Synergy Between Remote Sensing and Machine Learning for Crop Yield Prediction

Luca Sartore[1,2], Arthur Rosales[2], David Johnson[2],
Mary Frances Dorn[3], Clifford Spiegelman[4]

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

lsartore@niss.org

GASP 2019
September 23, 2019

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

USDA NASS Cropland Data Layer

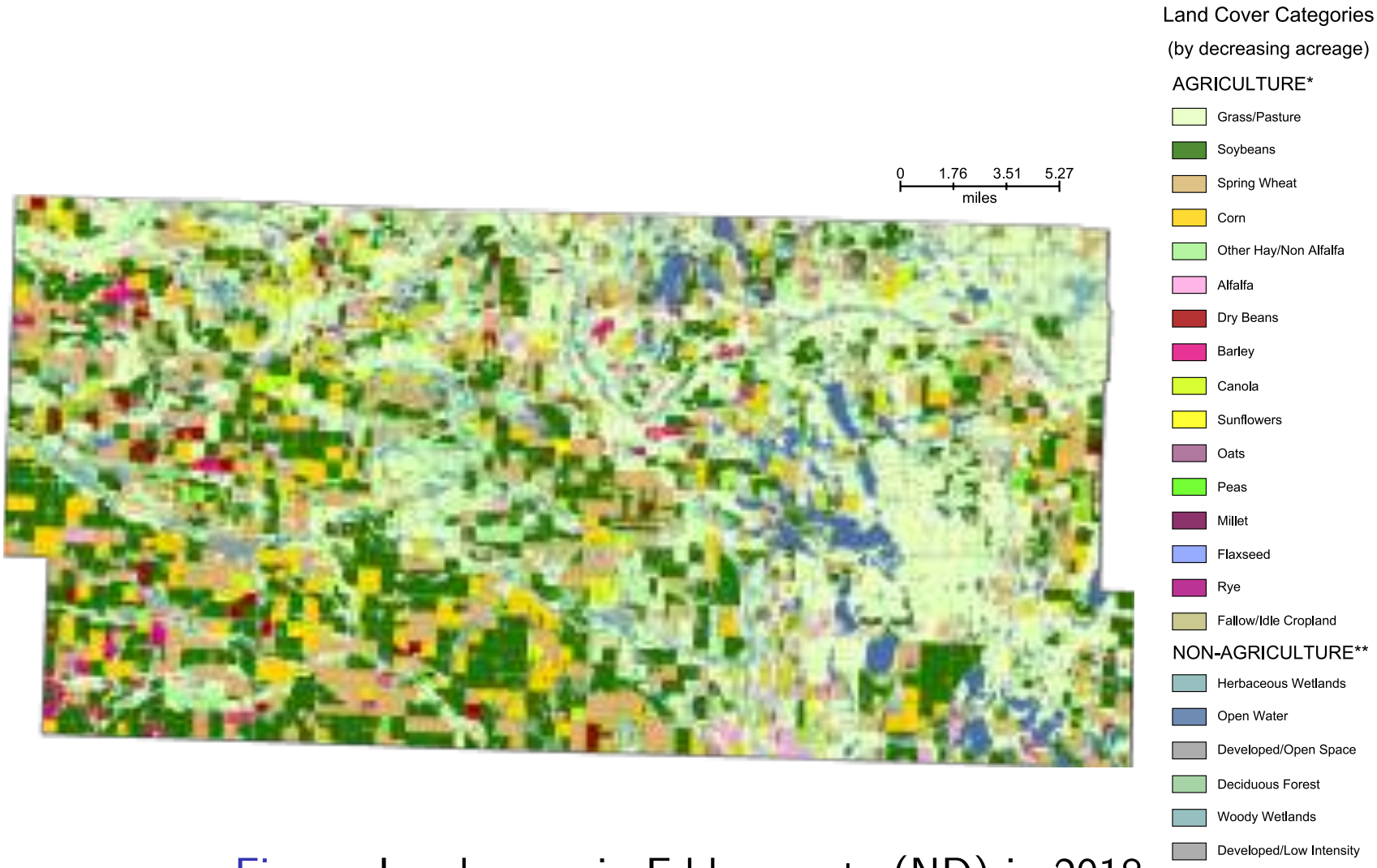


Figure: Land cover in Eddy county (ND) in 2018

Javascript and Google Earth Engine[1]

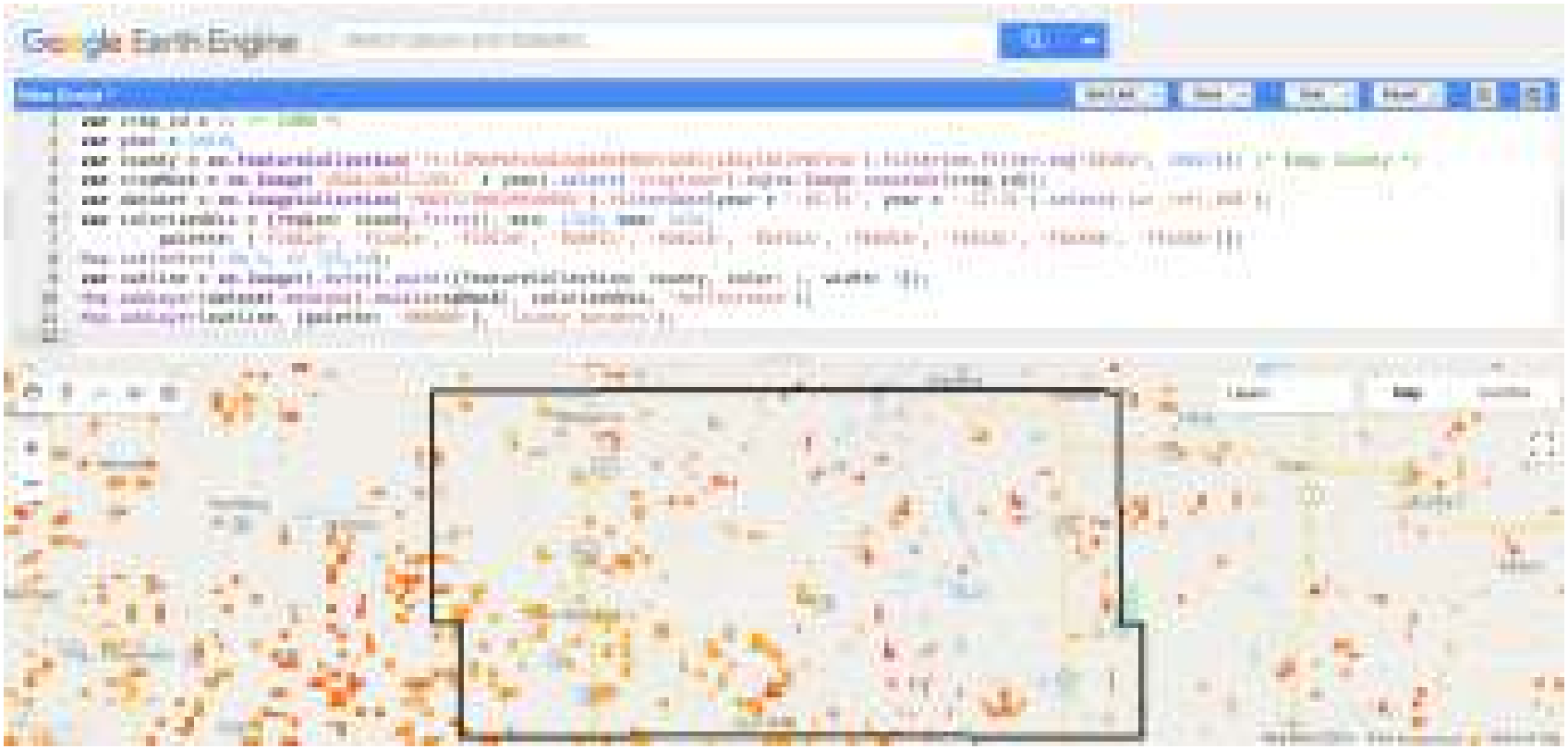


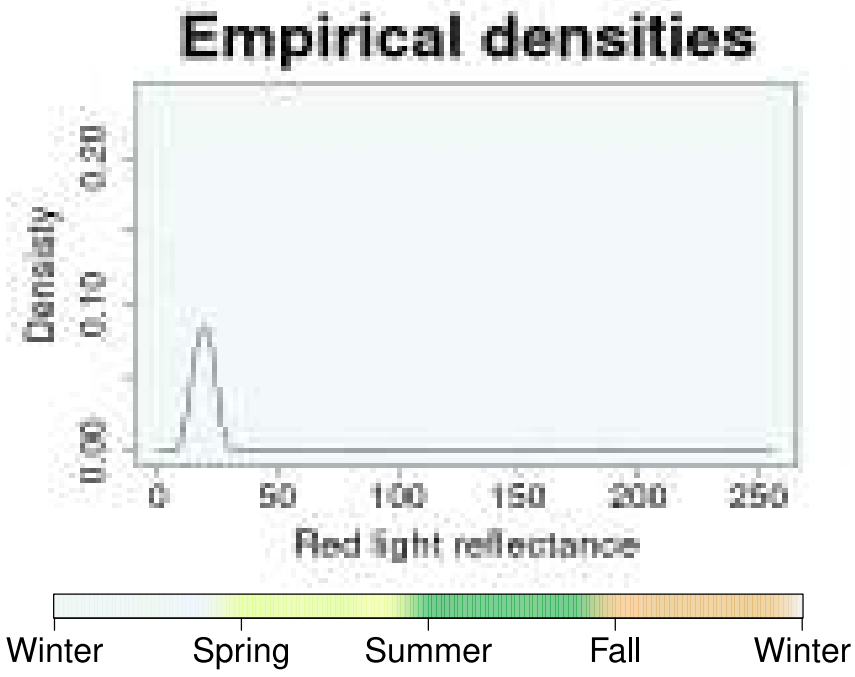
Figure: Median reflectance in 2018 over corn fields

[1] `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

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

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:

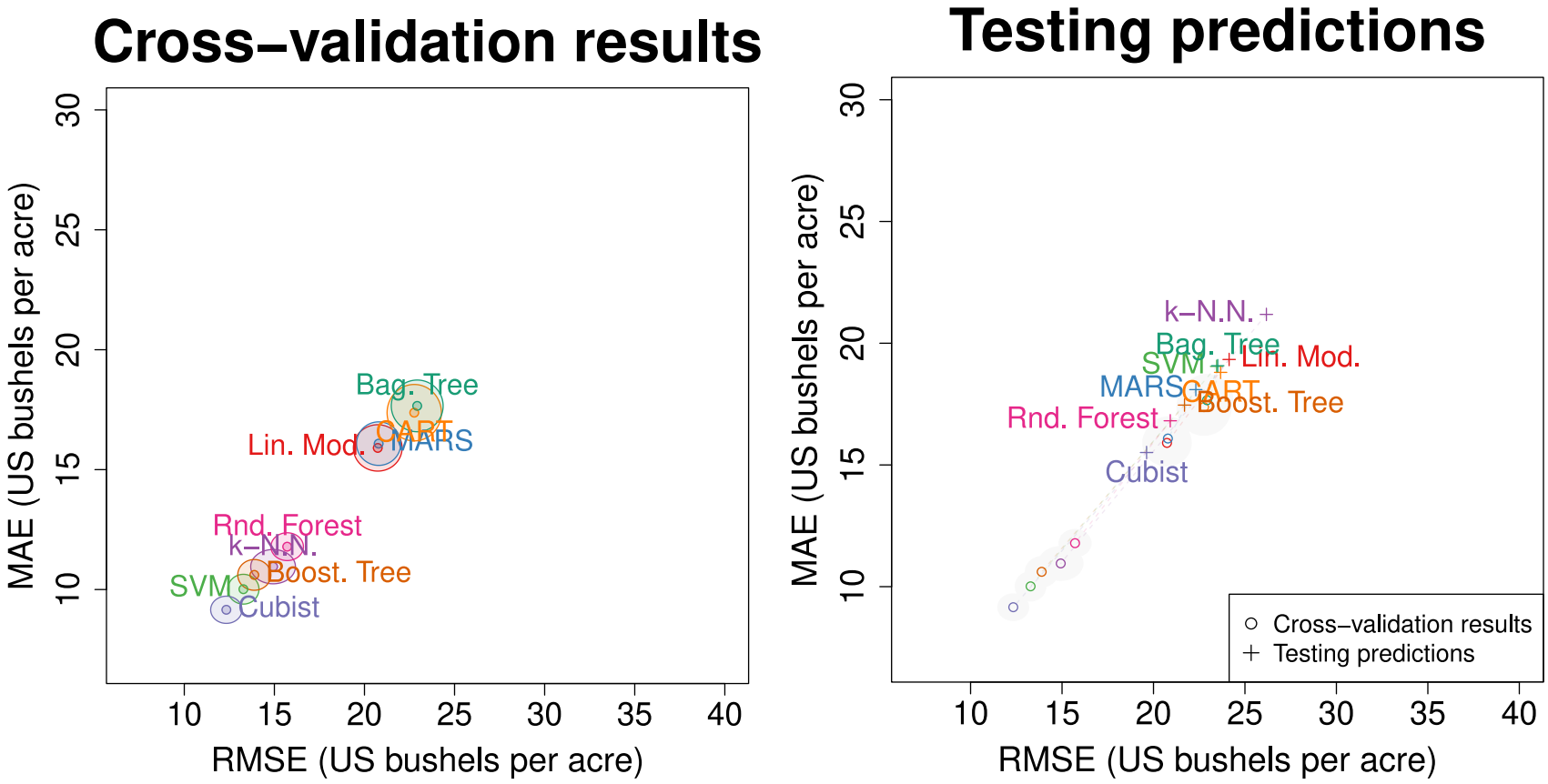
- 58 HPC nodes with 40 CPU cores and 128GB RAM
- 5 high-memory nodes with 120 CPU cores and 1.48TB RAM

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

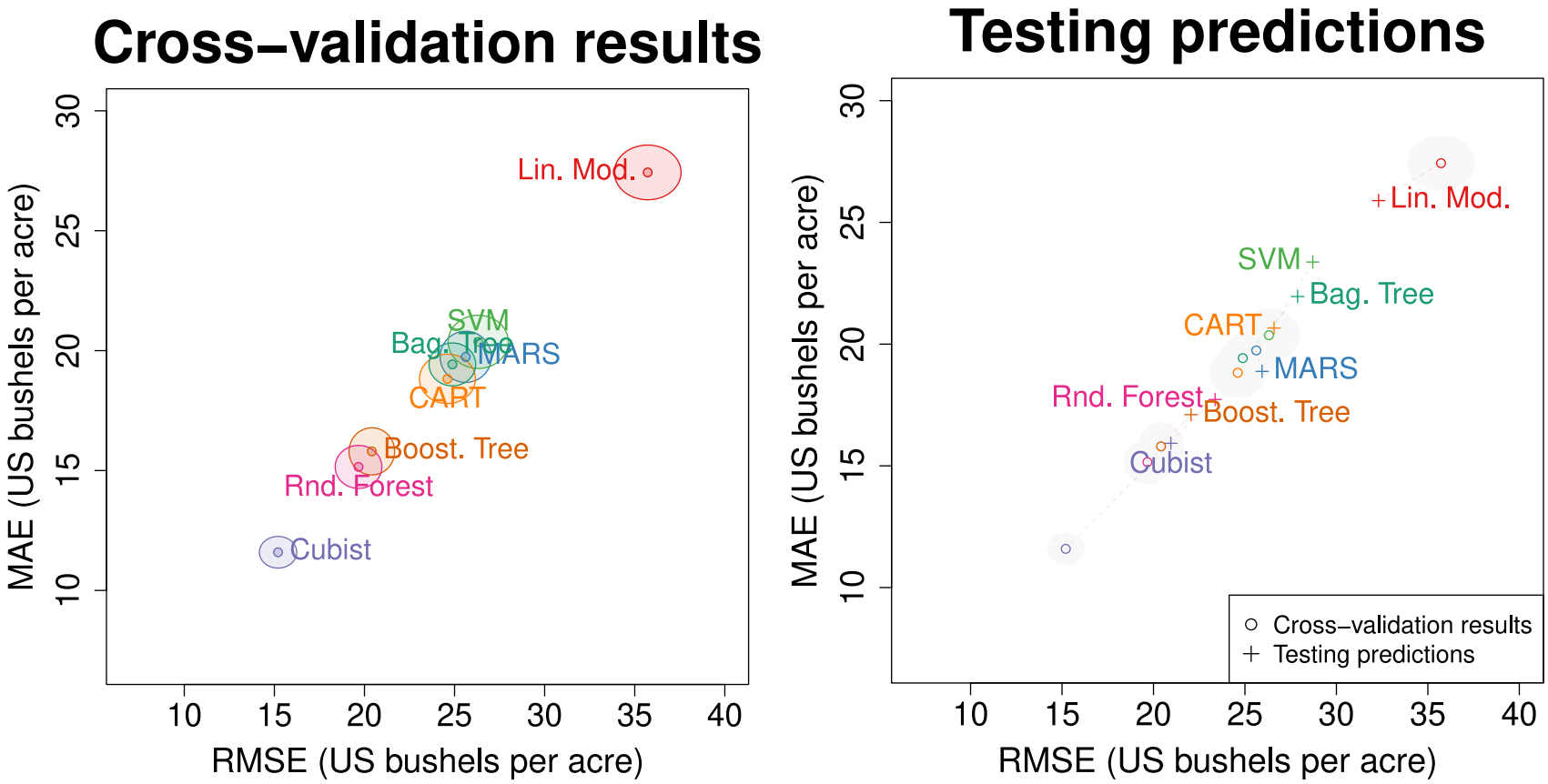
PART IV

RESULTS AND CONCLUDING REMARKS

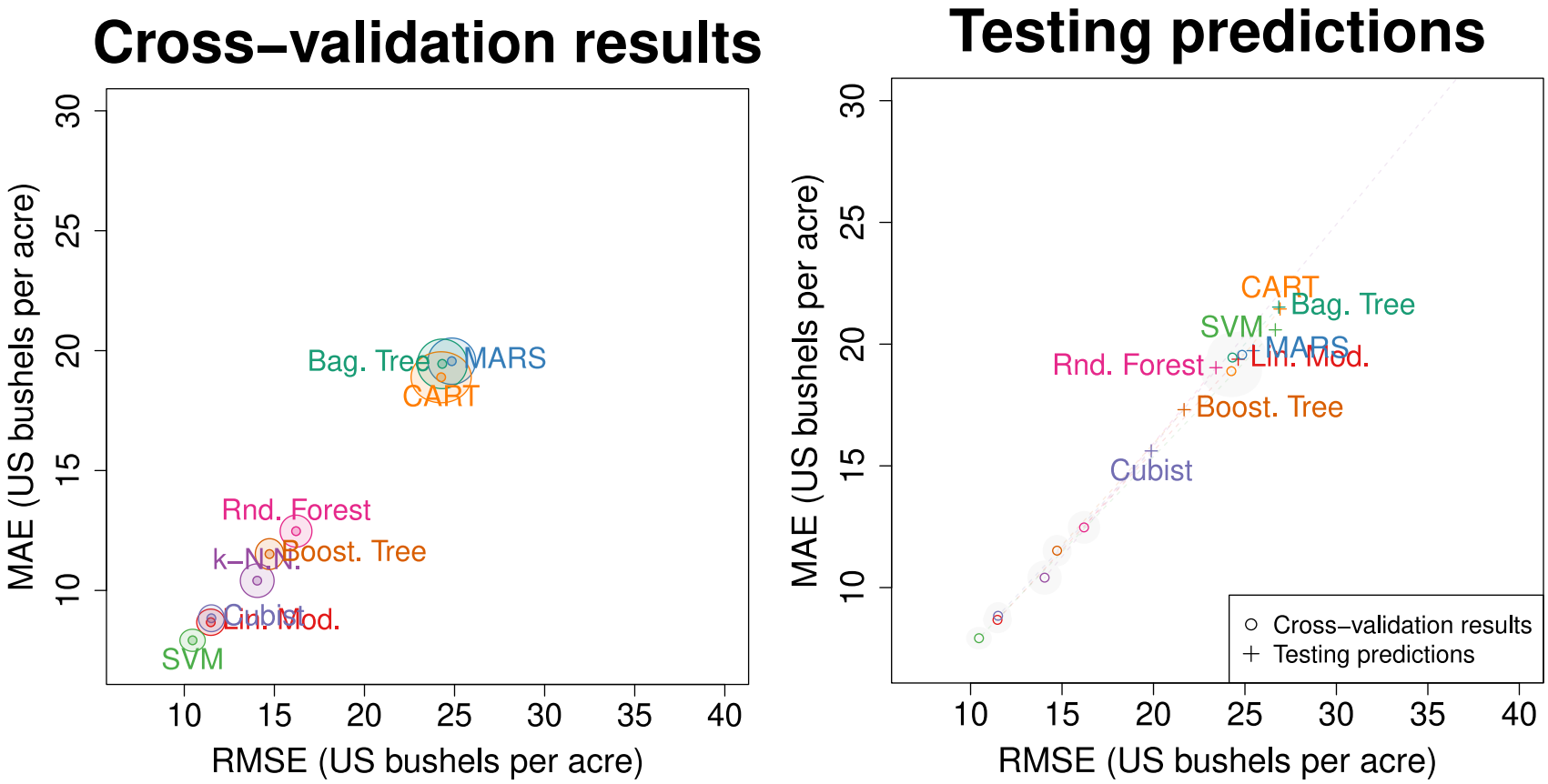
Multidimensional scaling



Principal components



Independent components



Performance evaluation of Cubist

	MDS	PC	IC
C.V. MAE	9.156	11.592	**8.840**
C.V. RMSE	12.333	15.205	**11.502**
C.V. MAE S.D.	0.289	0.337	0.285
C.V. RMSE S.D.	0.444	0.530	0.364
Test MAE	**15.511**	15.927	15.613
Test RMSE	**19.621**	20.950	19.877

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

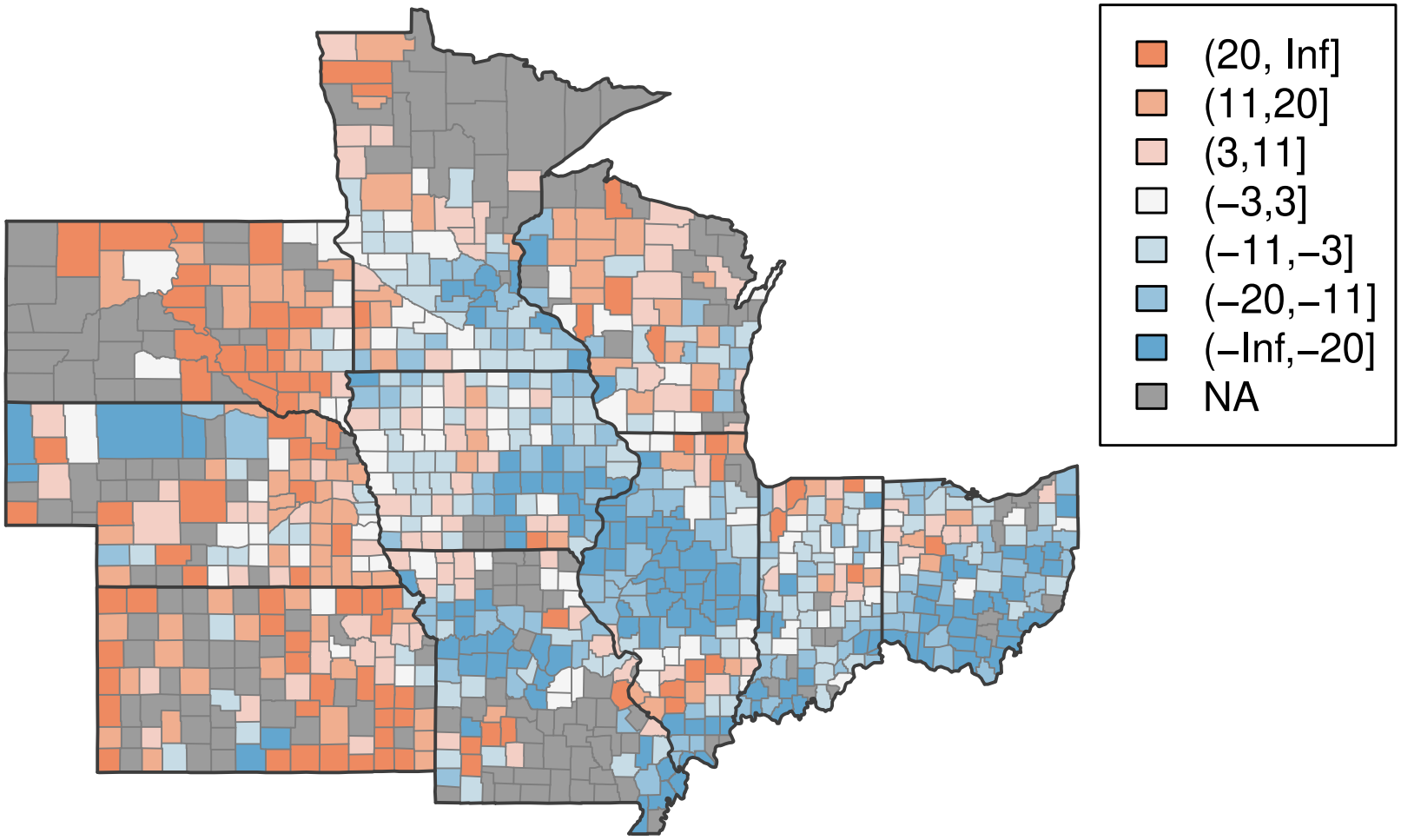


Figure: The map shows unexplained spatial dependence

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

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Friedman, J. H. et al. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 19(1):1–67.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27.

Hyvarinen, A. (1999). Survey on independent component analysis. *Neural computing surveys*, 2(4):94–128.

Johnson, D. M. (2016). A comprehensive assessment of the correlations between field crop yields and commonly used modis products. *International Journal of Applied Earth Observation and Geoinformation*, 52:65–81.

Jolliffe, I. (1986). *Principal component analysis*. Springer.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129.

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*, volume 26. Springer.

Navarro, J. A. (2017). First experiences with Google Earth Engine. In *GISTAM*, pages 250–255.

Rulequest (2006). *Data Mining with Cubist*. RuleQuest Research Pty Ltd., St. Ives, NSW, Australia.

Zhao, J., Shi, K., and Wei, F. (2007). Research and application of remote sensing techniques in chinese agricultural statistics. Beijing. Paper presented at the Fourth International Conference on Agricultural Statistics.

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