An open-source pipeline for remote sensing of crop yields under environmental change in Sub-Saharan Africa

Proposers and clients:

Tamma Carleton
Assistant Professor, UCSB
tcarleton@ucsb.edu
415-816-7870

Jonathan Proctor
Postdoctoral Fellow, Harvard University
proctor1@fas.harvard.edu
650-815-1054

Project objectives:

Food security in sub-Saharan Africa depends heavily on local agricultural productivity and is increasingly under threat from current climate variability and future global climate change (Porter et al., 2014). Despite the importance of this sector for wellbeing across the continent, agricultural productivity in most of sub-Saharan Africa is monitored and made publicly available only at the national level (here), and measured with substantial error (Lobell et al., 2019). Current attempts to leverage spatial statistics and/or satellite imagery to meet this data need are static in nature (e.g., here) and/or cover a small region only (e.g., Lobell et al., 2019). Thus, the imprecision and sparsity of these statistics substantially limits our understanding of how environmental changes influence agricultural productivity and food security within this region.

The core objective of this project is to fill this important data gap by making fine-resolution predictions of agricultural crop yields over time across all of sub-Saharan Africa using publicly available satellite imagery and an existing machine learning algorithm previously developed by the clients (Rolf et al., 2021). This new data product and its open-source codebase will be disseminated through an in-development API that is being built by the clients for related work.

A secondary objective of the project is the dissemination of an intermediary data product with the potential for widespread use in other research and policy applications. As discussed below, the Rolf et al. (2021) approach to summarizing the information within satellite imagery (i.e. generating features $x_1, x_2, \ldots, x_n$ from raw imagery) enables model training and prediction using linear regression and is agnostic to the prediction task at hand. Therefore, by "featurizing" publicly available satellite imagery over sub-Saharan Africa and releasing these features publicly, this project will lay an important foundation for predicting other key social and environmental indicators across the continent, such as irrigation, desertification, or malnutrition. The clients will ensure these "features", as well as the predicted crop yield metrics, are integrated into the public-facing and freely accessible API.

Significance of the project:

Due to limited agricultural data across sub-Saharan Africa, policy-makers, businesses, and researchers are currently unable to accurately forecast short-run food security risks posed by

weather events or to generate long-run climate change projections. For example, one of the clients is a co-author on a large-scale effort to project the impacts of climate change on global agricultural output using the largest subnational crop yield database ever assembled (Hultgren et al., *in prep.*). Figure 1 indicates that despite substantial data collection efforts, nearly all of Africa is missing from the database, limiting the authors' ability to conclusively assess the risk of climate change on food security across the globe's poorest continent.

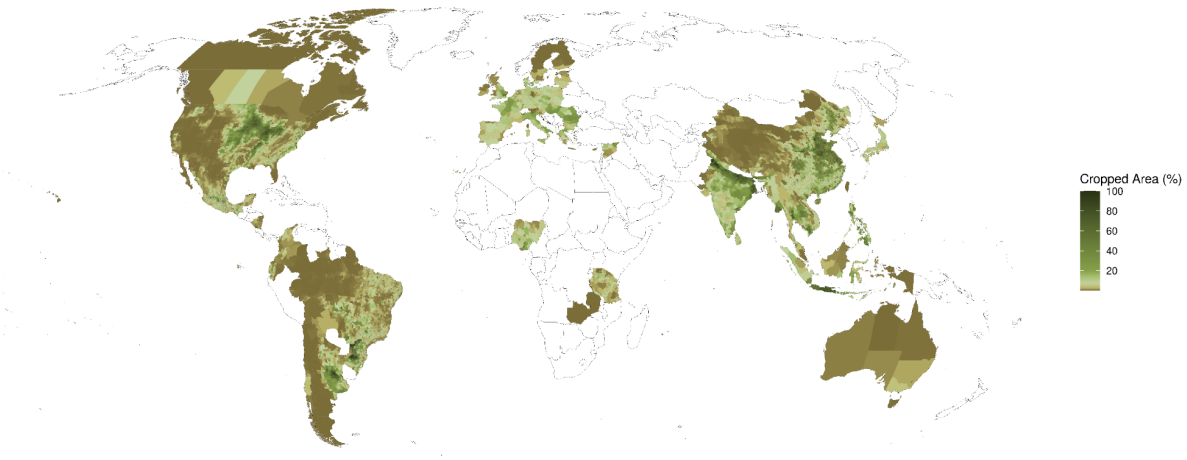

Figure 1: Recent data collection survey demonstrates dearth of agricultural data across Africa. Figure shows publicly available subnational staple crop yield data coverage from a recent large-scale survey. Countries with available subnational yield data for maize, soybean, wheat, rice, sorghum and/or cassava are colored, with subnational units colored by their cropped area in percent. All countries in white have no publicly available staple crop data, according to a recent data collection effort conducted by Hultgren et al. (2021).

Making measurements of agricultural productivity across sub-Saharan Africa would enable the study of how temperature, rainfall, and in turn climate change, may influence the region. Instead of exporting our understanding of how the U.S., Brazil, or Europe respond to changes in environmental factors, this data product would enable the research community to estimate effects specific to, and thus appropriate to, the African context. Such a dataset would be broadly useful within the research community.

This project sits within a broader research agenda aimed at democratizing access to social and environmental monitoring using satellite imagery. The clients are jointly pursuing this agenda as part of a team that developed a new unsupervised machine learning approach to featurize satellite imagery and generate predictions of a wide range of on-the-ground conditions. This system, called MOSAIKS, has the potential to dramatically lower the barriers to entry into the growing field of satellite-based prediction by creating a simple, computationally-cheap method for prediction in which users never need to access, manage, or process raw imagery.

The clients have developed this proposal with the aim of of extending the existing MOSAIKS architecture to a new, highly impactful context -- food security and environmental change in

sub-Saharan Africa. A valuable co-benefit of the project design is that the students will produce "featurized" versions of publicly-available Landsat imagery across sub-Saharan Africa over time. Essentially, the students will have collapsed the unstructured information contained within large quantities of imagery into a tabular dataset of geolocated features that can be used, in combination with linear regression, to predict a wide range of other outcomes, including but not limited to agricultural output. This has enormous potential to enable future researchers and data scientists to dramatically improve monitoring of many social and environmental variables over time across the world's most data-scarce region. As shown in the original MOSAIKS paper (Rolf et al., 2021), generating new predictions using existing features (i.e., those provided by the students on this project) takes on the order of a few minutes on a basic laptop. This innovation will dramatically lower barriers to entry into remote sensing in Sub-Saharan Africa.

Background

For decades, vegetation and crop health and productivity have been monitored using vegetation indices such as the normalized difference vegetation index, the enhanced vegetation index, and more recently solar induced fluorescence (e.g. Quarmby et al., 1993) . However, these have yet to demonstrate high accuracy in predicting yield, particularly across sub-Saharan Africa. Recently, machine learning approaches paired with high resolution imagery have sought to improve predictions with encouraging success (Burke and Lobell, 2017; Lobell et al., 2019) yet these cover only limited regions (e.g., one country), and have not led to publicly available datasets or open-source codebases for the broader research and policy communities to use. Our aim is to develop and make public such predictions and tools, at scale.

In a largely parallel body of research, machine learning approaches that don't adapt their features to the task at hand have generally failed to obtain performance competitive with deep learning methods in computer vision tasks such as satellite remote sensing. Rolf et al., 2021, however, demonstrates that models based on novel unsupervised featurization approaches can match the performance of deep learning methods. A secondary aim of this project is to test whether this high performance of the MOSAIKS features generalises to a new satellite source and region, namely Landsat data in Sub-Saharan Africa.

Equity

Satellite-based measurements have the potential to improve the development, implementation and enforcement of environmental regulations (Burke et al., 2021), yet low-income regions are consistently under-represented in both the production and consumption of remote sensing products. This occurs despite the fact that these regions are likely to benefit the most from such measurements, given the paucity of other systematically collected data (Yu et al., 2014 and Haack et al., 2016). In general, these inequities in access to the rich information contained within satellite imagery arise due to large barriers to entry into the remote sensing field, driven by high computational, data storage, expertise, and financial resource costs.

The MOSAIKS technology and its application in this project turn this model on its head by leveraging an unsupervised featurization algorithm that separates users from raw imagery. As documented in Rolf et al., this method lowers the computational cost of generating imagery-based predictions by many orders of magnitude, while being simple and easy to implement. This project will both make final output data available covering a region of the globe in which data are widely known to be scarce and of low quality *and* will release featurized images that will enable and empower people in data-poor regions to make predictions themselves for new tasks that are not studied here. Through these avenues, we hope to dramatically increase the equity of environmental monitoring processes and outputs.

Data

This project will combine publicly available satellite imagery from Landsat 7, as accessed via Google Earth Engine, with administrative data on crop yields from within and outside Africa, as collected and made publicly available by the clients.[1] The project will rely heavily on a featurization codebase written (along with colleagues) by the clients and made publicly available (in a cloud computing environment that can be tested by students or evaluators of this proposal) here.

Possible approaches

The general approach to this problem has already been developed and executed using different imagery and for different prediction tasks in Rolf et al., 2021. Therefore, there is a clear plan of action for the project, despite the context being quite novel. The key area of divergence from prior work is the temporal dimension -- here, we aim to create dynamic estimates of crop yield using Landsat 7 images over time, while no temporal analysis was done in Rolf et al. (2021).

The implementation steps are as follows:

1. Use the "random convolutional features" approach outlined in and coded up by Rolf et al. to featurize 8-day Landsat 7 satellite imagery over Nigeria, Tanzania, and Zambia (locations within Africa for which training data on subnational crop yields are available -- note that other countries' data are available and could be used for training if desired, such as Brazil, the United States, and India, among others).
2. Merge geo-located imagery features to administrative records of annual crop yields. Collapse 8-day features to annual measures using growing season masks (provided by the clients here) to include only growing season imagery. Note that there is likely more innovation to be done with respect to efficiently and effectively using the temporal dimension; this will be a key area of co-development between students and clients.

[1] Note that the linked repository for crop yield data excludes African countries as they had not yet been collected for the Proctor (2021) study. Crop yields for Nigeria, Tanzania, and Zambia from Hultgren, Carleton, et al. (2021) are stored on the Pod cluster at UCSB's Center for Scientific Computing. They are available for use but not yet linked in a public repository as Hultgren, Carleton, et al. (2021) is under review. Tamma Carleton can provide any additional details on the African data, including transfer of files to MEDS staff, if necessary.

3. Run cross-validated ridge regression to predict maize yield using the MOSAIKS features, pooling across all years.
4. Generalize the model to other staple crops after a pipeline for maize is built (time-permitting).
5. Apply random convolutional features across imagery covering all of sub-Saharan Africa, and use the prediction models built for each crop to generate image-level predictions of crop yield across the continent.
6. Work with the clients to integrate this output, and the intermediate features, into a public-facing API (already in development; API construction or maintenance does not need to be performed by the students as part of this proposal)

Deliverables

As discussed above, deliverables include:

1. Annual predictions of maize yield across Sub-Saharan Africa from 2000-2020 at 0.01 degree (~1km) resolution. Additional crop yields will be added if time permits. If computation at this scale is infeasible in the allotted time, annual predictions of maize yield in the three training countries (Nigeria, Tanzania, and Zambia) will be sufficient.
2. Annual averages of MOSAIKS features over the growing season across Sub-Saharan Africa at 0.01 degree (~1km) resolution. As above, if computation at this scale is infeasible in the allotted time, features computed in the three training countries (Nigeria, Tanzania, and Zambia) will be sufficient.
3. A clean, replicable, well-documented codebase that builds on existing MOSAIKS infrastructure (here).
4. [Time permitting] A report presenting a correlational analysis between estimated crop yields and high-resolution, publicly available climate indicators (i.e., temperature and precipitation).

Supplementary Materials

References

Burke, M., Driscoll, A., Lobell, D. B., & Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. Science, 371(6535).

Burke, M., & Lobell, D. B. (2017). Satellite-based assessment of yield variation and its determinants in smallholder African systems. Proceedings of the National Academy of Sciences, 114(9), 2189-2194.

Haack, B. & Ryerson, R. Improving remote sensing research and education in developing countries: approaches and recommendations. Int. J. Appl. Earth Observation Geoinf. 45, 77 (2016).

Hultgren, Andrew, Tamma Carleton, Michael Delgado, Diana Gergel, Michael Greenstone, Trevor Houser, Solomon Hsiang, et al. "The Impacts of Climate Change on Global Grain Production Accounting for Adaptation." *In prep.*, (2021).

Lobell, D. B., Azzari, G., Burke, M., Gourlay, S., Jin, Z., Kilic, T., & Murray, S. (2020). Eyes in the Sky, Boots on the Ground: Assessing Satellite‐and Ground‐Based Approaches to Crop Yield Measurement and Analysis. American Journal of Agricultural Economics, 102(1), 202-219.

Porter, J.R., L. Xie, A.J. Challinor, K. Cochrane, S.M. Howden, M.M. Iqbal, D.B. Lobell, and M.I. Travasso, 2014: Food security and food production systems. In: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change Field, C.B., V.R. Barros, D.J. Dokken, K.J. Mach, M.D. Mastrandrea, T.E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken, P.R. Mastrandrea, and L.L.White (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 485-533.

Proctor, J. (2021). Atmospheric opacity has a nonlinear effect on global crop yields. *Nature Food*, *2*(3), 166-173.

Quarmby, N. A., Milnes, M., Hindle, T. L., & Silleos, N. (1993). The use of multi-temporal NDVI measurements from AVHRR data for crop yield estimation and prediction. International Journal of Remote Sensing, 14(2), 199-210.

Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., ... & Hsiang, S. (2021). A generalizable and accessible approach to machine learning with global satellite imagery. Nature communications, 12(1), 1-11.

Yu, L. Meta-discoveries from a synthesis of satellite-based land-cover mapping research. Int. J. Remote Sens. 35, 4573 (2014).

Budget

This project does not require any additional funding. The only possible cost would be computational, as the students need to transform imagery into random convolutional features, which is a somewhat costly computational step (see Rolf et al., 2021 for a detailed breakdown of computation costs). All other computational steps in the analysis (i.e., geospatial merging, linear regression and prediction) do not require large computational resources (as documented in Rolf et al., 2021). This featurization step, however, can be conducted using UCSB's Aristotle cloud infrastructure, UCSB's Pod or Knot cluster, or directly on Google Earth Engine, where Landsat 7 images are already made available.

Client support

The clients commit to supporting MEDS students throughout the course of the capstone project by providing: input data (as described above); a well-documented existing codebase that can be used to directly build from for this new context; expertise in both agricultural systems and the computational infrastructure of MOSAIKS, an architecture the clients co-developed; and regular troubleshooting guidance and general project support.