Detecting irrigation phenologically

Problem statement

Irrigation represents 70% of global freshwater use and is simultaneously an important component of agricultural productivity and food security, and a driver of groundwater depletion, drying of rivers, soil salinization, and ecological degradation. In Brazil, the use of irrigation has increased over the past few decades thanks to the development of improved transportation networks. Our hypothetical client, the Brazilian government, would like to know where irrigation is growing fastest in order to provide infrastructural support to irrigation networks and to address water scarcity and environmental issues, such as accountability measures for agribusinesses that use irrigation.

Data

Training data come from the Brazilian Agricultural Research Corporation (Embrapa), which reports on the location of center pivot irrigation in 2014. This is combined with a map of agricultural areas from the Mapbiomas project, which allows us to randomly select 12,000 agricultural points across Brazil. Half of these agricultural training points are irrigated; the other half are not irrigated. For each training point, a variety of geographic, phenological, and spectral reflectance features were extracted. Geographic information included latitude, longitude, state, and region. Phenological and spectral reflectance information were extracted from cloud-filtered MODIS imagery obtained from August 1, 2013 to July 31, 2014, which represents the agricultural year. MODIS imagery is used to calculate the Enhanced Vegetation Index (EVI), a measure of greenness and a proxy for the degree of crop growth, every eight days over the agricultural year. Phenological data included the date of maximum EVI, the length of the crop cycle, the amplitude of the EVI curve, and the EVI value at the peak, minimum, and inflection points of the crop's EVI curve. These were calculated using a timeseries analysis on the EVI values through the agricultural year. Finally, surface reflectance information consisted of the timeseries in EVI every eight days; each observation day is assigned a MODIS observation number, with 1 corresponding to August 5, 2013 and 92 corresponding to July 28, 2014. Together, these features describe the timing and rate of growth of agricultural crops, which are expected to change under the presence of irrigation.

Data Cleaning

The training data are created on the cloud computing platform Google Earth Engine, and exported as a csv file for analysis in Python. These data are then gap filled and outliers are removed.

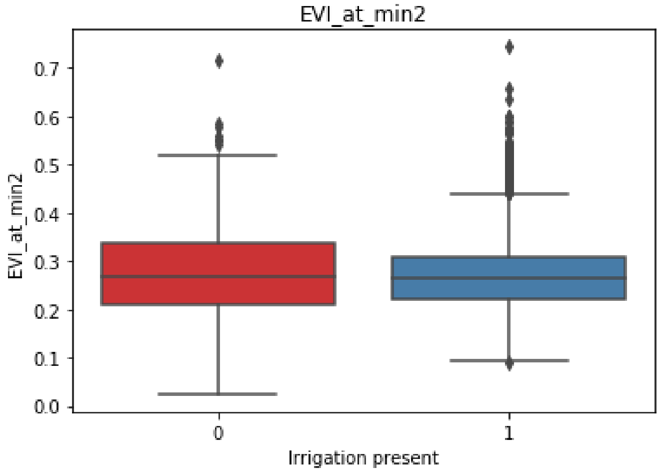
Gaps in the training data resulted from cloudiness in the MODIS imagery or from phenological dates that extended beyond the allowed agricultural year, which resulted in a NaN value for the EVI at that phenological date. These gaps were filled differently depending on the type of feature. Gaps in timeseries features (representing the EVI measured over the course of the agricultural year) were filled with linear interpolation using the closest non-missing points on the timeseries. Missing values in the phenological features were imputed with their mean. However, missing-ness in phenological features is likely meaningful for classification, we add a Boolean 'missingness' indicator for each phenological feature.

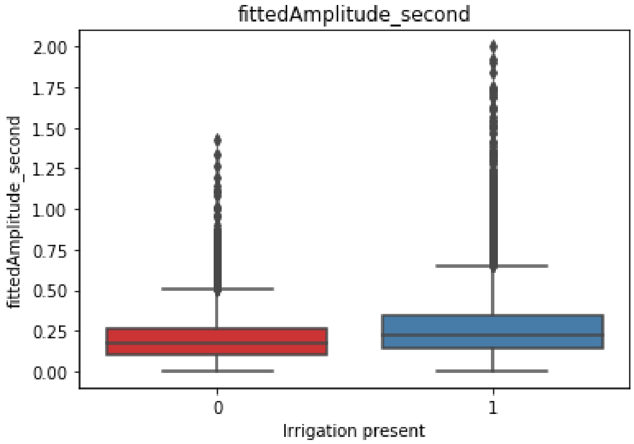
Missing-ness in the value of minimum EVI, for example, comes from a minimum EVI date that is estimated to be outside of the agricultural year, meaning that the timeseries analysis method used to estimate the date of the minimum EVI is inappropriate for the timeseries. This indicates that the point behaves differently from the assumptions made during timeseries analysis, a potentially useful piece of information.

Finally, outliers were eliminated from the phenological features based on reasonable limits. Points with EVI values that exceed 3 were eliminated because it is unlikely that an agricultural crop can exceed this greeness. Similarly, points whose crop cycle lengths exceeded 200 days and EVI amplitudes higher than 3 were also eliminated because they are highly unlikely for agricultural crops, and likely arise from error in the MODIS data or timeseries analysis.

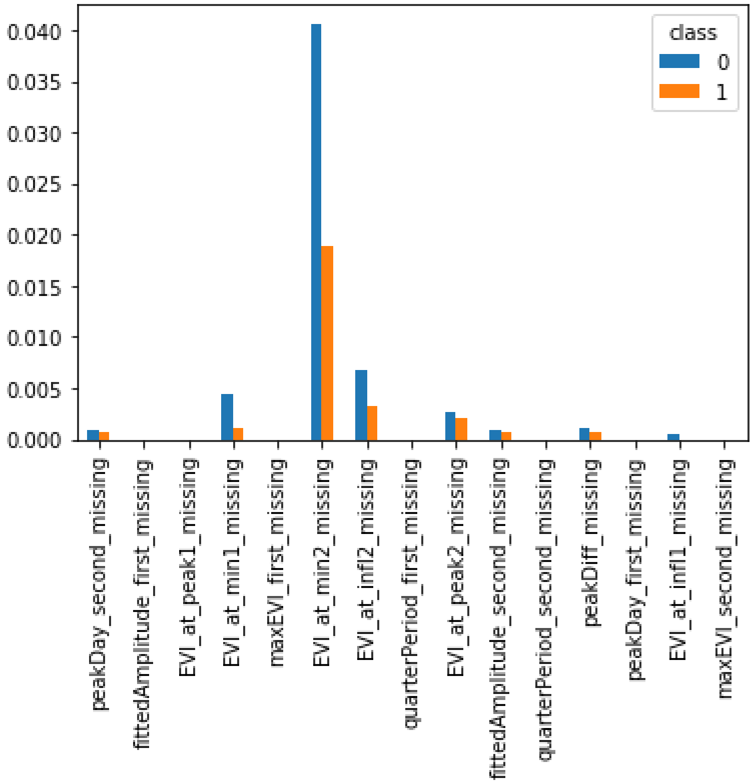
Exploratory Analysis

The goal of exploratory analysis is to develop intuition about the features that separate irrigated points from rainfed points. First, we use boxplots to find phenological parameters whose values differ for irrigated and rainfed points. The figures below show two phenological parameters: EVI_at_min2 (the EVI value at the minimum of the second crop in the agricultural year) and fittedAmplitude_second (the amplitude of EVI for the second crop in the agricultural year). In general, the phenological parameters that appear to differentiate irrigated from rainfed points capture the height and timing of the peaks. The phenological parameters that don't differentiate irrigated from rainfed points tend to correspond to EVI values at off-peak times. We can conclude that the most effective phenological parameters to separate irrigated from rainfed points will target the height and timing of the peak, not the cycle length or EVI values at off-peak times.

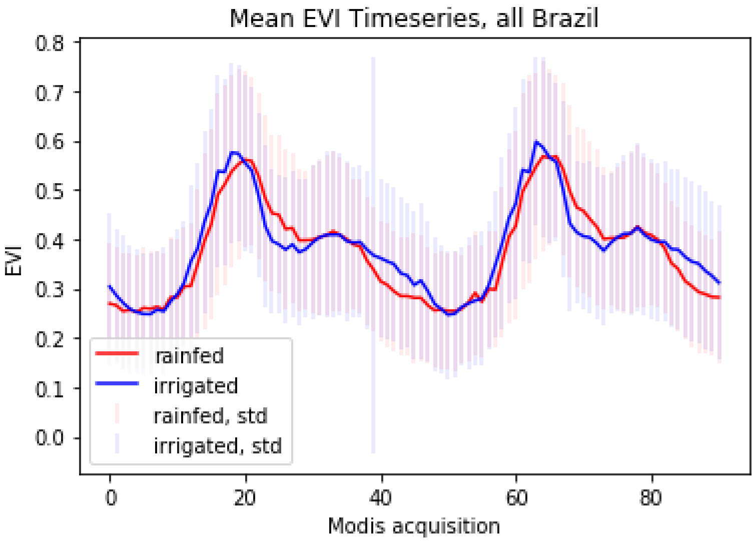




We can also look at whether the missingness of phenological parameters can help separate the two classes. As shown in the figure below, there are slightly more missing phenological values when there is no irrigation, so missingness may be an important predictor.

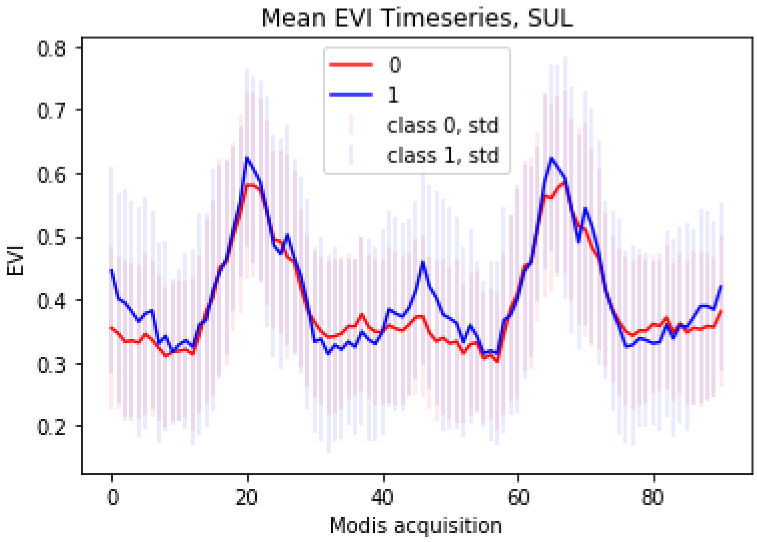


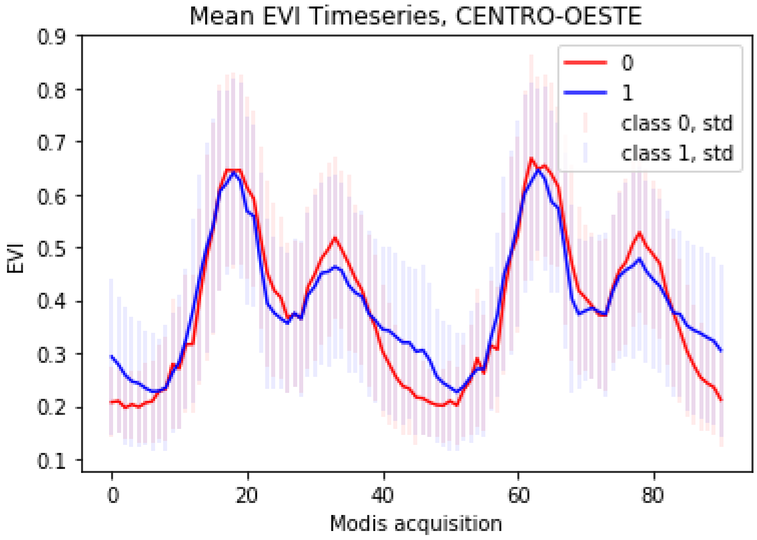
In addition to phenological features, another set of potentially important features is the EVI timeseries itself. Below, we compare the average EVI timeseries profile for irrigated and rainfed points across Brazil to highlight potential times of the year during which the two classes experience different EVI values. Here, time is displayed as Modis acquisition, but spans from August 1, 2013 to July 31, 2014.

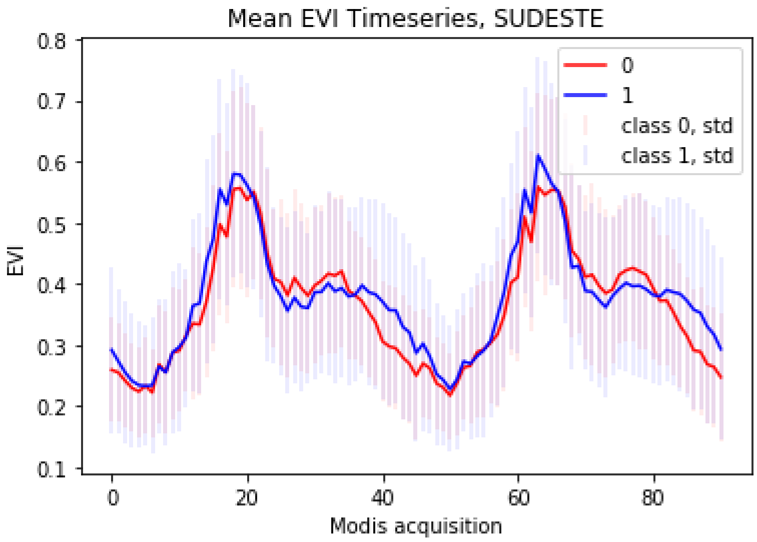


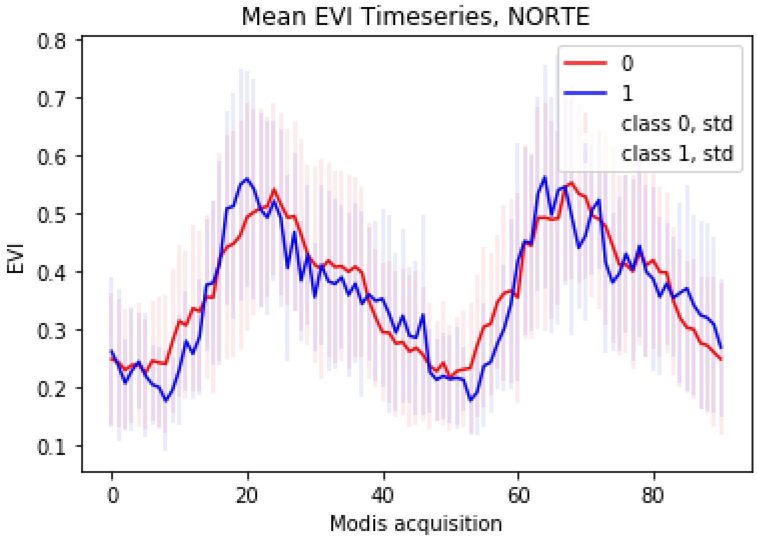
There is a slight difference in the EVI timeseries between irrigated and nonirrigated points. Irrigated points peak earlier and higher, and dip sooner for each peak. This timeseries indicates that certain acquisition times, corresponding to the peak and falling limb of each peak, are better able to distinguish between the two classes. Not all acquisition times are equally useful.

Next, we separate Brazil linto five regions and repeat the EVI timeseries plots. We expect that the different seasonality experienced in each region will impact the timing of agriculture, and therefore the times of year that best separate the two classes will differ by region. The plots below show that each region may have a different set of times that are best able to differentiate the two classes. In the south, southeast and center-west, the middle of the agricultural year (around modis acquisition 45) has higher EVI for irrigated points; in the northeast, the peak date is shifted for both the first and second peaks. This analysis shows us that each region should be treated differently because the phenological differences between the two classes differ in space.

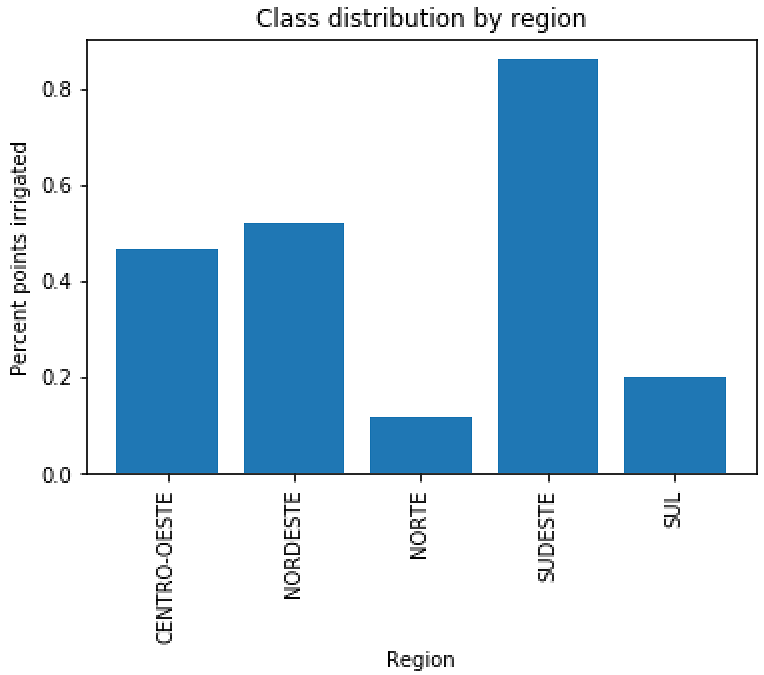


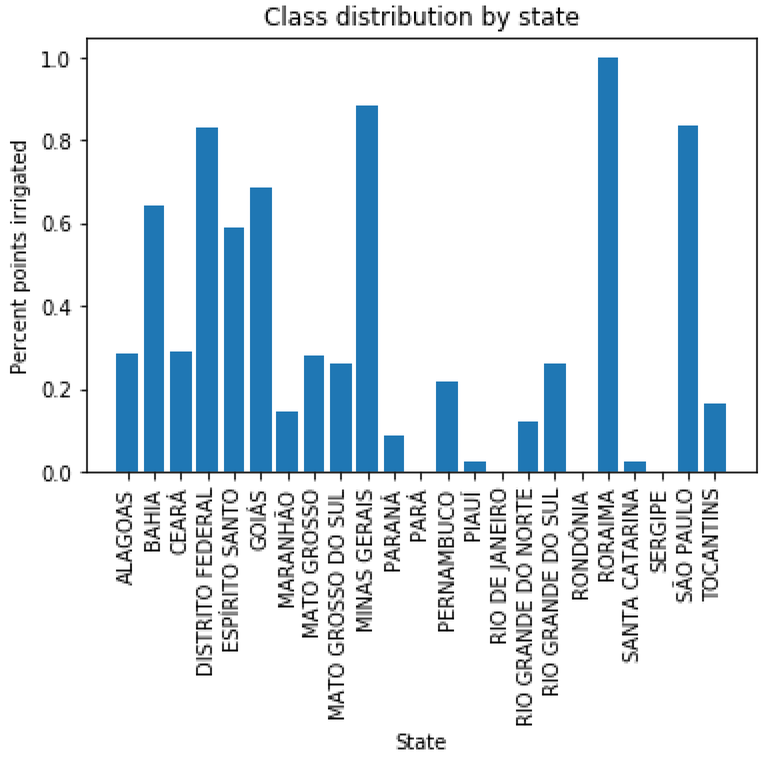


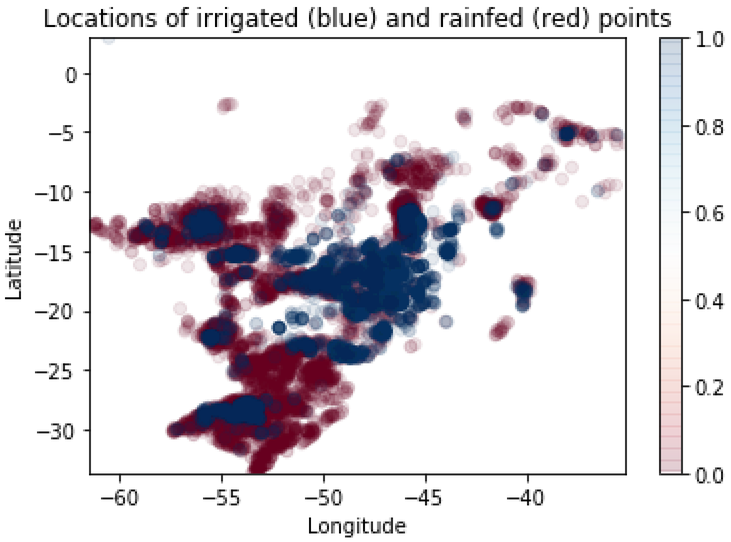




Since location is an important factor, we next look at the distribution of irrigation by location by region and state. The percent of training points irrigated in each area is summarized in the figures below. There is a stark imbalance in irrigation over Brazil, especially at the state level: on state, Roraima, has nearly 100% irrigation, while Para, Rio de Janeiro, and Sergipe have none. Including geographic features would therefore improve model performance, but would create a model that's less capable of predicting the spread of irrigation to new regions. In the final figure below, blue and red points indicate rainfed and irrigated training points, respectively.


Class distribution by region
Percent points irrigated
0.0
0.2
0.4
0.6
0.8
CENTRO-OESTE
NORDESTE
NORTE
SUDESTE
SUL
Region


Class distribution by state
Percent points irrigated
0.0
0.2
0.4
0.6
0.8
1.0
ALAGOAS
BAHIA
CEARÁ
DISTRITO FEDERAL
ESPÍRITO SANTO
GOIÁS
MARANHÃO
MATO GROSSO
MATO GROSSO DO SUL
MINAS GERAIS
PARANÁ
PARÁ
PERNAMBUCO
PIAUÍ
RIO DE JANEIRO
RIO GRANDE DO NORTE
RIO GRANDE DO SUL
RONDÔNIA
RORAIMA
SANTA CATARINA
SERGIPE
SÃO PAULO
TOCANTINS
State



A much higher percentage of points are irrigated in the midlatitude region than the north or the south. This could be back the midlatitude region has more established agriculture, whose managers can afford to invest in expensive irrigation equipment. The north has by far the lowest irrigation rate, likely because the wet season tends to last longer in the north (closer to the Amazon), and irrigation is less of a necessity. The northern area also has less established agriculture and less expansive transportation network, making it difficult to build irrigation structures. Since the physical location of the points impacts how the two classes can be distinguished, it will be important to keep in mind these regional imbalances in irrigation frequency.

Because the wet season in Brazil is quite long, there's not as much phenological or EVI difference between irrigated and nonirrigated pixels. In other words, this classification is much harder to do in Brazil than in the Sahara, where artificial water supply would make a large difference in crop growth. However, our exploratory analysis shows that EVI at certain times of the year and certain peak characteristics can help to differentiate the two classes. Geographic features would improve accuracy for predictions in the training year, 2014, but would make the model less relevant for predicting changes in irrigated locations over time.

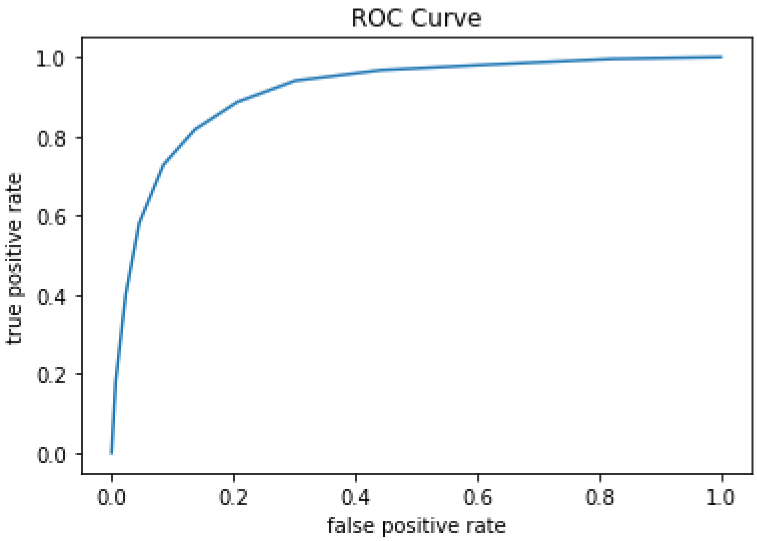
Modeling

Exploratory data analysis suggests that different regions have different phenological behavior, which may be best captured by a tree-based classifier. We test several specifications of random forest models, exploring the effect of excluding geographic predictors, tuning hyperparameters, tuning the classification probability threshold, and fitting separate models to each region of Brazil. In all following models, we use one hot encoding for categorical variables and a mean value imputer for missing numerical values.

Model 1

We begin by trying the simplest model possible: random forest model using all predictors (geographic, phenological, timeseries, and missingness features). A single model is fit to all points in Brazil.

We only tune one parameter, the classification probability threshold, to give a true positive rate of at least 0.85. This allows us to capture as much of the irrigation as possible, resulting in a conservative estimate of irrigated area. This comes at the expense of falsely classifying rainfed fields as irrigated, but ensures that the irrigated areas receive enough infrastructural support and regulation.

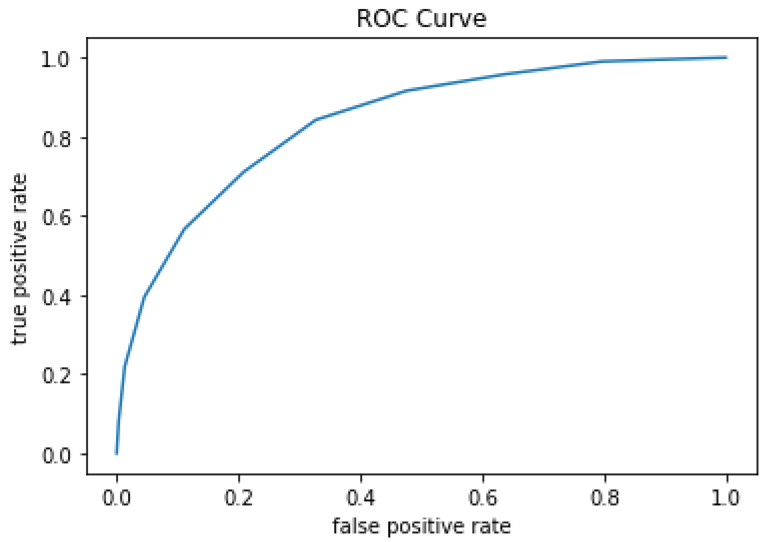


As shown in the ROC curve above, a threshold of 0.4 gives a high true positive rate of 0.88, and an acceptably low false positive rate of 0.21.

Because this model uses geographic predictors, it may be less relevant for predicting irrigation in the future as irrigated expands to new locations.

Model 2

In Model 2, we eliminate geographic predictors in the random forest model, but as with Model 1, train a single model for Brazil and only tune the classification probability threshold to maximize true positive rate. The ROC curve is shown below.



When we eliminate geographic features, our prediction accuracy lowers from 0.84 to 0.76. This decline may be worth the extra predictive power if irrigation expands geographically. A threshold of 0.4 gives a true positive rate of 0.84, and false positive rate of 0.32.

Model 3

In the third model, we tune multiple hyper-parameters using randomized search, including the missing value imputation strategy (mean or median), the number of estimators to use for each branch, the maximum depth of the trees, and the minimum samples per leaf. As before, we train a single model for all points in Brazil and exclude geographic predictors.
The best hyperparameters are:

- n_estimators = 1400
- min_samples_split = 5
- min_samples_leaf = 1
- max_features = auto
- max_depth = 90
- bootstrap = False
- imputer strategy =mean

The tuned model is now used to determine an appropriate classification probability threshold. A threshold of 0.5 gives true positive rate of 0.84 and false positive rate of 0.18. We find that accuracy of a model fit without geographic features increases to 0.83, compared to 0.76 without tuning. This performance is on par with the accuracy seen with the untuned model that uses geographic features.

Model 4

Although the use of geographic information may bias the model, it may still be useful to incorporate regional information. Because different regions have different precipitation seasonality, the Modis acquisition dates that best separate the classes are different, and it may be useful to fit a separate classifier per region. For Model 4, we use the tuned hyperparameters to fit a separate model for each region. Because each region has a different proportion of irrigated training points, we tune a separate classification probability threshold for each region. This allows us to keep the true positive rate steady at 0.85 for all locations.

The classification probability thresholds for each region that produce a true positive rate of about 0.85 are:

- south: 0.2
- center-west: 0.4
- southeast: 0.7
- northeast: 0.7
- north: 0.1

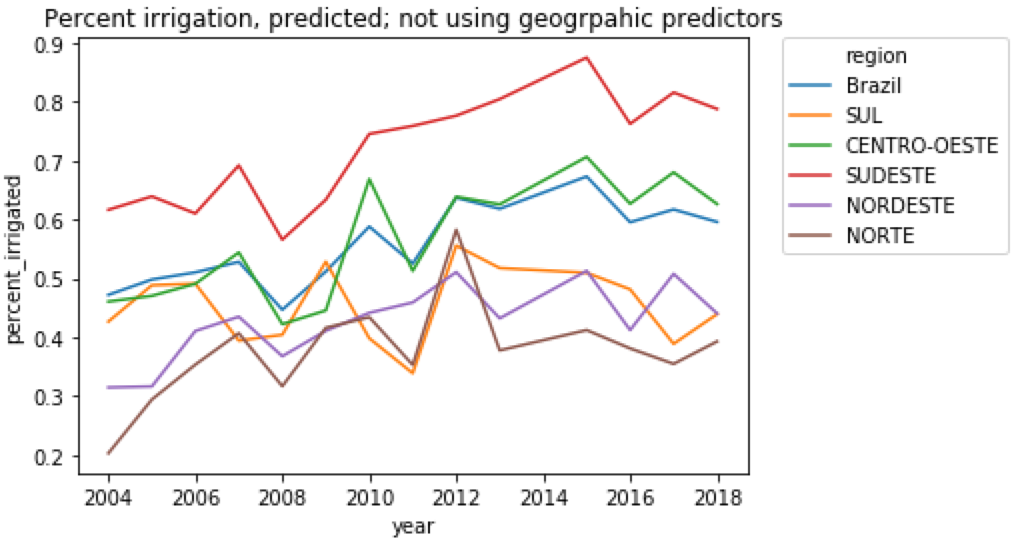
We use different thresholds for each region because the training points are imbalanced. The southeast has a very high proportion of irrigated points, so the threshold was set higher (at 0.7) to control the number of false positives; the north has a low proportion of irrigated points, so the threshold was set lower (at 0.1) to increase the true positives. These two thresholds produce the same false positive rate of 0.7 for the two regions. We choose to increase the true positive rate at the expense of increasing false positive rate for a conservative estimate of irrigation water use in different regions of Brazil.

Predictions

Predictions are made on the training point locations for 2004 to 2014. The client is interested in predicting the percent of agricultural fields that are irrigated in each of Brazil's five regions, which would allow them to target irrigation infrastructure and regulations regionally.

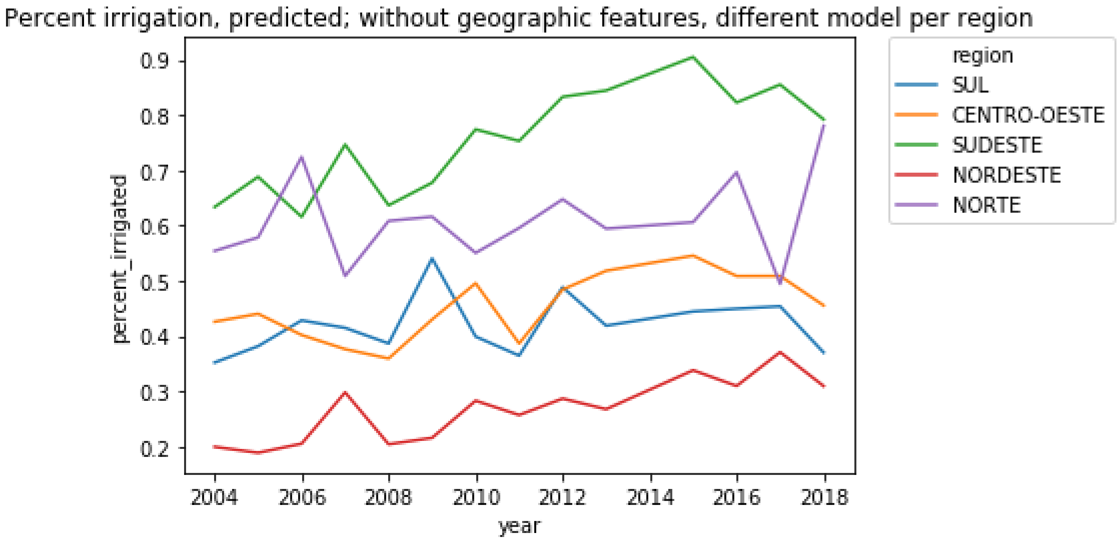
Predictions from Model 3

Predictions from Model 3, in which one tuned model is fit to all points in Brazil, is shown below. These predictions indicate that percent irrigation increased from 2004 to 2015 in all regions, then declined from 2015 to 2018 in most places. The spatial pattern in irrigation is also apparent: irrigation is very common in the south east and least common in the north. Both the north and southeast experience a large incerase in irrigated area in the 2000s, while irrigation stayed relatively constant in the south. There's bias in the percent irrigation that we predict: areas with high irrigation frequency in 2014 (southeast) are underestimated, and areas with low irrigated frequency in 2014 (north) are overestimated. However, we can still draw insights about the temporal trends within each region.



Predictions from Model 4

Predictions from Model 4, in which a different model is fit for each region, is shown below. While predictions from Model 3 and Model 4 produce different magnitudes for the percent irrigated within each region, they agree on the time trends: the northeast is still increasing irrigation, while the southeast is experiencing a decline in irrigation. The north and the south have relatively flat trends overall, but high interannual variability in percent irrigated.



Conclusions

We choose Model 4 as our final model for two reasons. First, it doesn't include geographical features, which allows us to predict pattern of irrigation spread in space and doesn't force the model to place weight on the spatial pattern observed in the training year. Second, the model still incorporates location information by fitting a separate model to each region. We are able to control for the imbalance in training data by allowing a different classification probability threshold in each region.

Region	Threshold	False positive rate	False negative rate	True positive rate	True negative rate	Percent irrigation in training data
South	0.2	0.42	0.25	0.75	0.57	20%
Center-west	0.4	0.54	0.12	0.88	0.45	47%
Southeast	0.7	0.70	0.09	0.91	0.30	86%
Northeast	0.7	0.57	0.14	0.85	0.42	53%
North	0.1	0.70	0.13	0.87	0.30	13%

The table above summarizes the accuracy metrics for each region. True positives are high for all regions, but comes at the cost of extremely high false positives, especially in the north and southeast. The cost of high falsely detected irrigation can be mitigated by conducting ground surveys of fields predicted as irrigated. Because these field surveys are expensive, they should be targeted in regions with a higher false positive rate, such as the north and southeast. Understanding the spread in false positive rates across Brazil can improve the design of ground validation surveys.