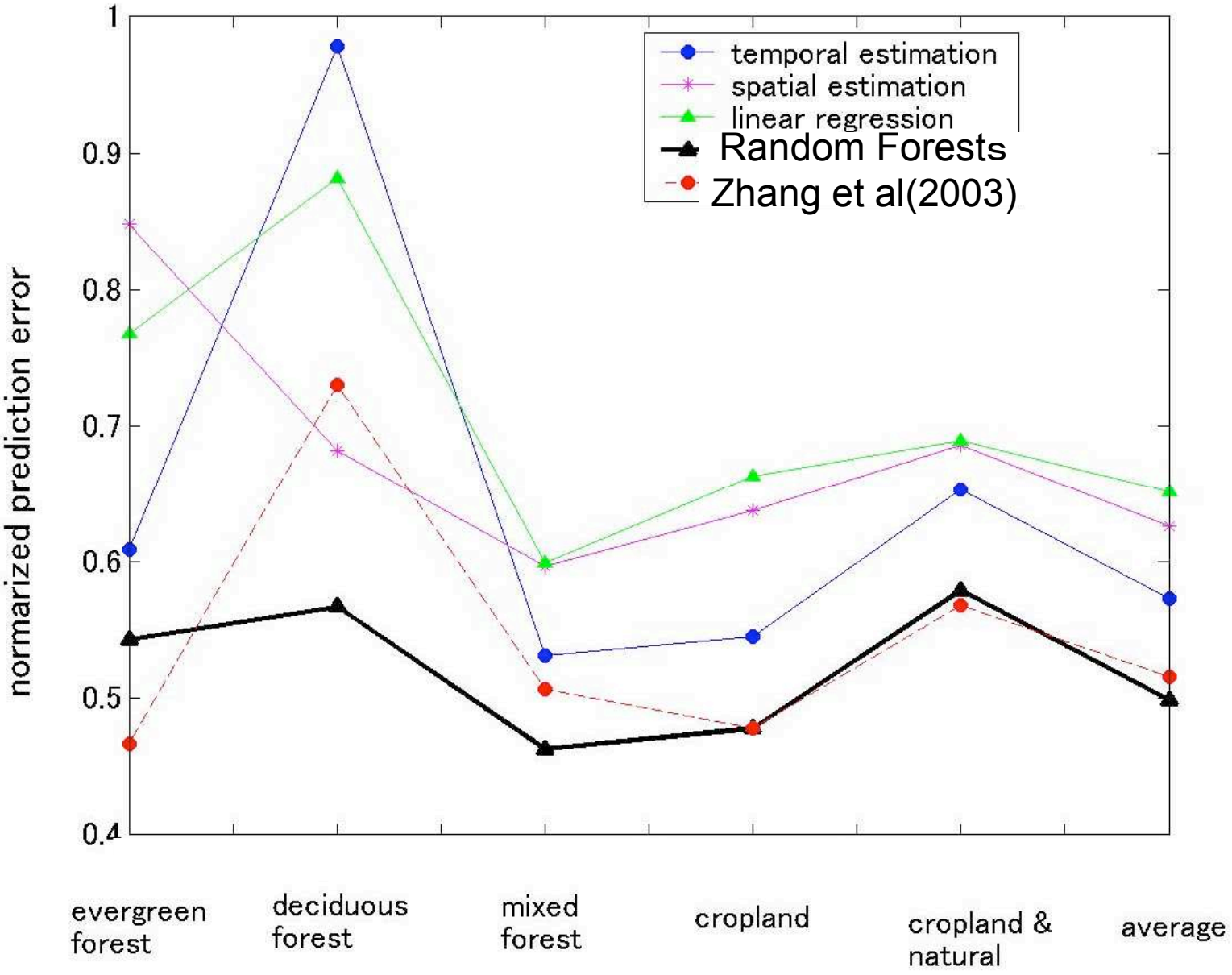
- MODIS data of Northeastern United States
  - EVI, “snow flag”, land cover type (IGBP) for each pixels
  - 16 days interval (January 2001 to December 2001)
  - A spatial resolution: 1km per pixel (1200 × 1200 pixels)
- Six datasets for evaluation
  - evergreen forests, deciduous forests, mixed forests, cropland, cropland and natural vegetation mosaics, and the mixture of five groups (about 1000 samples in each and 5000 for the mixture).



Evaluation Method

- The half of the data set was used for training of random forests and the rest was used for evaluation .
- Other Methods in comparison
  - Logistic model fitting via MAP (temporal prediction).
  - Four neighbor average (spatial prediction).
  - Linear regression by using all the attributes.

- Accuracy Measure
  - Normalized prediction error.


temporal estimation
spatial estimation
linear regression
Random Forests
Zhang et al(2003)
normarized prediction error
1
0.9
0.8
0.7
0.6
0.5
0.4
evergreen forest
deciduous forest
mixed forest
cropland
cropland & natural
average

Conclusion

- The effectiveness of MAP for temporal modeling for such dataset as MODIS EVI is confirmed.
  - No preprocessing (e.g., segmentation and filling of missing values) and required and still the solutions are obtained robustly.
- Random Forests is found to be effective to improve the temporal estimate.
  - Once the predictor is created, temporal estimation is improved promptly.
- Future study
  - Application to multiple-year data set, and detection of spatio-temporal pattern change.
  - Automatic detection of sub-seasonal pattern.
  - The models considering spatial distribution (e.g., mixture model for model parameters in the spatial domains) .

Acknowledgement

- The author deeply appreciates Prof. Padhraic Smyth of UCI, Prof. Mark Friedl and Dr. Xiaoyang Zhang of Boston University for their suggestion from the earliest stage of this study.

Discussion Questions

- 1. What data mining or statistical methods were used?
  - Maximum a Posteriori, Random Forests
- 2. How were the techniques developed?
  - b) modified from another application or research project
- 3. What is the importance of the science question?
  - How we can deal noises, missing data and construct the temporal model robustly.
- 4. What scientific results were obtained that would have been
- difficult or impossible without data mining and/or statistics?
  - Robust fitting without pre-processing and improvement of estimation
- 5. What were the obstacles?
  - Missing values, noises.
- 6. Did this work result in a geoscience publication? If not, why not?
  - Not yet.

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.