A new machine learning-based methodology for assessing the impact of agricultural interventions in Nepal

Friday, June 14, 2019
Background on Interventions

- **Agricultural Intervention**: Fostering the modernization of agricultural systems by deliberate and strategic actions on the part of governments/organizations.

  Agricultural interventions are performed with the main goals of improving yield, increasing profit for farmers, and improving nutrition for the poor.

  Assessing agricultural interventions is difficult for many reasons, including cost, time needed, and remoteness of the Nepalese countryside.

  This research is an effort to assess a new impact evaluation approach that addresses the challenges of traditional assessment methods, improves on other contemporary methods, and builds off other ongoing initiatives.
Yield prediction with satellite imagery

- **Multi-Linear Regression (MLR)**
  - Drummond (2003) – used MLR to predict yield of soybean fields in Missouri, USA
    - $R^2$ for MLR = 0.31, ANN = 0.45
  - Gonzalez-Sanchez (2014) – used MLR to predict 10 different crops in Sinaloa, Mexico
    - $R^2$ MLR = 0.25, ANN = 0.21

- **Random Forest (RF)**
  - Fukuda et al. (2013) – used Random Forests to estimate mango fruit yields in Thailand in response to water supply under different irrigation regimes
    - $R^2$ = 0.69
  - Jeong (2016) – used Random Forests to estimate global wheat and maize yield
    - $R^2$ RF = 0.90, MLR = 0.49
Specific Application – SEIRS

- Synthetic Counterfactual Variables and Impact Assessment (SEIRS)
- USAID/CIAT project
  - Terra-I – uses MODIS and TRMM to detect forest changes from NDVI time series and precipitation data
- SEIRS also used historical NDVI and rainfall data to back-predict NDVI, on cropland instead of forests
- This prediction, made using a convolutional neural network, served as a counterfactual to compare the observed result to

Credit: Pete Richards, USAID
How my work builds off SEIRS

<table>
<thead>
<tr>
<th></th>
<th>SEIRS</th>
<th>My Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of inputs</td>
<td>2 – NDVI, precipitation</td>
<td>Many - NDVI, precipitation, elevation, phenology cluster, treatment type...</td>
</tr>
<tr>
<td>Satellite imagery used</td>
<td>MODIS 250m spatial resolution 1-2 day temporal resolution</td>
<td>Landsat 30m spatial resolution 16 day temporal resolution</td>
</tr>
<tr>
<td></td>
<td>Type of neural network</td>
<td>Recurrent Neural Network (RNN)</td>
</tr>
<tr>
<td></td>
<td>Convolutional Neural Network (CNN)</td>
<td></td>
</tr>
</tbody>
</table>
Datasets

- Landsat constellation (5, 7, 8)
  - Calculate NDVI time series from Landsat imagery
  - Smooth and de-spike time series
  - Clustering done on smoothed/de-spiked NDVI time series to simulate phenology
- Planet/Digital Globe - validation
- CHIRPS precipitation – pentad, 5km
- SALDAS temperature – dekad, 250m
- NASA SRTM elevation – 90m (2014)
- World Bank field intervention dataset

* - South Asia Land Data Assimilation System – Ben Zaitchik, Johns Hopkins University
Study Area

Area of Study

Elevation (meters)
- <0
- 0-500
- 500-1000
- 1000-2000
- 2000-3000
- 3000-4000
- >4000

China
Nepal
India
Collect Earth Online was used for accuracy assessment of World Bank field dataset, 30 meter grid around the centroid of each field to simulate Landsat scale.

Overall, 67% of the plots (~1900) were classified as 100% agricultural cover. These fields are the ones that have been used for the analysis.
Prior to clustering, data was **de-spiked and smoothed** for optimal clustering.

Not all the same crop or same phenology cycle represented in this dataset.

The fields were **separated into clusters based on phenology**. Each cluster was used as a parameter to input into neural network.
The machine learning approach will aim to **back-predict NDVI** using the aforementioned datasets:

- PCA conducted for multicollinearity

Machine learning approach – used [neuralnet package in R](https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf), recurrent neural network, back propagated error

Because we know what the observed NDVI values are, we can compare the predictions from the machine learning approach to the observed NDVI values.

The overall relationship between the predicted NDVI and actual NDVI can then be determined using statistical methods (**R², MAE, RMSE**).
Neural Network Optimization

- First, the optimal number of hidden layers and neurons in each hidden layer was found.

- Literature suggests that the best way to find the optimal setup is by trial and error, but there are constraints:
  - One or two hidden layers is enough for the vast majority of applications.
  - Don’t have more neurons in any individual hidden layer than the total number of inputs.

- The best configuration found was one hidden layer, five neurons in the hidden layer, and this was the configuration used for comparison.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>$R^2$</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>:5:</td>
<td>0.170361</td>
<td>0.133822</td>
<td>0.1729</td>
</tr>
<tr>
<td>:4:</td>
<td>0.170889</td>
<td>0.134382</td>
<td>0.1677</td>
</tr>
<tr>
<td>:3:3:</td>
<td>0.171059</td>
<td>0.134503</td>
<td>0.1661</td>
</tr>
<tr>
<td>:2:5:</td>
<td>0.171226</td>
<td>0.13481</td>
<td>0.1644</td>
</tr>
<tr>
<td>:3:2:</td>
<td>0.171265</td>
<td>0.134713</td>
<td>0.164</td>
</tr>
</tbody>
</table>

Table: The five configurations with the best statistics out of the 30 neural network configurations tested.
Comparing MLR, RF, ANN

<table>
<thead>
<tr>
<th></th>
<th>$R^2$</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-Linear Regression</td>
<td>.255</td>
<td>.162</td>
<td>.124</td>
</tr>
<tr>
<td>Random Forest</td>
<td>.666</td>
<td>.109</td>
<td>.076</td>
</tr>
<tr>
<td>Neural Network</td>
<td>.173</td>
<td>.170</td>
<td>.134</td>
</tr>
</tbody>
</table>
Thus far, artificial neural networks have not yet shown an improvement over the established multi-linear regression and random forest approaches.

The artificial neural network approach could be improved with increased repetitions during the training process, decreasing the threshold for the partial derivatives of the error function (which is the stopping criteria for each repetition in the training process).

Future work will include holding space and/or time constant for ANN approach. Ex: How does the neural network work on each district separately? How did neural network do in different years?

Additional future work includes incorporating in-situ data and additional datasets (soil composition, moisture, in-situ yield data) if available, further testing of other neural network approaches, testing this methodology over different areas of the world.
Funding Acknowledgement

Funding for this research is provided by NASA SERVIR through NASA Cooperative Agreement NNM11AA01A.
References

Image Citations
Stanford University, Pete Richards

Data & Methodology Citations
Landsat, Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), Planet Labs, Digital Globe, Famine Early Systems Warning Network (FEWSNET), Synthetic Counterfactual Variables and Impact Assessment (SEIRS), NASA Shuttle Radar Topography Mission (STRM)

Major Literature Citations
World Bank Development Impact Evaluation Group (DIME), Food and Agricultural Program of the United Nations (FAO), World Food Programme (WFP). Full list of references available upon request.