Synergy Between Remote Sensing and Machine Learning for Crop Yield Prediction

Luca Sartore¹,², Arthur Rosales², David Johnson², Mary Frances Dorn³, Clifford Spiegelman⁴

¹National Institute of Statistical Science (NISS)
²United States Department of Agriculture
   National Agricultural Statistics Service (USDA NASS)
³Los Alamos National Laboratory (LANL)
⁴Texas A&M University, College Station (TAMU)

lsartore@niss.org

GASP 2019
September 23, 2019

“. . . providing timely, accurate, and useful statistics in service to U.S. agriculture.”
Disclaimer and acknowledgments

The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA, Los Alamos National Laboratory, or US Government determination or policy.

- This research was supported in part by the intramural research program of the U.S. Department of Agriculture, National Agricultural Statistics Service.
- This research used resources provided by the SCINet project of the USDA Agricultural Research Service, ARS project number 0500-00093-001-00-D
Presentation Outline

1. Motivation and scientific background
2. Modeling approaches
3. Case study (preliminary study)
4. Results and concluding remarks
PART I

MOTIVATION AND
SCIENTIFIC BACKGROUND
Modeling crop yield with remote sensing data

USDA National Agricultural Statistics Service (NASS) produces county level estimates of crop yield by combining several sources of information.

Remote sensing technology provides a variety of data to assess the status of the agriculture.

Several challenges arise from:
- Land-cover and crop identification
- Non-parametric modeling

Prediction accuracy and computational efficiency are major concerns.
Use of satellite imagery in agriculture

Remote sensing have been used in several countries to estimate crop production or yield (Zhao et al., 2007)

“Greenness” of plants is characterized by a spectral signature that can be used to determine the state, structure and composition of the crop

Satellite imagery of planted areas shown strong associations between the spectral signature and crop production or yield

Johnson (2016) investigated correlations between MODIS data and crop yield at the county level for several commodities
PART II

MODELING APPROACHES
Current steps to predict crop yield

1. Discriminating of crop planted in the fields within each county
2. Estimating the average value for the variables of interest at the county level
3. Combining remote sensing variables with historical yield data
4. Training non-parametric models
5. Performing model selection
6. Producing predictions

Instead of averaging field level data points at the county level, more information is considered by using empirical distributions
Figure: Land cover in Eddy county (ND) in 2018
Javascript and Google Earth Engine

Figure: Median reflectance in 2018 over corn fields

https://code.earthengine.google.com/
Using approximate densities as covariates

The distributions of the variables of interest can be used as **functional covariates** rather than computing their expected values.

For computational reasons, these density functions can be approximated as **histograms**.
Measuring distances between densities

The **symmetric Kullback-Leibler distances** (SKLD) between histograms is computed as

\[
\text{SKLD} = \frac{1}{2} \int_{\mathbb{R}} \left[ f_A(x) - f_B(x) \right] \log \left( \frac{f_A(x)}{f_B(x)} \right) \, dx
\]

Standard data frames for model training are obtained by

1. **Multidimensional scaling** (MDS, Kruskal, 1964)
2. **Principal components** (PC, Jolliffe, 1986)
3. **Independent components** (IC, Hyvarinen, 1999)

applied on the SKLD matrix
PART III

CASE STUDY
Case study

Public USDA data on corn yield, and NASA data related to
- Visible red light reflectance
- Near-infrared reflectance
- Afternoon surface temperature

from 2008 to 2017 for most of the counties in IL, IN, IA, KS, MN, MO, NE, OH, SD, WI (Corn Belt states) are considered

Data from 2008 to 2016 are used for **training** (to estimate parameters) and **validation** (to avoid over-fitting)

Data from 2017 are used for **testing** the prediction models (to evaluate the extrapolation on new data)
A closer look to the data

- **Response Variable** (unidimensional response)
  - Corn yield

- **Covariates** (3 unidimensional and 138 functional covariates)
  - Time (expressed in calendar years)
  - Latitude (county centroid in angular degrees)
  - Longitude (county centroid in angular degrees)
  - Approximate densities for
    - 46 visible red light measurements (%)
    - 46 near-infrared measurements (%)
    - 46 afternoon surface temperature measurements (K)

Histograms with 256 bins are obtained each for each 8-days period within a year
Models for studying yield with SKLDs

- Linear
- Multivariate adaptive regression splines (MARS, Friedman et al., 1991)
- \( k \)-nearest neighbors
- Support vector machines (SVM, Cortes and Vapnik, 1995)
- Regression trees (CART, Breiman et al., 1984)
  - Bagging (Breiman, 1996) and boosting (Freund et al., 1996)
- Random forests (Breiman, 2001)
- Cubist (Rulequest, 2006)

Ten-fold cross-validation is performed 5 times by randomly assigning each record to the ten groups (for more details, see Kuhn and Johnson, 2013)
Computational Environment

All computations are performed in R, Javascript and C

Specific shell-scripts have been produced to execute parallel processes on several nodes

USDA ARS resources related to the SCINet project (https://www.ars.usda.gov/scinet/) have been used:

- 58 HPC nodes with 40 CPU cores and 128GB RAM
- 5 high-memory nodes with 120 CPU cores and 1.48TB RAM

About four days elapsed for predicting all counties in the corn belt
PART IV

RESULTS AND CONCLUDING REMARKS
Multidimensional scaling

Cross–validation results

Testing predictions

MAE (US bushels per acre)

RMSE (US bushels per acre)

USDAGASP 2019 – Luca Sartore
Principal components

Cross–validation results

Testing predictions

MAE (US bushels per acre)

RMSE (US bushels per acre)

MAE (US bushels per acre)

RMSE (US bushels per acre)

○ Cross–validation results
+
Testing predictions

GASP 2019 – Luca Sartore

USDA
Independent components

Cross-validation results

Testing predictions

MAE (US bushels per acre)

RMSE (US bushels per acre)

MAE (US bushels per acre)

RMSE (US bushels per acre)

- Bag. Tree
- MARS
- CART
- Rnd. Forest
- Boost. Tree
- SVM
- Lin. Mod.
- Cubist

○ Cross-validation results
+ Testing predictions
Performance evaluation of Cubist

<table>
<thead>
<tr>
<th></th>
<th>MDS</th>
<th>PC</th>
<th>IC</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.V. MAE</td>
<td>9.156</td>
<td>11.592</td>
<td>8.840</td>
</tr>
<tr>
<td>C.V. RMSE</td>
<td>12.333</td>
<td>15.205</td>
<td>11.502</td>
</tr>
<tr>
<td>C.V. MAE S.D.</td>
<td>0.289</td>
<td>0.337</td>
<td>0.285</td>
</tr>
<tr>
<td>C.V. RMSE S.D.</td>
<td>0.444</td>
<td>0.530</td>
<td>0.364</td>
</tr>
<tr>
<td>Test MAE</td>
<td><strong>15.511</strong></td>
<td>15.927</td>
<td>15.613</td>
</tr>
<tr>
<td>Test RMSE</td>
<td><strong>19.621</strong></td>
<td>20.950</td>
<td>19.877</td>
</tr>
</tbody>
</table>

The use of independent components would be selected by cross-validation, but multidimensional scaling performs best during extrapolation.
Prediction errors from the best predictive model

Figure: The map shows unexplained spatial dependence
Conclusion

- The use of **approximate densities** as functional covariates allows to consider a full stochastic process at the field level.
- **IC analysis** produces better results during cross-validation, but its ability during extrapolation is not fully clear.
- **Computer clusters** and non-standard coding techniques for
  - Data storage
  - Analyses
- Further research should be conducted for the **model evaluation** of
  - neural networks
  - spatio-temporal dependencies
- Developing an algorithm that is **robust to measurements error on the covariates**
Selected References


Thank you!

Questions?

Luca Sartore, PhD  lsartore@niss.org
Arthur Rosales  arthur.rosales@usda.gov
David Johnson  david.m.johnson@usda.gov
Mary Frances Dorn, PhD  mfdorn@lanl.gov
Clifford Spiegelman, PhD  cliff@stat.tamu.edu