Landsat Image Time Series Processing using HTCondor on UW-CHTC and OSG Resources

Matthew Garcia, Ph.D.
Prof. Philip A. Townsend

Dept. of Forest & Wildlife Ecology
University of Wisconsin–Madison

HTCondor Week
24 May 2018
So you think you need to model your data…
NDII

Mean Phenology (Rsq = 0.594)

Single pixel time series

KTTC statistical outliers: red

Retained Landsat dates: black

Fitted curve: blue $\rightarrow r^2 \sim 0.6$
NDII mean phenology

$\mu = 0.618$

$\mu = 0.062$
PLS: Projection to Latent Structures, a.k.a.
PLSR: Partial-Least-Squares Regression

Similar to PCA, but…

- maximizes covariance, instead of minimizing correlation
- incorporates the response variable, not just the predictors

Unlike OLS regression, does not assume predictors are error-free

Similar to Multiple Linear Regression, but handles predictor collinearity
→ able to handle many predictor variables with few response variables

$$
\begin{bmatrix}
W_{1,1} & \cdots & W_{m,1} \\
\vdots & \ddots & \vdots \\
W_{1,n} & \cdots & W_{m,n}
\end{bmatrix}
\begin{bmatrix}
\beta_1 \\
\vdots \\
\beta_m
\end{bmatrix}
=
\begin{bmatrix}
V_1 \\
\vdots \\
V_n
\end{bmatrix}
$$
Computational Details

Weather/climate pixels @ 480-m resolution
Landsat pixels @ 30-m resolution

→ Geographic chunks of collected pixels
  (1 weather/climate + 16 x 16 Landsat)
  ~1.5 MB/chunk collected input data
  → ~50 MB/chunk raw output data

~130 million Landsat pixels over 5 footprints
~70,000 – 140,000 chunks per footprint
~624,000 total chunks

Ideal task for distributed processing:
→ UW CHTC for pre-processing
→ OSG for statistical modeling
→ UW CHTC for post-processing
Mean Phenology: \textit{NDII} fitted curve error statistics and goodness-of-fit

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fitted_curve_error_stats.png}
\caption{Fitted curve RMSE and $r^2$ maps.}
\end{figure}
mean phenology

residuals via PLSR

full phenology model

RMSE

\[ \mu = 0.062 \]

RMSE

\[ \mu = 0.735 \]

RMSE

\[ \mu = 0.030 \]

\[ r^2 \]

\[ \mu = 0.618 \]

\[ r^2 \]

\[ \mu = 0.451 \]

\[ r^2 \]

\[ \mu = 0.944 \]
Statistical model:
~624K chunks @ ~12.6 h/chunk = ~8 Mh

Overall processing time:
~13 million computing hours
~5.6 Mh on OSG nodes
~5.1 Mh on CHTC resources
~2.3 Mh on other UW clusters
Thank you!

http://matthewgarcia.tech