NASA AIST NNX15AH48G | NNX17AC89G

Hierarchical modeling for climate-vegetation dynamics

Kamalika Das
ESTF 2017

Team: Marcin Szubert[*], Josh Bongard[*], Anu Kodali[+], Sangram Ganguly[+]
[*]University of Vermont
[+]NASA Ames Research Center

State-of-art in modeling

- Physics based modeling and perturbation theory
  - Assumptions of equilibrium restricted to ideal conditions
- Statistical correlation analysis
- Machine learning: linear and fixed order nonlinear models
  - Restricted model complexity fails to capture real dynamics

Proposed approach: GP Regression Trees

- Genetic programming based symbolic regression
  - Evolutionary optimization based technique that allows for discovery of free-form equations
  - Leads to discovery of unknown physical processes
- Hierarchical partitioning along the lines of classification and regression trees (CART)
  - Single model unable to capture spatio-temporal variations
  - Multiple models discover nonlinear boundaries where the physical processes change

Symbolic regression

- Data driven, white box, nonlinear modeling
  - Divide available observational data into training and validation
  - Distills equations of arbitrary form and complexity

$$Y = -0.01log(e^{X_8}(0.03e^{4X_6+X_8+2X_9}((X_5+X_6)^2 - X_2 - X_3)^2 + 0.2e^{X_{10}}))$$

  - Starts with random instantiation of a population
  - Stochastic optimization based search of the space for “better” model
  - Accurate vs parsimonious models
  - Solutions form a Pareto front
  - Requires searching through thousands (often millions) of candidate solutions; can be parallelized

Hierarchical partitioning

- Step 1: Induce model tree
  - Pick one of the predictor variables, X_i
  - Pick a value of $X_{i,}$ say s_i, that divides the training data into two (not necessarily equal) portions
  - Measure the MSE of a 2nd order polynomial regression equation on that partition
  - Evaluate for different values of $X_{i,}$ and s_i to minimize errors in initial split
  - Repeat process for each split, until desired number of partitions are reached

Output: A binary tree with 2^{depth} leaf nodes

- Step 2: Symbolic regression based equation discovery at each leaf node partition

Case study: Tropical rainforests

- Amazon + Congo + Indo-Malay rainforests
  - Largest terrestrial carbon sink (40-50%)
  - Amazon: Major droughts in 2005, 2010, 2015
  - Congo: Ongoing drought since 2000
  - Huge discrepancies in research conclusions

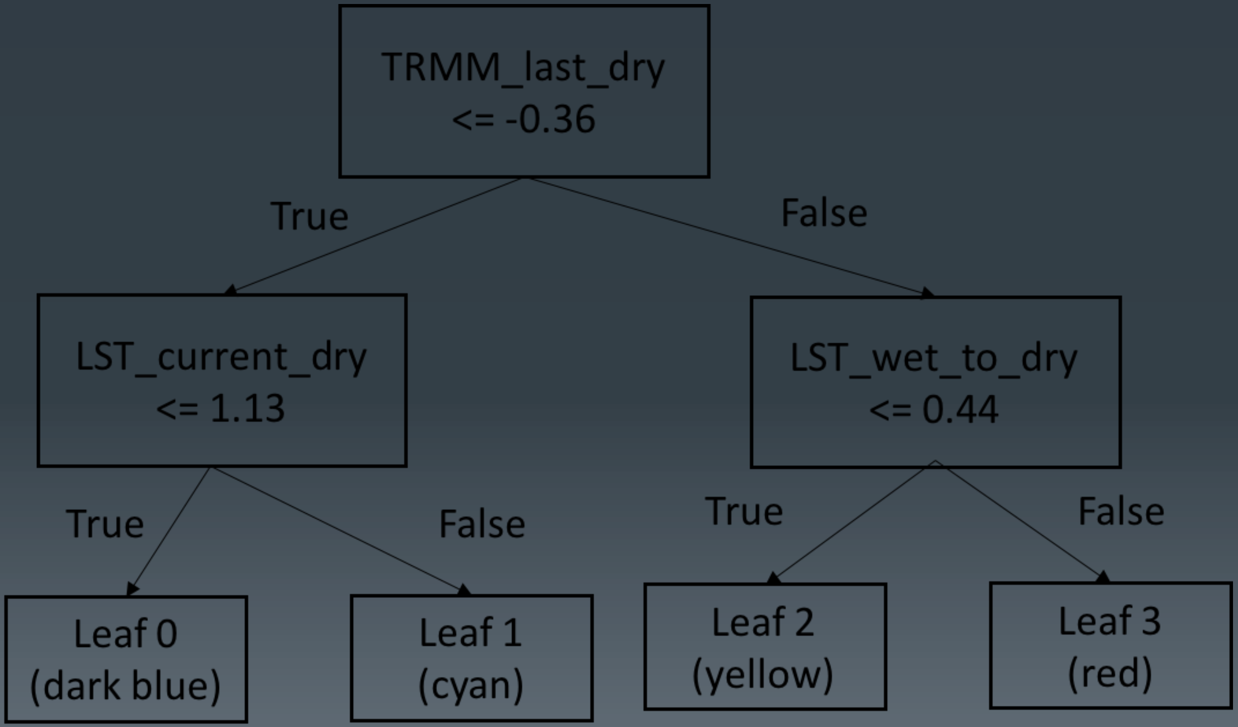
Data sets

- NDVI: Normalized Difference Vegetation Index (MOD13Q1: Terra MODIS, Collection 6)
- Land Surface Temperature (day temperature) (MOD11A1: Terra MODIS, Collection 6)
- Precipitation (TRMM Product 3B43V7)
- Elevation (GTOPO30)
  - Derived feature: Slope (Horn’s method)
- Landcover Mask (LCT from MOD12Q1.051, Collection 5)

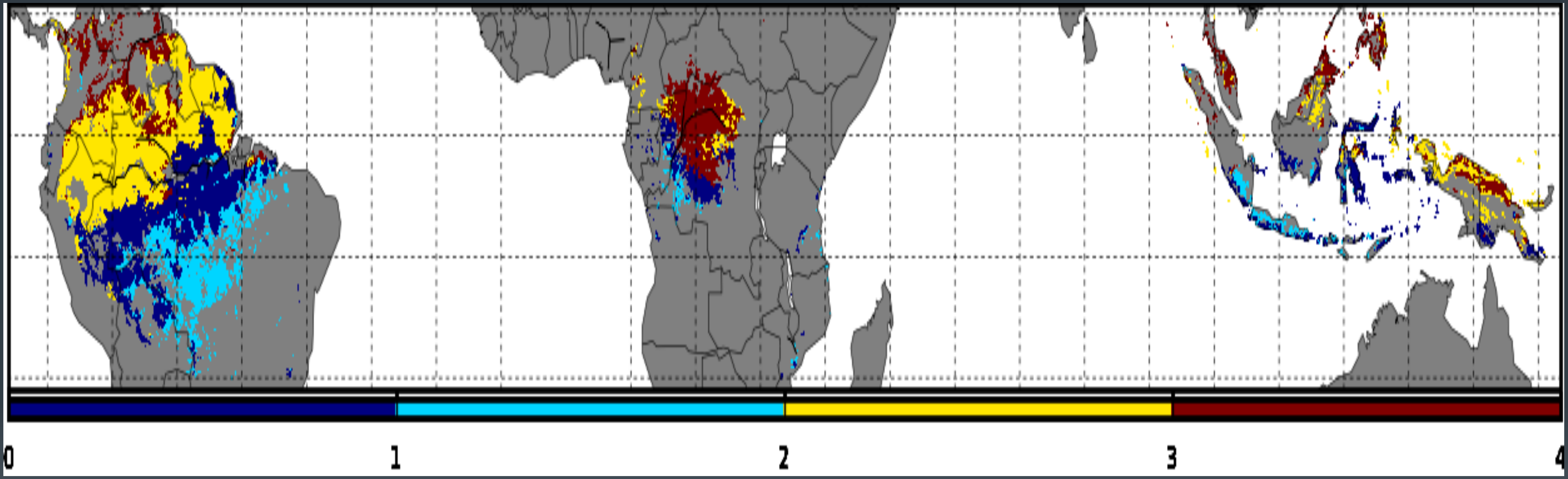
- Dependent variable: vegetation (NDVI)
- Independent variables: seasonal precipitation and land surface temperature going back up to a year in time, elevation, and slope

$\mathrm{NDVI}_k =$ f(LST_i ; TRMM_i ;Elev; Slope), where $k =$ $\mathrm{current}_D$ and I (D_{current} ;D_{last} ;WD;W;DW)

Results and analysis: Global rainforests



Results and analysis



Southern Amazon basin: Temperature dominated, negligible effect of precipitation

Transitional savannahs: Dominated by wet season rainfall

Northern Amazon basin: Temperature dominated, wet season rainfall plays a role in greenness

Congo & Amazon northern fringes: Complex model; lack of defining characteristics

Conclusion and future work

- The equation discovery technique leads to identification of new physical processes while learning from the data
- Method is independent of the domain and can be used for any number of applications
- Identifies the biggest influencers

- Future work
- Causality analysis

Publications & presentations

Workshops & Posters

1. Regression based modeling of vegetation and climate variables for the Amazon rainforest. Ankush Khandelwal, Anuradha Kodali, Marcin Szubert, Sangram Ganguly, Joshua Bongard, Kamalika Das. AGU Fall Meeting, San Francisco, December 2015.
2. Predicting the future of the Amazon rainforests using regression analysis. Kamalika Das, Anuradha Kodali, Marcin Szubert, Joshua Bongard, Sangram Ganguly. ESTF 2016, Annapolis, MD.
3. Quantifying how climate affects vegetation in the Amazon rainforests. Kamalika Das, Anuradha Kodali, Marcin Szubert, Sangram Ganguly, Joshua Bongard. AGU Fall Meeting, San Francisco, December 2016.

Conferences:

1. Reducing antagonism between behavioral diversity and fitness in semantic genetic programming. Marcin Szubert, Anuradha Kodali, Sangram Ganguly, Kamalika Das, Joshua Bongard. Genetic and Evolutionary Computation Conference (GECCO), Denver,CO July 2016.
2. Semantic forward propagation for symbolic regression. Marcin Szubert, Anuradha Kodali, Sangram Ganguly, Kamalika Das, Joshua Bongard. International Conference on Parallel Problem Solving for Nature (PPSN), Edinburgh, Scotland, September 2016.
3. Understanding climate-vegetation interactions in global rainforests through a GP-tree analysis Submitted to European Conference on Machine Learning and Practices in Knowledge Discovery 2017.

Journals

1. Amazonian Forests show Resiliency to Extreme Droughts. S. Ganguly, A. Kodali, M. Szubert, J. Bongard, M. H. Costa, R. Nemani, K. Das. Submitted to Nature Climate Change.
2. Identifying the biggest climate controls for the Amazon rainforests. Marcin Szubert, Anuradha Kodali, Sangram Ganguly, Joshua Bongard, Kamalika Das. To be submitted to Nature Communications.

Code

- https://bitbucket.org/marcin_sz/aist_gp
  - Data pipeline
  - Hierarchical GP