Remote Sensing of Environment 222 (2019) 303–317

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques

Sherrie Wang[a,b,*], George Azzari[a], David B. Lobell[a]

[a] Department of Earth System Science, Center on Food Security and the Environment, Stanford University, United States of America
[b] Institute for Computational and Mathematical Engineering, Stanford University, United States of America

ARTICLE INFO

Keywords:
Classification
Unsupervised learning
Agriculture
Landsat
Land cover
Machine learning
Google Earth Engine
Big data
Remote sensing

ABSTRACT

Crop type mapping at the field level is necessary for a variety of applications in agricultural monitoring and food security. As remote sensing imagery continues to increase in spatial and temporal resolution, it is becoming an increasingly powerful raw input from which to create crop type maps. Still, automated crop type mapping remains constrained by a lack of field-level crop labels for training supervised classification models. In this study, we explore the use of random forests transferred across geographic distance and time and unsupervised methods in conjunction with aggregate crop statistics for crop type mapping in the US Midwest, where we simulated the label-poor setting by depriving the models of labels in various states and years. We validated our methodology using available 30 m spatial resolution crop type labels from the US Department of Agriculture's Cropland Data Layer (CDL). Using Google Earth Engine, we computed Fourier transforms (or harmonic regressions) on the time series of Landsat Surface Reflectance and derived vegetation indices, and extracted the coefficients as features for machine learning models. We found that random forests trained on regions and years similar in growing degree days (GDD) transfer to the target region with accuracies consistently exceeding 80%. Accuracies decrease as differences in GDD expand. Unsupervised Gaussian mixture models (GMM) with class labels derived using county-level crop statistics classify crops less consistently but require no field-level labels for training. GMM achieves over 85% accuracy in states with low crop diversity (Illinois, Iowa, Indiana, Nebraska), but performs sometimes no better than random when high crop diversity interferes with clustering (North Dakota, South Dakota, Wisconsin, Michigan). Under the appropriate conditions, these methods offer options for field-resolution crop type mapping in regions around the world with few or no ground labels.

1. Introduction

Growing demand for food from an increasing global population necessitates close monitoring of agricultural activities, especially in regions of the world where food security remains elusive. Major steps along the road to closing the yield gap and achieving sustainable food security include accurately forecasting crop yields, understanding farm management practices, establishing links between crop choices and nutritional outcomes, evaluating the impact of changing policies and aid, and predicting with more certainty how climate change will affect agriculture. Helpful to all of these higher objectives is first knowing which crop types are growing in farmers' fields, not only at an aggregated regional level but at the level of the individual plot.

Traditionally, crop type information has been obtained from field surveys and censuses. This is true worldwide: in the United States, the Department of Agriculture's (USDA) National Agricultural Statistics Service (NASS) and Farm Service Agency (FSA) both collect field-level crop information (among much other data) via personal interviews with producers, and use it to monitor and forecast production throughout the growing season (Common Land Unit (CLU) Information Sheet, n.d.; June Area, n.d.). In Sub-Saharan Africa, crop type information included in the World Bank's Living Standards Measurement Study-Integrated Surveys on Agriculture (LSMS-ISA) Initiative has allowed for the study of how input use and soil quality affect yield and household income (Liverpool-Tasie et al., 2017; Bhargava et al., 2018). Such surveys, however, have their limitations: they are expensive and time-consuming to conduct, are therefore updated infrequently and cover small spatial extents in much of the world, and have been shown to contain biases due to flawed human recall (Carletto et al., 2015; Gourlay et al., 2017).

With the advent of accessible satellite data, many researchers see an opportunity to augment surveys or lessen the burden of data collection

* Corresponding author at: Department of Earth System Science, Center on Food Security and the Environment, Stanford University, United States of America.
E-mail address: sherwang@stanford.edu (S. Wang).

https://doi.org/10.1016/j.rse.2018.12.026
Received 9 May 2018; Received in revised form 23 November 2018; Accepted 17 December 2018

by creating low-cost crop type maps using features derived from remotely sensed spectral differences in vegetation and other surfaces over time (Waldner et al., 2015). These efforts have been aided by the success of machine learning methods such as support vector machines, random forests, and, increasingly, neural networks (Cai et al., 2018; Fu et al., 2017), which allow for large-scale automated analysis of satellite imagery. Some crop type maps created by previous works include the following:

- Global crop distribution maps like M3-Crops (Monfreda et al., 2008), MIRCA2000 (Portmann et al., 2010), and SPAM (You et al., 2014) report cropland percentages (out of total land area) in grid cells for major crop types in a single year. Each grid cell spans a 10 km × 10 km area, which is an improvement from coarser units such as countries and sub-national provinces. Other global crop distribution maps like CropWatch (Wu et al., 2015) report at a regional level. Both options, while potentially useful inputs or validation for creating high resolution crop type maps, trade resolution for scale and are alone too coarse to characterize individual fields.
- Country-wide and regional moderate resolution crop type maps created from publicly available satellite imagery such as those acquired by MODIS, Landsat, and Sentinel. Examples include the USDA's Cropland Data Layer (CDL) for the United States (Boryan et al., 2011; USDA National Agricultural Statistics Service Cropland Data Layer, n.d.); the Agriculture and Agri-Food Canada's Annual Crop Inventory in Canada (Fisette et al., 2013); since 2015, Sen2-Agri maps across Europe and in parts of Africa (the latter with low accuracy) (Inglada et al., 2015); and maps of a subset of crops, provinces, and years in China (Dong et al., 2016; Shao et al., 2001; Clauss et al., 2016; Liu et al., 2018), Brazil (Cohn et al., 2016; Zhong et al., 2016a), India (Heller et al., 2012), and Indonesia (Nuarsa et al., 2012). Depending on the size of crop fields in a region, these maps may enable the study of individual fields, and allow for the development of more detailed regional models. However, they are not currently available globally.
- High and very high resolution cropland and crop type maps based on satellite or aerial sources. Some examples include work in the US (Yang et al., 2007), Ethiopia (McCarty et al., 2017), and Turkey (Turker and Ozdarici, 2011), among others. These maps have the potential to capture detailed field-level information, but to date remain highly localized and challenging to create due to the large data storage and computational requirements. Global imagery is also not publicly available, as providers are primarily commercial or (in the case of UAVs) geographically limited.

In regions where maps of specific crop types are available across substantial areas and years, they exist thanks to large amounts of periodically collected ground annotations, which are needed to label the satellite imagery given to supervised machine learning algorithms. So while the resolution of maps has increased as researchers make use of ever better technology from the European Space Agency's Sentinel 2 (Inglada et al., 2015) to commercial satellites and UAVs (Yang et al., 2007), regularly-updated crop type maps have remained limited in geographic scope to the United States, Canada, parts of the EU, and other small pockets around the world. Even in the US, where data management and analysis are advanced, CDL exists across all states only back to 2008 (USDA National Agricultural Statistics Service Cropland Data Layer, n.d.), and in more idiosyncratic areas maps are made usually for a single year in which researchers obtained ground data.

For the foreseeable future, we will continue to lack timely and inexpensive ways of obtaining widespread ground data that can be used to label satellite imagery — especially in regions of the world where food insecurity makes having it most pressing. In this absence, two options exist to create crop type maps at national and sub-national scales: transfer supervised methods trained on available (and hopefully similar) ground labels, or turn to unsupervised methods, a class of machine learning algorithms that can learn structures from, and groups within, a dataset without labels.

Toward applying supervised methods in a data-scarce setting, some studies have used historical or little available within-year data to construct ideal crop phenologies or simulate training samples for classification. Hao et al. (2016a) used historical crop labels to extract ideal time series profiles for major crop types, and classified crops at 30 m resolution for the subsequent year. In a second work, Hao et al. (2016b) used historical labels again to nominate and label training samples for a more recent year. Ghazaryan et al. (2018) fit a harmonic regression to a small number of field samples' Landsat and Sentinel time series, generated simulated training samples with the aid of crop calendar information, and then trained a random forest on these simulated samples. These studies showed that in the regional settings of Xinjiang; Kansas, US; and Central Ukraine, respectively, supervised learning can classify new areas or years with high accuracy if given enough ground label coverage to accurately model or simulate crop phenologies.

While supervised methods such as random forests and SVMs have been used extensively to create crop maps, unsupervised methods remain relatively unexplored. Recent work by Gumma et al. (2016) and Xiong et al. (2017) used unsupervised classification with hundreds of clusters to map rice in India and crop types throughout Africa, respectively. These hundreds of clusters were then labeled with crops types under human supervision by matching cluster spectra to ideal crop spectra.

The sparse use of unsupervised methods is perhaps for good reason: these methods are often vulnerable to outliers, high dimensionality, and noisy features, and require the user to input information (such as number of clusters to search for) not known about the data (Tan et al., 2005; Hastie et al., 2009). But their advantage, if they can be implemented with success, is clear: we would gain the ability to create crop type maps where little to no field-level ground data exists — that is, globally.

This study develops and tests automatic methods for creating crop type maps in settings without field-level data by both transferring supervised learning across distance and time and using unsupervised learning guided by statistics on crop area. Our study area is the Midwestern US from 2010 to 2016, where ample high-quality validation data exists in the form of CDL. We used 30 m spatial resolution Landsat 5, 7, and 8 imagery with features computed using Google Earth Engine (GEE) (Gorelick et al., 2017). We then performed a Fourier transform, or harmonic regression (Ghazaryan et al., 2018; Jakubauskas et al., 2002), in GEE to extract features from Landsat time series at each pixel. Our novel contributions are threefold:

1. Using these harmonic features, we trained random forest models to classify crops in one state and applied them to eight other states for each year, and in one year and applied to six other years in each state. Results confirm that supervised methods generalize well within regions where crop compositions and phenologies remain similar, which we quantify through growing degree days (GDD), and not outside.
2. We used two unsupervised learning algorithms (k-means and Gaussian mixture model (GMM)) to cluster Landsat pixels into crop types. Results show that unsupervised methods, especially the GMM, can classify crop types with high accuracy if given appropriate features and low crop diversity.
3. Lastly, we used aggregate county-level crop statistics from the USDA National Agricultural Statistics Service (NASS) to automatically determine the number of clusters to search for, and to automatically assign crop types to clusters output by unsupervised methods.

While we recognize that the US Midwest is in many ways an ideal testing ground, our ultimate goal is to facilitate the development of crop type maps without a need for large quantities of field-level ground data

in areas of the world where their collection is prohibitively expensive and time-consuming, yet crop type maps would benefit regional and global food security. This large-scale case study is a step in that direction.

2. Data

In this section, we describe the dataset used, explain how we sample our data, and provide some summary statistics.

2.1. Study area

To thoroughly evaluate supervised and unsupervised approaches to classifying crop types, we tested these methods in the data-rich setting of the Midwestern United States. We performed our analysis for the years 2010–2016 on nine states covering an area of 1.62 million squared kilometers: Illinois, Indiana, Iowa, Nebraska, North Dakota, South Dakota, Minnesota, and Wisconsin (Fig. 1). We chose these states to cover the majority of grain-producing regions of the US and span a wide variety of growing conditions on which to test the scalability and robustness of our methodology. The majority of the study area grows corn and soybean, which together make up more than 74% of the crop area and are grown in highly overlapping regions throughout all nine states (USDA National Agricultural Statistics Service Acreage, n.d.). Corn is planted from late April through May and harvested from late September to early November, while soybeans are planted later from

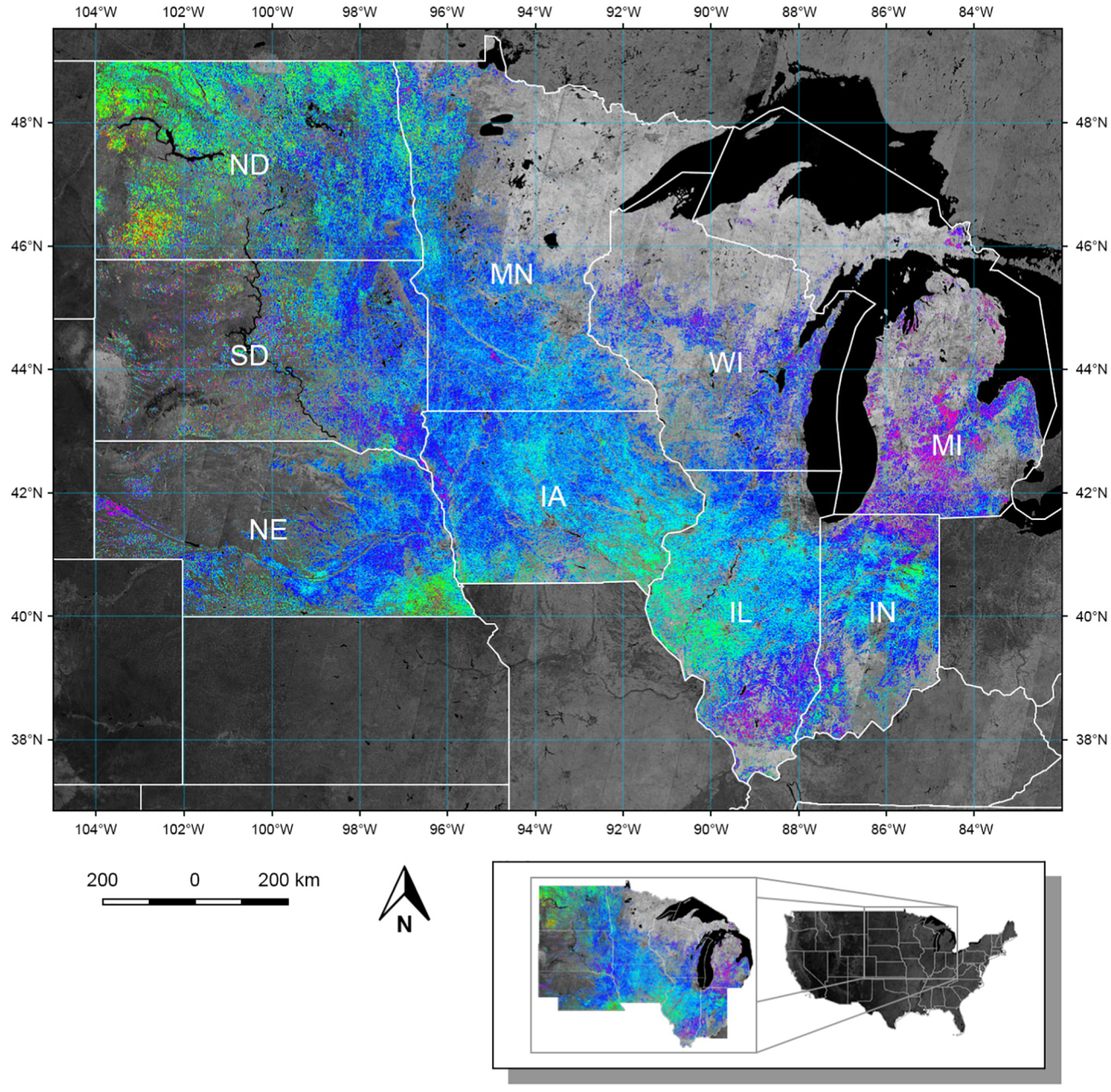

Fig. 1. The study area spans nine states in the US Midwest: Illinois (IL), Indiana (IN), Iowa (IA), Michigan (MI), Minnesota (MN), Nebraska (NE), North Dakota (ND), South Dakota (SD), and Wisconsin (WI). Harmonic coefficients fit to the time series of NDVI in 2016 are visualized in false color across the region's cropland using Google Earth Engine, and show phenological characteristics varying across space. Phase of the first-order harmonic terms is mapped to color hue, amplitude to saturation, and the constant term to value. The grayscale background shows NDVI computed from Landsat 8 cloud-masked median composites for the region in 2016. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

early May into June, and harvested from late September to late October (USDA National Agricultural Statistics Crop Progress and Condition, n.d.). Other crops including wheat and alfalfa are also grown in parts of the region. Spring wheat is grown in North Dakota, South Dakota, and Minnesota, and is planted from April to May and harvested from late July to early September (USDA National Agricultural Statistics Crop Progress and Condition, n.d.). Alfalfa is primarily found in North Dakota, South Dakota, Minnesota, and Wisconsin; harvest occurs from late May to early October (USDA National Agricultural Statistics Crop Progress and Condition, n.d.).

2.2. Remote sensing data

The Landsat Program is a series of Earth-observing satellites jointly managed by the USGS and NASA. We used Google Earth Engine to obtain imagery over our study region in the period 2010–2016 from Landsat 5, 7, and 8 Surface Reflectance Tier 1 collections (Gorelick et al., 2017). Images are moderate spatial resolution (30 m) and taken every 16 days by each satellite. The study region of 1.62 million km^2 corresponds to 1.80 billion Landsat pixels and a total of 23,166 images from January 1 to December 31 of all seven years. The January 1 to December 31 time window was used across all pixels to enable shared interpretation of harmonic regression coefficients (Section 3.2), and in the study region encompasses a single growing season for the majority of crop types. Yearly cloud-free image frequencies at each pixel are visualized in Fig. S1. The mean cloud-free image count at a pixel across all years is 25, with 2012 a notably lower year with a mean of 12. The dip in 2012 occurred because Landsat 5 Thematic Mapper ended operations in November 2011 and Landsat 8 was not launched until 2013. Our feature extraction procedure (Section 3.2) requires at least 5 points to fit a second-order harmonic regression and obtain coefficients.

Although the Landsat 5 TM, Landsat 7 ETM+, and Landsat 8 OLI instruments measure slightly different bands and wavelengths, they share six surface reflectance bands: blue, green, red, near infrared, shortwave infrared 1, and shortwave infrared 2 (Roy et al., 2014). From these, we derived ten vegetation indices (VIs) (Table S1) commonly used in the literature to characterize vegetation and cropland in particular: EVI (Liu and Huete, 1995), GCVI (Gitelson et al., 2005), MODCRC (Sharma et al., 2016), NBR1 (Key and Benson, 2006), NBR2 (Key and Benson, 2006), NDTI (Deventer et al., 1997), NDVI (Tucker, 1979), SNDVI (Huete, 1988; Rondeaux et al., 1996), STI (Deventer et al., 1997), and TVI (Broge, 2001). Since feature choice greatly determines the efficacy of machine learning algorithms, and we did not know *a priori* which bands and VIs would help discern crop types most, we explored a large number of VIs to see which perform best on the crop classification task, and selected a small subset of features for our models using the procedure described in Section 3.3.

Of the study region's 1.80 billion Landsat pixels, 674 million are classified as cropland by CDL on average between 2010 and 2016. This is a large dataset size on which to train random forests and perform clustering during method development; to reduce computational time, we sampled a subset of pixels uniformly at random within each state and year from regions indicated as cropland by CDL. The final dataset totals 912,081 pixels across all states and years. A summary of the dataset for 2016 can be found in Table 1.

2.3. Regional crop data

The USDA's National Agricultural Statistics Service (NASS) conducts hundreds of yearly surveys on a number of metrics, including food production, farm wages, and crop prices. We acquired data on the planted area of various crop types at a county level from the NASS Quick Stats database (USDA National Agricultural Statistics Service Quick Stats, 2018), and used this county-level aggregate data to determine which crops to classify in our supervised and unsupervised methods as follows.

Because (i) supervised algorithms tend only to perform well on predicting classes from which they have seen many examples, (ii) CDL accuracies are low for crops grown on small percentages of the cropland in a state, and (iii) unsupervised methods cannot recover a cluster when too few points belong to it, we only tasked our algorithms with classifying samples into the major crops of a state. A crop was considered *major* in a given region if it comprises over 10% of the crop area in said region. Smaller crop classes were grouped together into a single "other" class. Our supervised and unsupervised methods therefore classified crops into major classes plus the "other" class. Accuracies and other metrics were calculated with respect to these classes as well. The threshold of 10% was chosen based on experiments showing that crop clusters that comprised < 5–10% of our datasets could not be recovered by unsupervised algorithms.

Of the nine Midwestern states in our study area, all states except North Dakota and Wisconsin had only corn and soybean as major crops. North Dakota also produces a significant amount of spring wheat, and Wisconsin produces a significant amount of alfalfa.

Lastly, the NASS data also helps us assign crop labels to the clusters output by unsupervised learning. This procedure is described in Section 3.6.

2.4. Field data

CDL is a raster geo-referenced land cover map collected by the USDA for the entire continental United States (USDA National Agricultural Statistics Service Cropland Data Layer, n.d.). It is offered at 30 m resolution, so that each Landsat pixel has a corresponding CDL label, and includes 132 detailed class labels spanning field crops, tree crops, developed areas, forest, water, and more. In our dataset, 95% of crop pixels fall into one of six classes: corn, soybean, spring wheat, winter wheat, alfalfa, or non-alfalfa hay.

For the remainder of this paper, we treat CDL labels as ground truth and use them to evaluate the performance of various unsupervised and supervised methods. The quality of our evaluation thus depends on the quality of the CDL labels themselves. CDL is created yearly using imagery from Landsat and the Disaster Monitoring Constellation (DMC) satellites, and a decision tree algorithm trained on ground samples. The accuracy of CDL labels varies by class; for the top classes in our dataset, accuracies detailed in the CDL metadata exceed 95% (Table 1). Since the CDL decision tree is validated on randomly sampled fields and does not account for spatial correlation, we expect the reported accuracies to represent an upper limit for CDL (i.e. the true accuracy of the decision tree for regions lacking any ground samples would likely be less than reported).

3. Methods

For a graphical overview of the methods presented in this paper, see Fig. 2.

3.1. Image processing

Remotely sensed imagery often contains anomalous pixels due to atmospheric conditions or instrument malfunction. We removed low quality pixels in Google Earth Engine (GEE) by masking them out using the pixel Quality Assessment (QA) band provided with the Landsat 5/7/8 product. This QA band is generated by Landsat's Cloud Cover Assessment algorithm, which involves a series of spectral tests and a decision tree model (LDCM CAL/VAL Algorithm Description Document, n.d.). Pixels of an image labeled as "clear" by the QA band were kept; image pixels with all other QA values ("water", "cloud shadow", "snow" and "cloud") were removed and not used in subsequent harmonic regressions on the time series at a given location. The number of clear observations at each pixel is visualized in Fig. S1 for all seven years.

Once harmonic features were calculated and samples exported from

Table 1
Summary statistics of data in each state for the year 2016. Number of samples refers to the number of Landsat pixels sampled from the cropland of each state. The major crops are those that comprise over 10% of crop area in a state. The number of classes is the number of major crops plus one (for the "Other" class). CDL producer's and user's accuracies provided by NASS are printed for each crop respectively.

State	Num counties	Num samples	Major crops	Num classes	CDL producer's accuracy	CDL user's accuracy
Illinois	102	16,706	Corn, soybean	3	98.4%, 97.7%	98.6%, 97.9%
Iowa	99	21,872	Corn, soybean	3	98.5%, 98.0%	98.8%, 98.4%
Indiana	92	20,877	Corn, soybean	3	97.9%, 97.4%	97.5%, 97.7%
Nebraska	93	18,875	Corn, soybean	3	99.1%, 98.0%	98.3%, 98.6%
North Dakota	53	10,432	Corn, soybean spring wheat	4	94.6%, 96.6% 93.3%	95.6%, 94.6% 87.7%
South Dakota	66	9261	Corn, soybean	3	96.8%, 97.1%	96.0%, 96.1%
Minnesota	87	11,098	Corn, soybean	3	98.2%, 98.0%	98.2%, 97.0%
Wisconsin	72	10,894	Corn, soybean, alfalfa	4	97.2%, 95.0% 90.8%	95.5%, 95.9% 86.2%
Michigan	83	6548	Corn, soybean	3	95.7%, 95.0%	95.6%, 95.0%
Total	747	126,563	–	–	–	–

GEE (Section 3.2), we removed points more than 2.58 standard deviations (or 2% of samples) away from the data mean as measured by the Mahalanobis distance. Mahalanobis distance is the multi-dimensional generalization of how many standard deviations away from the mean a sample is. It is defined for a sample $\mathbf{x}$ as

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

where $\boldsymbol{\mu}$ is the mean of the dataset and $\mathbf{S}$ is the covariance matrix. We removed points where $D_M > 2.58$ because unsupervised methods such as k-means and Gaussian mixture model can be sensitive to outliers; they seek to minimize Euclidean distance or maximize likelihood, both of which can be severely influenced by single outlying points (Tan et al., 2005; Hastie et al., 2009).

Lastly, since k-means is affected by feature scaling, the data were normalized before being used as input to any algorithms. Within each state, the features were subtracted by their means and divided by their

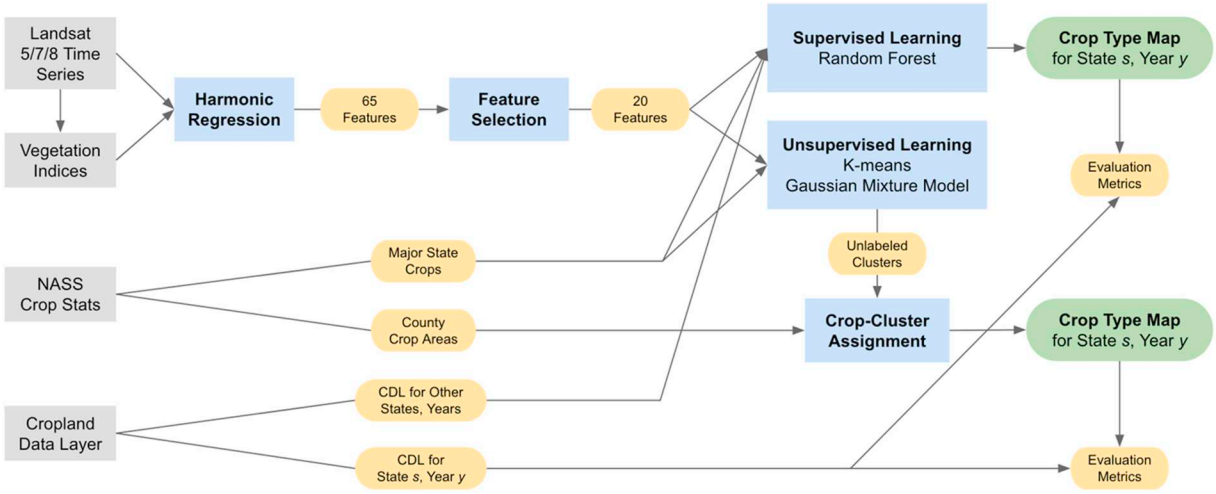

Fig. 2. A flowchart overview and example walk-through of the methods presented in this paper. (a) Explanatory diagram showing the workflow of our data sources, algorithms, and results. (b) An example application of the methods explored in this work shown for the region surrounding Decatur, Illinois, USA. (i) Landsat imagery of fields (RGB bands shown here) and vegetation indices derived from Landsat bands are input to harmonic regression. (ii) CDL pixel-level labels are treated as "ground truth" for training and validation. (iii) We extract pixel-level harmonic regression coefficients (shown in false color for NDVI mapped to HSV color space) from time series and use them as features in machine learning models. (iv) We transferred a random forest algorithm trained on nearby regions to classify pixels into crop types in this region. (v) In the absence of labeled pixels for transfer learning, k-means and other unsupervised methods can be used to cluster the pixels. Then, aggregate statistics can be used to assign crop labels. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

standard deviations. For supervised methods, the mean and standard deviation were calculated only on the training set; for unsupervised methods, the entire dataset was used to compute these statistics. We chose to process data at the state level rather than within agro-ecological boundaries (van Wart et al., 2013) for two reasons. First, for unsupervised methods, aggregate crop statistics are available for state and county boundaries rather than agro-ecological zones, making it more direct to cluster and assign crops to clusters at the state level. An early exploration of clustering within agro-ecological zones also did not show strictly better results (Section 4.3). Second, for supervised methods, ground labels, when available, are likely to be obtained within administrative boundaries rather than within agro-ecological ones.

3.2. Feature extraction

Since multiple Landsat images are taken over the same location over the course of a growing season, and the phase and amplitude of plant phenology can be used to differentiate crop types (Odenweller and Johnson, 1984; Foerster et al., 2012; Gomez et al., 2016), a method to extract features from discrete time series is required to apply machine learning models to temporal information. Parametric models that have been used to extract time series features and successfully classify crop types include the double-sigmoid (Zhong et al., 2016b) and harmonic regression (Ghazaryan et al., 2018; Jakubauskas et al., 2002). We opted for the latter, as harmonic regression has the advantage of being directly deployable in GEE using the built-in linear regression function, allowing for seamless large-scale application. The harmonic regression, or discrete Fourier transform, decomposes a function of time into its frequencies, yielding a compact representation of the time series of a pixel's reflectance or vegetation index.

We viewed the ith Landsat band and VI as a time-dependent function $f_i(t)$ and performed the harmonic regression

$$f_i(t) = c_i + \sum_{k=1}^{n} [a_{ik}\cos(2\pi\omega kt) + b_{ik}\sin(2\pi\omega kt)] \quad (1)$$

where a_{ik} are cosine coefficients, b_{ik} are sine coefficients, and c_i is the intercept term. The independent variable t represents the time an image is taken within a year expressed as a fraction between 0 (January 1) and 1 (December 31).

Unlike a continuous Fourier transform in which more harmonic terms result in a closer fit to the function, n must be chosen in the regression above to balance closeness of fit and overfitting to points in the sample. We picked $n = 2$ by minimizing mean-squared error on a hold-out set of points for time series across a variety of crops. We did the same for ω and found the best fit to our data when $\omega = 1.5$. The final regression at each Landsat pixel for band or VI i is therefore

$$f_i(t) = c_i + a_{i1}\cos(3\pi t) + b_{i1}\sin(3\pi t) + a_{i2}\cos(6\pi t) + b_{i2}\sin(6\pi t) \quad (2)$$

After performing this regression, we extracted coefficients a_{i1}, a_{i2}, b_{i1}, b_{i2} and c_i for 13 bands and VIs, giving us a total of 65 features on which to classify and cluster different crop types. Since the model has five parameters to fit, we needed at least five cloud-free observations at a pixel to extract coefficients. An example regression is shown in Fig. 3 for GCVI on a corn pixel time series.

3.3. Feature selection

The accuracy and computational cost of many machine learning methods suffer when the dimensionality (number of features) of the input data is high — this phenomenon is known as the "curse of dimensionality" (Hastie et al., 2009). Unsupervised methods such as k-means and GMM, which weigh all dimensions equally when computing the distances between samples and cluster centers, may be especially affected if some dimensions are irrelevant for discovering the true clusters. To reduce the downsides of having a large number of features without sacrificing information, we used a combination of time series correlation analysis, random forest feature importance, and univariate statistical tests with mutual information criteria to narrow down the 13 bands and VIs to four: NIR, SWIR1, SWIR2, and GCVI (Fig. S3). GCVI ($NIR/green - 1$) was designed to capture chlorophyll content in crops, and the NIR and green spectral ranges were chosen using a dataset of corn and soybean canopies (Gitelson et al., 2005), so it makes sense that GCVI differentiates these crop types better than other VIs. We chose feature selection instead of dimensionality reduction techniques such as PCA to retain interpretability of our models. For the univariate statistical tests, we used Python `scikit-learn`'s `SelectKBest` implementation (Pedregosa et al., 2011). Our subsequent results are shown for models trained on a total of 20 features derived from four bands/VIs.

Many of the bands and VIs computed are highly correlated. Fig. S4 shows that the difference in accuracies of random forests trained on all 65 features and on the subset of 20 is practically insignificant at only a couple of percentage points in all states. In Iowa and Wisconsin, the difference in accuracy is also not statistically significant at the 5% level upon resamplings of the training and validation datasets. This suggests that information loss from reducing feature dimensionality had little impact on the ability to distinguish crop types.

3.4. Supervised learning

For the supervised classification of crop types, we trained random forest models on the coefficients extracted from harmonic regressions on Landsat time series. Random forests are an ensemble machine learning method comprised of many decision trees in aggregate (Breiman, 2001), and offer great ease of use along with high performance. They are used frequently in the field of remote sensing to perform land cover classification and crop type mapping (Azzari and Lobell, 2017; Gislason et al., 2006), and have been shown to yield higher accuracies than maximum likelihood classifiers, support vector machines, and other methods for crop type mapping (Ok et al., 2012; Inglada et al., 2015; Gomez et al., 2016; Ghazaryan et al., 2018).

We used the default parameters of Python's `scikit-learn` (Pedregosa et al., 2011) package `RandomForestClassifier` with the exception of increasing `n_estimators` (the number of decision trees in the random forest) from 10 to 100 to reduce model variance. We observed significant increases in prediction accuracy up to 100 trees and not beyond. The final hyperparameters used are summarized in Table S2.

We first applied random forests within a state to provide an upper bound on the performance that we can expect to achieve by transferring supervised methods across geography and time, as well as with unsupervised methods. The samples were randomly split 80%–20% into a training set and a test set, respectively. We did not enforce a minimum distance between training and test samples, as our sampling coverage is quite sparse at 127k/674M ≈ 0.019% of the total cropland area.

Next, to understand the generalizability of these supervised models across geographic and temporal distance — and therefore whether they are appropriate to use in data-poor environments — we conducted the following experiments.

- To see whether models trained in one geographic region (e.g. one where we have data) transfer to other regions, for each year we trained a random forest on one state and tested it on the remaining eight.
- To see whether models transfer across years, for each state we trained a random forest on one year between 2010 to 2016 and tested ita on the remaining six.

3.5. Unsupervised learning

We explored two unsupervised learning methods for clustering the

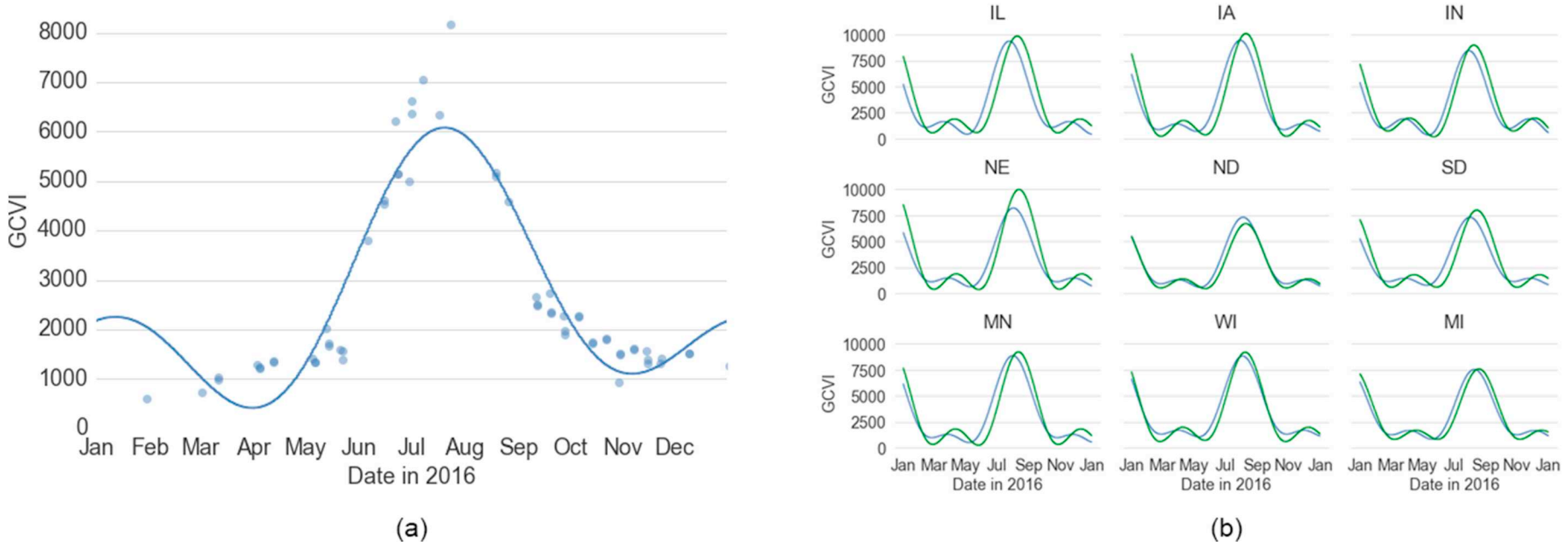

Fig. 3. Harmonic regressions for a single time series and averaged for crops across states. (a) Harmonic regression of GCVI values with $n = 2$ and $\omega = 1.5$ for one pixel of corn, sampled from Illinois in 2016. (b) Average GCVI harmonic regressions for corn (blue) and soybean (green) in the nine states of our study area. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

data in high dimensional space into crop types. Both clustering methods have Python `scikit-learn` implementations (Pedregosa et al., 2011).

- **K-means** partitions m samples into k clusters by alternately assigning samples to the nearest cluster centroid as measured by Euclidean distance and updating cluster centroids to the mean of samples assigned to the cluster. It is a common data analysis algorithm and has been applied frequently as a step within land-cover classification pipelines (Loveland et al., 2000; Xiong et al., 2017).
- **Gaussian Mixture Model (GMM)** is a probabilistic model that assumes samples are generated from k Gaussian distributions with unknown parameters. The expectation-maximization algorithm alternates between soft assignment (giving a probability of belonging to each Gaussian) of samples and estimating each Gaussian cluster's mean and covariance from its assigned samples. GMMs have been used for land-cover classification on remote sensing data in recent years (Ju et al., 2003; Tao et al., 2016).

Both k-means and GMM require the number of clusters k as input. We set k to be the number of major crops in a region — where "major" is defined as comprising over 10% of the region's total crop area planted — plus one "other" class to encompass the remaining crops. Note also that these clustering algorithms may find local rather than global optima and are dependent on how clusters are initialized. To ensure good clusters are found, we ran each algorithm 10 times and returned the clustering with minimum within-cluster sum of squares (k-means) or highest likelihood (GMM). The remaining hyperparameters we employed for these two models were the `scikit-learn` defaults, with the exception of the GMM covariance type, which we found to perform best when set to 'tied' (all components share the same covariance matrix). Clustering hyperparameters are summarized in Table S3.

Since the clusters output by k-means and GMM are not labeled with crop type, we evaluated the clustering portion of our unsupervised learning pipeline separately from the crop-to-cluster assignment portion by computing the classification accuracy of the *best possible assignment*. The best possible assignment is defined as the one-to-one matching of crop classes (including the "other" class) to clusters that results in the highest classification accuracy across test samples. For example, if we clustered the crops "corn", "soybean", and "other" into three clusters "0", "1", and "2", there are six possible assignments of crops to clusters. We iterated over the six possibilities and found the assignment that maximizes the overall classification accuracy. This was the *optimal classification accuracy* that we associated with the clustering algorithm in a state and year, and the metric with which we evaluated the suitability of k-means and GMM for separating crop types.

3.6. Assigning crops to clusters

The output of unsupervised learning algorithms is the assignment of every sample (here Landsat pixels) to a cluster. The clusters themselves are unlabeled; the task remains of assigning a crop class to each cluster. To do this, we again made use of the NASS data on crop area. We matched crops to clusters by finding the assignment that minimizes the discrepancy between the NASS county-level crop areas and the prediction made by the crop-cluster assignment. The particular metric we used in the loss function was the log-ratio of major crop areas. That is, for each state we found the assignment A^* that satisfies

$$A^* = \arg\min_{A \in \mathbf{A}} \sum_{k=1}^{|K|} w_k \sum_{(i,j) \in C_k} \left(\log \frac{z_{ik}}{z_{jk}} - \log \frac{\hat{z}_{Aik}}{\hat{z}_{Ajk}} \right)^2 \quad (3)$$

where $\mathbf{A}$ is the set of all possible crop-cluster assignments, K is the set of counties in a state, C_k is the set of major crop pairs in the kth county, z_{ik} is the area planted for crop i in county k reported by NASS, and $\hat{z}_{Aik}$ is the number of samples in our dataset belonging to crop i in county k under crop-cluster assignment A. Weights w_k are equal to the crop area in county k divided by the state crop area. We considered crop area percentages as an alternative option to log-ratios, but log-ratios have the advantage of not requiring complete crop data in each county to be accurate.

For example, if we clustered the crops "corn", "soybean", and "other" into three clusters "0", "1", and "2", we iterated over the six possible crop-cluster assignments and found the assignment that minimizes the squared error between the NASS-reported county-level crop area log-ratios and the predicted county-level crop log-ratios given the crop-cluster assignment. We weighed the squared error from each county by its fraction of total state crop area, so that the NASS data from counties with larger crop areas matter more in crop-cluster assignment. Counties with more crop area tend to have more complete data.

3.7. Quantifying differences among regions

To better understand why supervised and unsupervised algorithms perform differently in different states (Section 4), we used the following metrics to quantify differences between regions.

- **Growing degree days (GDD)** is a heuristic used to measure the accumulated heat experienced by plants over the course of a year and can be useful in predicting plant development. While most precisely defined as an integral of daily temperatures above a baseline temperature, we computed GDD for a single day using a discrete approximation given by the expression

$$GDD = \max\left(\frac{\min(T_{max}, T_{cap}) + \max(T_{min}, T_{base})}{2} - T_{base}, 0 \right)$$

where T_{max} is the daily maximum temperature, T_{min} the daily minimum temperature, T_{base} the baseline temperature, and T_{cap} the temperature at which the daily maximum is capped. We used the standard baseline temperature $T_{base} = 10$ °C and cap temperature $T_{cap} = 30$ °C, and summed GDD from January 1 to December 31 of each year to obtain the final amount of accumulated heat experienced at each sample's location. GDD data was calculated using daily temperatures from the National Aeronautics and Space Administration's (NASA) Daymet V3 dataset (Thornton et al., 2018), which provides gridded estimates of daily weather parameters for North America at 1 km × 1 km spatial resolution.
The accumulation of GDD during the season varied widely for our region, from 1100 to 2600 GDDs (Celsius). In the supervised learning results (Section 4.2), we used GDD to split the study area and explain the transferability of random forests across regions. We binned samples across all states in each year by their GDD, dividing the range 1100 to 2600 into ten evenly spaced bins of 150 GDDs each. A random forest was trained on 80% of samples in each bin and tested on the remaining 20%, as well as on samples from the other eight bins. All bins had at least 500 samples, and we classified samples only into three crop classes that were represented in all bins: "corn", "soybean", and "other".
Then, for unsupervised learning results (Section 4.3), data from each state was divided into 20 equally sized bins by GDD, and a separate clustering algorithm was run on each GDD bin to characterize the conditions under which clustering succeeds or fails.
- **Shannon entropy** is a measure of the disorder or uncertainty of a probability distribution, defined for a discrete probability mass function as

$$H = -\sum_i p_i \log p_i$$

where p_i is the probability of event i occurring. The smallest possible value of H is zero, obtained when the probability distribution has a weight of 100% on a single outcome. Entropy increases as the

number of possible outcomes increases or if their probabilities are more uniform. We computed the entropies of the crop distributions in each state, in which p_i is the proportion of cropland area of crop type i as reported by the NASS dataset.

4. Results

We present the performance of our supervised and unsupervised methods on the crop type classification task, along with analyses to understand conditions under which these methods perform well.

4.1. Harmonic coefficients for crop classification

Harmonic regressions with $n = 2$ closely approximate the time series of Landsat reflectance and derived VIs over the course of a year (Fig. 3). Across all years, states, and bands of the study area, the median number of points in a time series is 25 and the median R^2 of the regression is 0.88. From Fig. 3, it is also qualitatively apparent that harmonic coefficients can capture different crop phenologies. Average harmonic regressions for corn and soybean indicate that, while the curves are temporally close, corn is consistently sown and harvested earlier than soybean. This observation is consistent with known sowing and harvest dates for these crops in the US.

A more quantitative gauge of how well harmonic coefficients can distinguish different crop types is their performance on within-state and within-year out-of-sample crop classification. These test accuracies are shown in the dark blue bars in Fig. 4 for the year 2016, using coefficients extracted from NIR, SWIR1, SWIR2, and GCVI bands. Results are similar for 2010–2015. Across six of the nine states, accuracies exceed 87%, while North Dakota, Wisconsin, and Michigan have lower accuracies of 82%, 80%, and 80% respectively. A classifier that always predicts the majority class would achieve accuracies of 56%, 58%, 51%, 58%, 33%, 37%, 43%, 51%, and 38% (on average over all years) in Illinois, Iowa, Indiana, Nebraska, North Dakota, South Dakota, Wisconsin, and Michigan, respectively.

The high classification performance of random forests indicates that the harmonic coefficients are excellent features for distinguishing crop types in the study area. They successfully capture the difference in phase and amplitude of crop phenology among corn, soybean, and other crops in the relevant bands and VIs to separate them. Difficulties in North Dakota, Wisconsin, and Michigan may be caused by lower CDL accuracy in these states (Table 1), as well as indicate that harmonic coefficients are worse at differentiating crop types in these states.

We note that we also tried using percentiles (5%, 25%, 50%, 75%, and 95%) of annual observations on the same 13 bands and VIs as features during model development, and found both supervised and unsupervised models performed worse. More details on how the percentile features were computed can be found in Supplementary materials Section S1.1. Harmonic features appear to be a good candidate for the task of crop type mapping, and this method of extracting features from remote sensing time series may also perform well in other applications.

4.2. Transferring supervised models across geography and time

To understand whether supervised models trained on a different geographic region can be used to classify crops in a region lacking ground data, random forests were trained on data from one state and applied to that of eight other states; results are shown in the light blue bars in Fig. 4 (a) for the year 2016. For most training states, the out-of-state test accuracies are near or over 70%, with North Dakota being a noticeable exception. In particular, models trained on any state besides

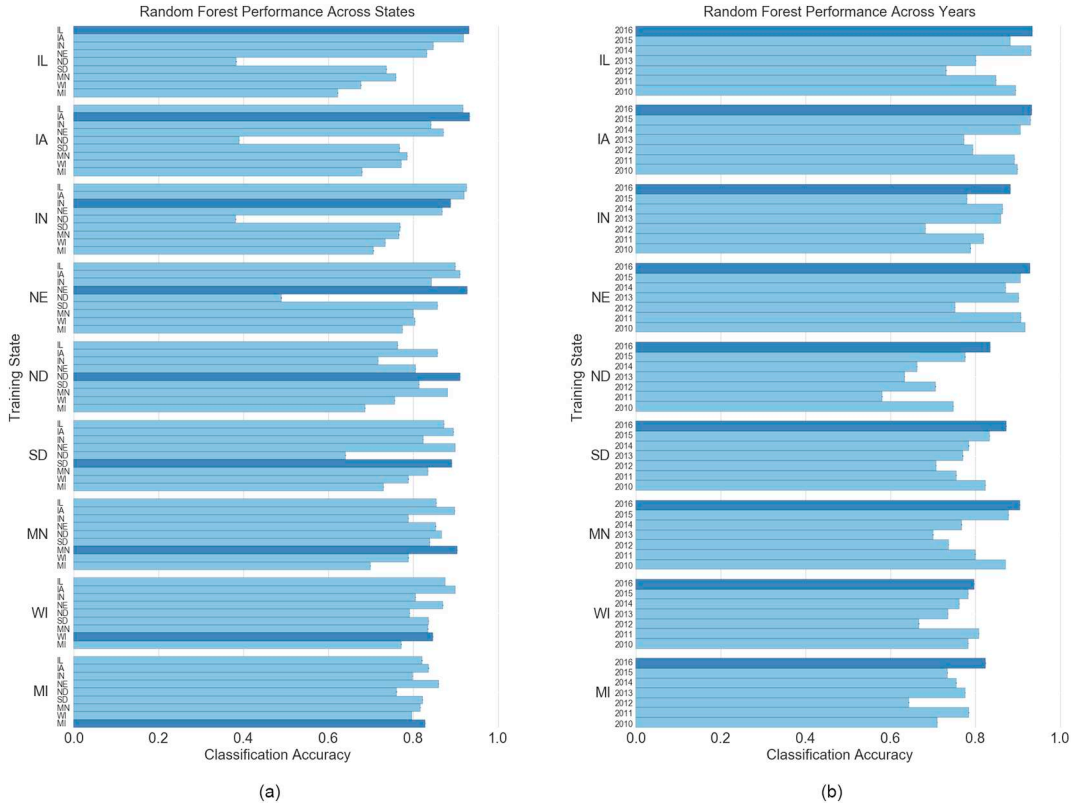

Fig. 4. Performance of random forest models transferred across states and years. (a) Out-of-state random forest performance for the year 2016. Each set of bars was trained on data from the same state, and each bar represents a different test state. Darker bars denote out-of-sample test accuracy within the same state as the training state. (b) Random forest performance across years 2010–2016, trained on 2016. Each set of bars was trained on data from the same state, and each bar represents a different test year. Darker bars denote out-of-sample test accuracy within the same year as training.

North Dakota apply with accuracies at or exceeding 80% to Illinois, Indiana, Iowa, and Nebraska. Models trained on any state besides Minnesota (especially states in the south of our study area) transfer poorly to North Dakota.

Next, to see whether models generalize across time, random forests were trained on a state in 2016 and tested on the same state for years 2010–2015 (Fig. 4 (b)). Cross-year classification accuracies are generally over 70%, with North Dakota again being an exception, implying high year-to-year variability in this state.

Given the phenological focus of the harmonic features, greater similarity of phenological patterns across two locations or years will lead to higher performance of a random forest model transferred from one region or year to the other. In the US Midwest, two key factors determining crop phenology are the date of sowing and the accumulation of thermal time, typically measured as growing degree days (GDD), after sowing (Mederski et al., 1973; Odenweller and Johnson, 1984). Several cases where the performance dipped, such as Illinois and Iowa in 2012, were associated with unusual timing of sowing, with 2012 having sowing dates several weeks before normal (USDA National Agricultural Statistics Crop Progress and Condition, n.d.). There were also fewer Landsat images and therefore worse quality features for 2012; Landsat 5 Thematic Mapper ended operations in November 2011 and Landsat 8 was not launched until 2013, resulting in an average of 12 readings in a time series rather than the usual 25.

To systematically study the role of GDD differences, we binned samples across all states by GDD and transferred random forests across bins (Section 3.7). Results in Fig. 5 indicate that the accuracy upon transferring a model to a new region correlates significantly with the difference in GDD between the training region and the test region. As expected, models tend to perform best when applied to other samples from the same bin. Accuracy decreases steeply as a model is applied to points more than 300 GDDs different from those of the training set. This decrease can be explained by a combination of (*i*) difference in crop class composition among GDD bins and (*ii*) within-class differences in features for various crops as GDD changes. Decision tree-based algorithms are sensitive to class imbalance and differences between training set and test set class compositions; a random forest trained on data whose crop composition is 50/40/10 corn/soybean/other (such as Illinois) is likely to perform worse on a test sample of 10/30/60 corn/soybean/other (such as North Dakota) than a test sample with the same 50/40/10 composition, all else equal. Reasons include over-estimating the hyperspace corresponding to the corn class simply because there are a large number of corn samples rather than due to the true underlying distribution of corn versus other classes, and that some trees in the forest may never see examples from the "other" class after bootstrapping due to its small sample size. We repeated the experiment with crop class percentages re-sampled to be the same (using CDL labels) across bins and observe greater model transferability, though prediction accuracy still decreases with difference in GDD.

4.3. Unsupervised learning for crop classification

Since supervised models can learn irregular decision boundaries, high classification accuracy using a random forest does not guarantee that a similarly high performance can be achieved using unsupervised methods. However, we show here that harmonic coefficients can in fact separate crop types by boundaries smooth enough to be captured by unsupervised methods, especially a Gaussian mixture model. A visualization of the clustering found by GMM in Illinois is shown in Fig. S2.

Before discussing the end-to-end performance of k-means and GMM, we point out that the classification accuracy of these algorithms depends on two separate steps: the first is how well the clustering algorithm is able to find clusters that correspond to crop types, which we measure using the classification accuracy of the best possible crop-to-cluster assignment (the *optimal classification accuracy*). The second step is whether our assignment algorithm recovers this optimal crop-cluster

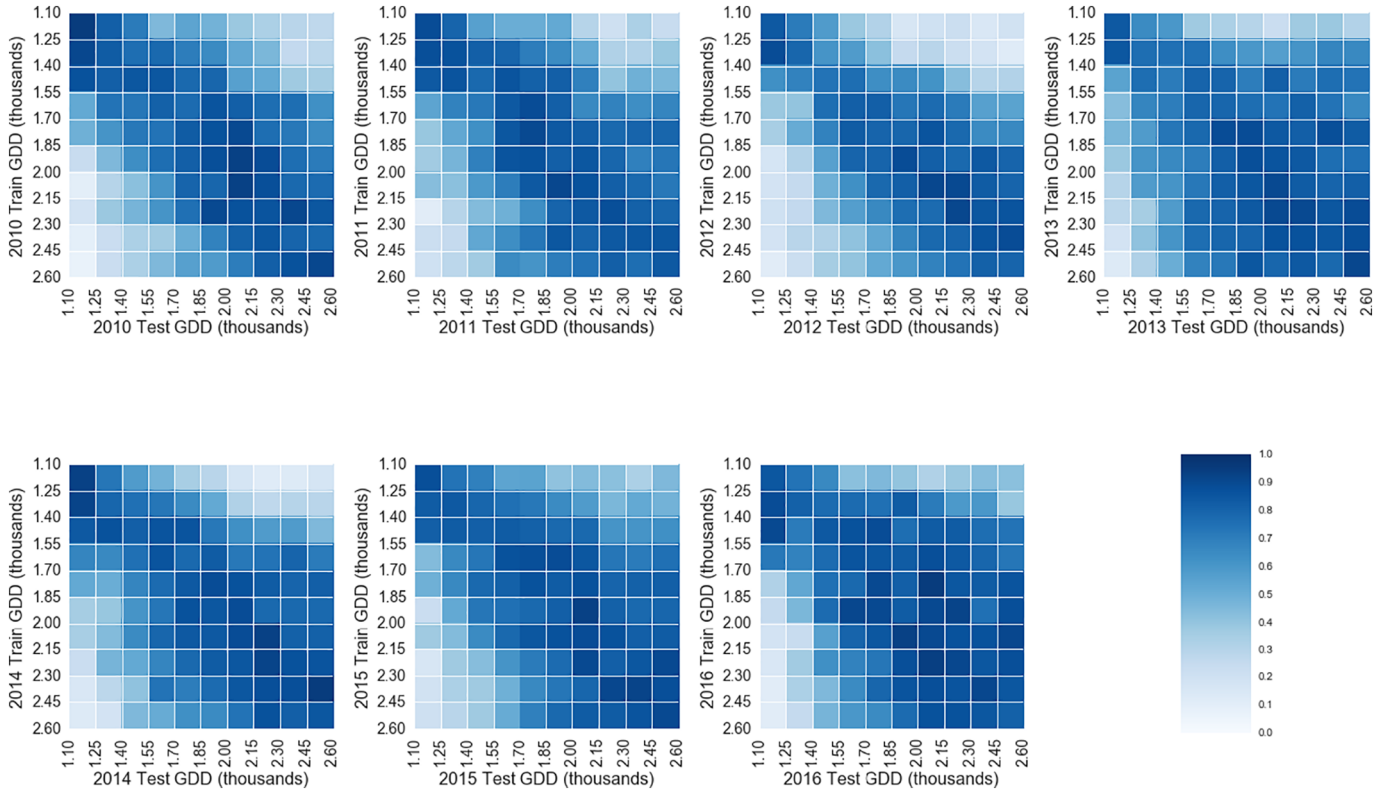

Fig. 5. For each year from 2010 to 2016, accuracy of transferring random forests trained on one region to other regions, based on similarity in growing degree days (GDD). The *y*-axes denote training set GDD bins and the *x*-axes test set GDD bins. GDD is calculated over an entire year and is shown in Celsius.

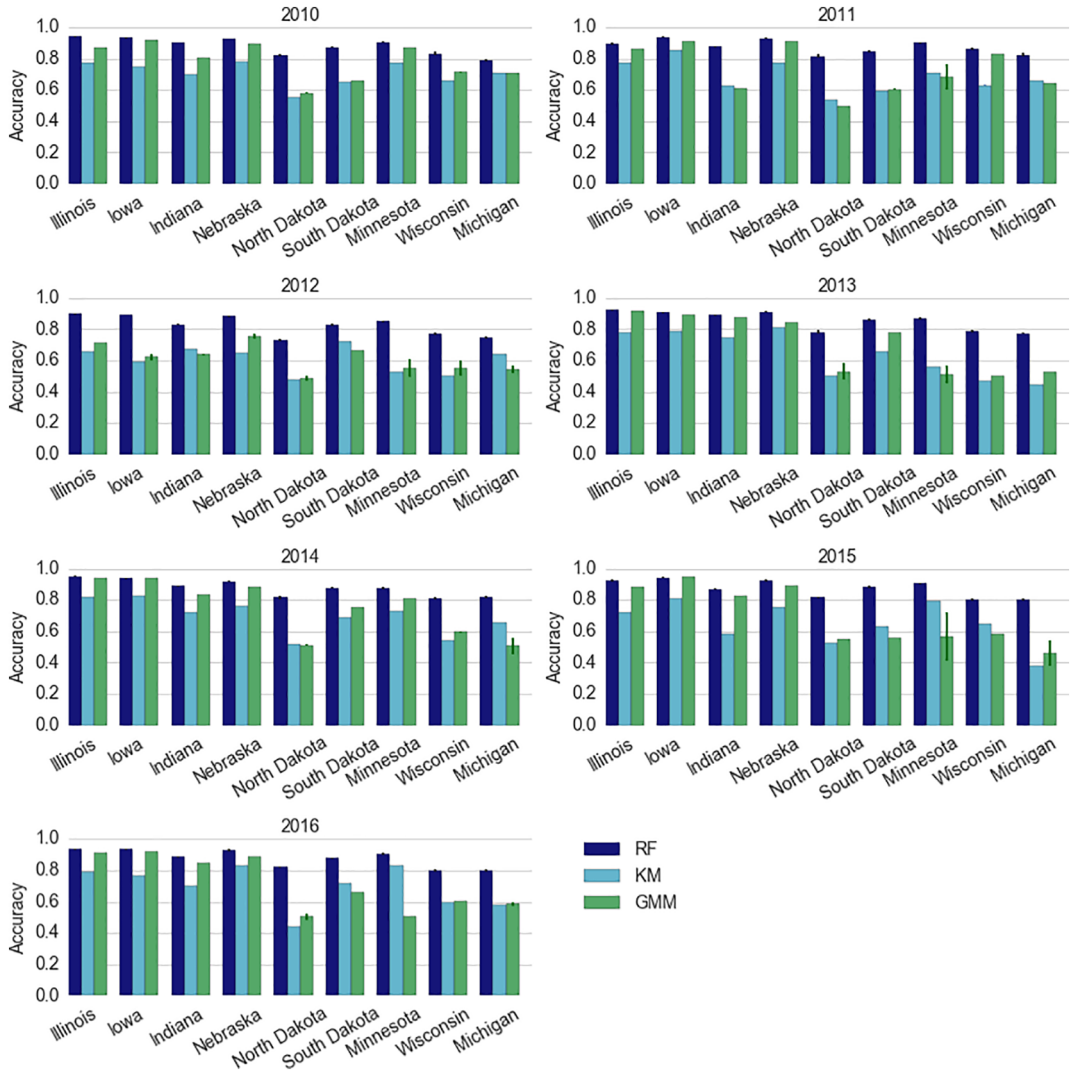

Fig. 6. Random forest, k-means, and Gaussian mixture model (GMM) crop classification accuracies across 9 US states for the years 2010–2016. Each random forest model is trained and tested on data from the same state and year. K-means and GMM clustering are likewise performed independently for data in each state and year, accuracies are shown for the best crop-to-cluster assignment.

assignment. This section describes the best-case results of the clustering algorithms, and Section 4.4 describes the results of our assignment algorithm.

Optimal classification accuracies for k-means and GMM clustering methods applied to all samples in a given state are summarized in Fig. 6 for 2010–2016. In Illinois, Iowa, Indiana, and Nebraska, GMM clustering consistently achieve accuracies around 85–90%, doing nearly as well as the corresponding out-of-sample random forest classifications. In contrast, GMM clustering performs inconsistently in South Dakota, Minnesota, and Wisconsin across years, and only ever reach accuracies below 60% in North Dakota and Michigan. K-means performs significantly worse in most states and years, which is unsurprising since k-means assigns samples to clusters by minimizing Euclidean distance, thereby assuming clusters to be spherical. It does not perform well when clusters are not spherical and vary in size and density (Tan et al., 2005), which is true in this case due to unbalanced crop classes and highly correlated bands and VIs. As mentioned previously, the noticeably lower accuracies across methods in 2012 can be attributed to fewer time series readings in that year due to only Landsat 7 being operational. Fewer Landsat images increase the variance of harmonic regressions within each crop type.

Analysis of clustering performance across twenty different GDD zones in all nine states reveals that unsupervised methods can separate the major crops (corn and soybean) in areas of low crop diversity, but have trouble finding clusters corresponding to crop types as crop diversity increases (Fig. 7). The crop diversity in each bin was measured by Shannon entropy, described in Section 3.7. The smallest possible value of Shannon entropy is zero, which can be achieved when 100% of a region's crop area grows a single crop. Fig. 7 (a) shows that as the crop entropy increases, the clustering accuracy (shown for GMM) decreases, with an R^2 of 0.602. Two example GDD bins from Minnesota are given in Fig. 7 (b); the first bin has a high entropy of 1.77 and low clustering accuracy of 0.56, while the second bin has a low entropy of 0.85 and a high accuracy of 0.91. North Dakota and Michigan have higher crop diversity than Illinois, Iowa, Indiana, and Nebraska, which contributes to their low clustering accuracies. Regions with crop distribution entropies below roughly 1.2 tend to see good performance with unsupervised learning techniques.

We also tried clustering within global agri-ecological zones (GAEZ) and within counties instead of on a state level, with the hopes that doing so might lower within-crop heterogeneity and therefore increase clustering accuracy. We found that performance is stable across different demarcations and scales; clustering accuracy remains high in GAEZs and counties corresponding to Illinois, Iowa, Indiana, and Nebraska, and low in GAEZs and many counties corresponding to North Dakota and Michigan (Fig. S5). Additionally, it is worth noting that

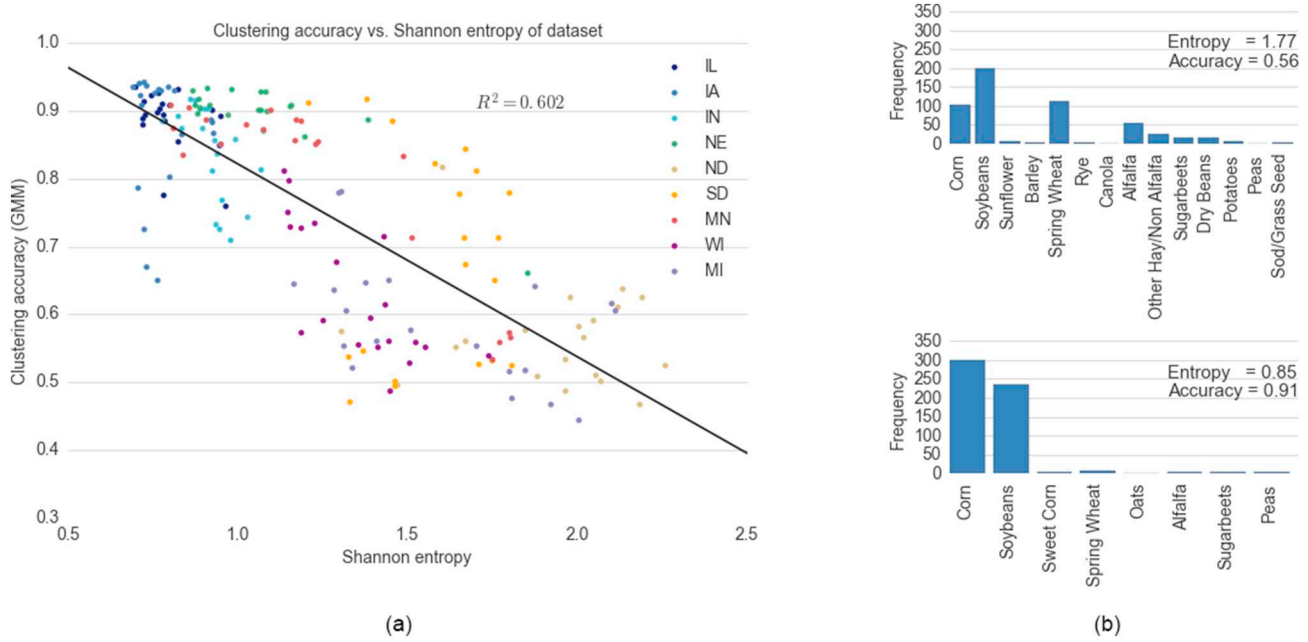

Fig. 7. Clustering accuracy decreases as the entropy of a region's crop distribution increases or, in other words, crop diversity increases. (a) Clustering accuracy for the Gaussian mixture model method is plotted against Shannon entropy for samples divided into 20 GDD zones in each of the 9 states. It appears that the higher the Shannon entropy of crop types in a GDD zone, the lower the clustering accuracy. Linear regression fit to the 180 points has an R^2 of 0.602. (b) Example crop distributions in two GDD zones in Minnesota. Top: GDD 10- to 15-percentile has a high crop type entropy of 1.77 and low clustering accuracy of 0.56. Bottom: GDD 85- to 90-percentile has a low crop type entropy of 0.85 and a high clustering accuracy of 0.91.

unsupervised methods are sensitive to outliers, which can greatly affect the sum of squared distances and likelihoods that guide k-means and mixture model clustering, respectively. Removal of outliers (Section 3.1) is therefore a crucial step to ensure the success of these unsupervised algorithms.

4.4. Automated assignment of crops to clusters using county-level statistics

Since k-means and GMM return unlabeled clusters of samples, the second part of our unsupervised learning algorithm is to find the crop-to-cluster assignment that yields the highest classification accuracy. External information is needed to accomplish this; previous works matched the spectra of clusters to ideal crop spectra (Gumma et al., 2016; Xiong et al., 2017), a procedure that demands high human involvement due to there being hundreds of unlabeled clusters and the generation of ideal crop spectra for crops across many sub-regions to encompass growing season differences. In contrast, statistics on crop area planted and harvested are already available worldwide (You et al., 2014; Monfreda et al., 2008). For the Midwest, we automate the assignment procedure using county-level crop area statistics provided by NASS (Section 3.6).

To see how well the algorithm using aggregate stats can assign crops to clusters, we compare the assignment it returns against the best possible assignment, defined in Section 3.5. We find that when the accuracy of the best possible assignment is high — which is the case in Illinois, Iowa, Indiana, and Nebraska — our algorithm outputs crop-cluster assignments identical to the best possible assignments (Fig. 8). In these states, the unsupervised algorithm is finding clusters that correctly correspond to crop types, which remains true at the county level. In the northern states where clustering is rarely or inconsistently successful in identifying crops, the assignment algorithm does not recover the best possible assignment (and even if it does, classification accuracy may be low). A logistic regression fit to crop-cluster assignment correctness versus whether the best possible assignment is recovered shows that we recover the best possible assignment 90% of the time when clustering accuracies are above 65.5% (Fig. S7).

Of the various losses we tried minimizing in the assignment algorithm, using the log-ratios of major crops (Eq. (3)) performed the best, and has the benefit of requiring only ratios of crop areas to be accurate in order to assign clusters successfully. Assignment based on percentages of crop area in each county also performed well, but may be more sensitive to inaccuracies in absolute crop statistics, which is likely to occur in countries where agricultural data management is less sophisticated than NASS.

To simulate the case where aggregate statistics are not available for the same year as the remote sensing data, we also assigned crops to clusters using previous years' statistics, and achieved the same accuracies for states where GMM found crop clusters, i.e. Illinois, Iowa, Indiana, and Nebraska (Fig. S6). Accuracies in the other five states remained low, and in some cases a worse crop-cluster assignment was recovered than the one recovered with 2016 NASS data. Aggregate statistics from past years can be used to recover best possible assignments because they do not change dramatically from year to year in the Midwest. As long as farmer crop choices are stable across years in a given region, aggregate statistics from past years can be used in the assignment stage of the unsupervised algorithm.

5. Discussion

The results of this study suggest that, in places where field-level ground data is difficult to acquire and not publicly available in large quantities, crop type maps may be generated by either (1) applying a supervised model trained elsewhere or (2) using regional statistics and an unsupervised learning algorithm.

Of the two options, if field-level data is available in a region or year similar to the area of interest, then transferring a supervised method is likely to yield more consistently accurate classification. The more similar crops in two regions and years are, the higher the accuracy will be when a supervised model like random forest is trained on one and applied to the other. In the US Midwest, we used GDD as a proxy for crop type distributions and crop phenologies. We found that the more similar two regions' GDD were, the better a random forest performed

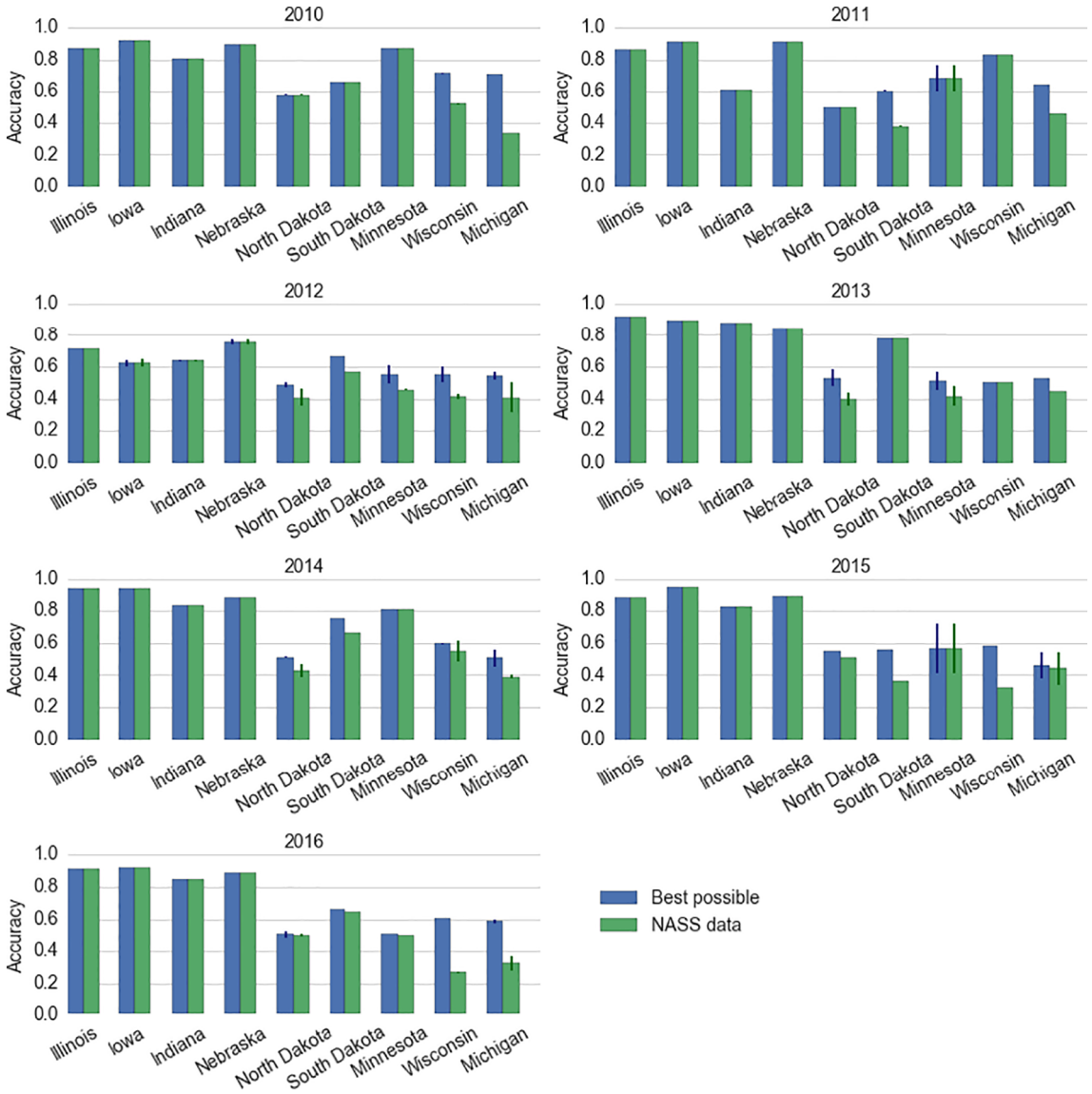

Fig. 8. For 9 states in years 2010–2016, accuracies of GMM clustering algorithm when evaluated under NASS crop statistics-based crop-cluster assignment versus evaluated under best possible crop-cluster assignment.

when transferred between them. Given a data-poor test region and target classification accuracy, one can use similarity in GDD to determine which data-rich regions to train a random forest on to achieve the desired accuracy in the test region. GDD is an appropriate metric in this study because temperature is a limiting factor in plant growth in the US Midwest. It would, however, be less appropriate in places where temperature does not play this limiting role. For a metric to serve as a good proxy for transferability of supervised models, it must derive from data that is easy to acquire and be highly correlated with model performance. This work does not investigate other metrics, but precipitation and geographic distance could also be good candidates, in isolation or in combination with GDD.

In many areas of the world, there is not enough data to train a supervised model on similar regions or other years. These areas with less advanced data infrastructure are often also areas where food insecurity makes mapping crop types via remote sensing especially helpful for understanding farmer decisions and developing strategies to close the yield gap. In these settings, unsupervised learning combined with aggregate statistics remains an option to try for creating crop type maps. We have shown that using Gaussian mixture models on harmonic coefficients can distinguish major crop types at a field-level when crop phenologies differ enough for crop types to be the dominant division across clusters, and when crop diversity (quantifiable via Shannon entropy) is low. We did not map crops that are grown over a small portion of statewide area, as they are difficult to recover using clustering methods and are merged into a single “other” class. GMMs are therefore suitable for applications like food security that are mainly concerned with the staple crops in a region, and not suitable for mapping less dominant crops.

Despite these findings, clear challenges remain. The limitations of unsupervised methods are already apparent in the US Midwest; k-means and GMM have trouble consistently recovering crop clusters in five out of the nine states. In many ways, the Midwestern states are a perfect testing ground: Landsat satellite data is available in large quantities, aggregate statistics at the county level are accurate and released annually, agricultural practices are fairly uniform and farmers have optimized sowing and harvest dates, and only two crops (corn and soybean) dominate the majority of the region. Mapping crop types in the developing world will be challenged by all of these factors being less ideal. The 2015 launch of Sentinel-2 and increasingly available subnational crop statistics worldwide may ameliorate some of the remote sensing data and crop-cluster assignment concerns, but additional

research in the remote sensing and machine learning communities are needed to find better features and more robust clustering algorithms to separate crop types in agriculturally heterogeneous environments.

6. Conclusion

In this study, we assessed the potential of (1) transferring supervised models trained on data from nearby regions and years and (2) using unsupervised learning in conjunction with aggregate crop statistics for the task of crop type mapping in settings where ground data is limited or unavailable. Starting with the US Midwest where data is available for validation, we fit harmonic regressions to Landsat time series and extracted the regression coefficients as features for our machine learning models. County- and state-level crop area data from the USDA NASS were used to determine the number of crop types to classify or cluster, as well as to automatically assign crop labels to clusters output by unsupervised algorithms.

We found that harmonic coefficients can separate crop types successfully when used in supervised or unsupervised methods. Random forests transfer with high accuracy to neighboring geographies when regional crop compositions and growing conditions (as measured by GDD) are similar. The Gaussian mixture model clusters samples into crop types with high accuracy when crop diversity in a region is low. Finally, automatic assignment of crop type to cluster using NASS statistics recovers the correct cluster assignments when clustering accuracies exceeded 65%.

The methods developed in this work enable the hindcasting of CDL in the US to fill in states and years without CDL maps, and potentially serve as a first step toward the creation of crop type maps in regions around the world that lack field-level labels but possess regular and accurate district-level aggregate statistics. Notably, these aggregate statistics need only be accurate in a relative sense, since we use the log-ratio of crop areas for cluster assignment and not the absolute area of each crop. As aggregate statistics are increasingly available for subnational administrative units, this approach should be widely useful. Classification in developing countries is likely to be challenged by factors including more complex mixtures of crops, cloudier growing seasons, wider ranges of sowing dates, and less accurate aggregate statistics. Where fewer Landsat observations are available per year, future studies could explore the use of Sentinel-2 imagery to complement Landsat data. Future work could also extend these approaches to the problem of in-season forecasting of crop area in data-scarce regions.

Declarations of interest

None.

References

Azzari, G., Lobell, D., 2017. Landsat-based classification in the cloud: an opportunity for a paradigm shift in land cover monitoring. Remote Sens. Environ. 202, 64–74 Big remotely sensed data: tools, applications and experiences.

Bhargava, A.K., Vagen, T., Gassner, A., 2018. Breaking ground: unearthing the potential of high-resolution, remote-sensing soil data in understanding agricultural profits and technology use in sub-Saharan Africa. World Dev. 105, 352–366.

Boryan, C., Yang, Z., Mueller, R., Craig, M., 2011. Monitoring US agriculture: the US Department of Agriculture, national agricultural statistics service, cropland data layer program. Geocarto Int. 26 (5), 341–358.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Broge, N.H., 2001. Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density. Remote Sens. Environ. 76 (2), 156–172.

Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., Li, Z., 2018. A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. Remote Sens. Environ. 210, 35–47.

Carletto, C., Gourlay, S., Winters, P., 2015. From guesstimates to GPStimates: land area measurement and implications for agricultural analysis. J. Afr. Econ. 24 (5), 593–628.

Clauss, K., Yan, H., Kuenzer, C., 2016. Mapping paddy rice in China in 2002, 2005, 2010 and 2014 with modis time series. Remote Sens. 8 (5).

Cohn, A.S., VanWey, L.K., Spera, S.A., Mustard, J.F., 2016. Cropping frequency and area response to climate variability can exceed yield response. Nat. Clim. Chang. 6, 601.

Common Land Unit (CLU) Information Sheet. Available at https://www.fsa.usda.gov/assets/usda-fsa-public/usdafiles/apfo/support-documents/pdfs/clu_infosheet_2017_final.pdf (accessed 2018-11-12; verified 2018-11-12).

Deventer, v., Ward, A., Gowda, P., Lyon, J., 1997. Using thematic mapper data to identify contrasting soil plains and tillage practices. Photogramm. Eng. Remote. Sens. 63, 87–93.

Dong, J., Xiao, X., Menarguez, M.A., Zhang, G., Qin, Y., Thau, D., Biradar, C., Moore, B., 2016. Mapping paddy rice planting area in northeastern asia with Landsat 8 images, phenology-based algorithm and google earth engine. Remote Sens. Environ. 185, 142–154 Landsat 8 Science Results.

Fisette, T., Rollin, P., Aly, Z., Campbell, L., Daneshfar, B., Filyer, P., Smith, A., Davidson, A., Shang, J., Jarvis, I., 2013. AAFC annual crop inventory. In: 2013 Second International Conference on Agro-Geoinformatics (Agro-Geoinformatics), pp. 270–274.

Foerster, S., Kaden, K., Foerster, M., Itzerott, S., 2012. Crop type mapping using spectral-temporal profiles and phenological information. Comput. Electron. Agric. 89, 30–40.

Fu, G., Liu, C., Zhou, R., Sun, T., Zhang, Q., 2017. Classification for high resolution remote sensing imagery using a fully convolutional network. Remote Sens. 9 (5).

Ghazaryan, G., Dubovyk, O., Löw, F., Lavreniuk, M., Kolotii, A., Schellberg, J., Kussul, N., 2018. A rule-based approach for crop identification using multi-temporal and multi-sensor phenological metrics. Eur. J. Remote Sens. 51 (1), 511–524.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recogn. Lett. 27 (4), 294–300 Pattern Recognition in Remote Sensing (PRRS 2004).

Gitelson, A.A., Vina, A., Ciganda, V., Rundquist, D.C., Arkebauer, T.J., 2005. Remote estimation of canopy chlorophyll content in crops. Geophys. Res. Lett. 32 (8). https://doi.org/10.1029/2005GL022688.

Gomez, C., White, J.C., Wulder, M.A., 2016. Optical remotely sensed time series data for land cover classification: a review. ISPRS J. Photogramm. Remote Sens. 116, 55–72.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: planetary-scale geospatial analysis for everyone. Remote Sens. Environ 202, 18–27.

Gourlay, S., Kilic, T., Lobell, D., 2017. Could the Debate be Over? Errors in Farmer-Reported Production and Their Implications for the Inverse Scale-Productivity Relationship in Uganda. The World Bank.

Gumma, M.K., Thenkabail, P.S., Teluguntla, P., Rao, M.N., Mohammed, I.A., Whitbread, A.M., 2016. Mapping rice-fallow cropland areas for short-season grain legumes intensification in South Asia using MODIS 250 m time-series data. Int. J. Digital Earth 9 (10), 981–1003.

Hao, P., Wang, L., Zhan, Y., Niu, Z., 2016a. Using moderate-resolution temporal NDVI profiles for high-resolution crop mapping in years of absent ground reference data: a case study of Bole and Manas counties in Xinjiang, China. ISPRS Int. J. Geo-Inf. 5 (5).

Hao, P., Wang, L., Zhan, Y., Wang, C., Niu, Z., Wu, M., 2016b. Crop classification using crop knowledge of the previous-year: case study in southwest Kansas, USA. Eur. J. Remote Sens. 49 (1), 1061–1077.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Heller, E., Rhemtulla, J.M., Lele, S., Kalacska, M., Badiger, S., Sengupta, R., Ramankutty, N., 2012. Mapping crop types, irrigated areas, and cropping intensities in heterogeneous landscapes of southern india using multi-temporal medium-resolution imagery. Photogramm. Eng. Remote. Sens. 78 (8), 815–827.

Huete, A., 1988. A soil-adjusted vegetation index (savi). Remote Sens. Environ. 25 (3), 295–309.

Inglada, J., Arias, M., Tardy, B., Hagolle, O., Valero, S., Morin, D., Dedieu, G., Sepulcre, G., Bontemps, S., Defourny, P., Koetz, B., 2015. Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. Remote Sens. 7 (9), 12356–12379.

Jakubauskas, M.E., Legates, D.R., Kastens, J.H., 2002. Crop identification using harmonic analysis of time-series AVHRR NDVI data. Comput. Electron. Agric. 37 (1), 127–139.

June Area. Available at https://www.nass.usda.gov/surveys/guide_to_nass_surveys/june_area/index.php (accessed 2018-11-12; verified 2018-11-12).

Ju, J., Kolaczyk, E.D., Gopal, S., 2003. Gaussian mixture discriminant analysis and sub-pixel land cover characterization in remote sensing. Remote Sens. Environ. 84 (4), 550–560.

Key, C., Benson, N., 2006. Landscape Assessment: Ground Measure of Severity, the Composite Burn Index; and Remote Sensing of Severity, the Normalized Burn Ratio. 01 LA 1-51. .

LDCM CAL/VAL Algorithm Description Document. Available at https://landsat.usgs.gov/sites/default/files/documents/ldcm_cvt_add.pdf (accessed 2018-08-30; verified 2018-08-30). https://landsat.usgs.gov/sites/default/files/documents/LDCM_CVT_ADD.pdf.

Liu, H.Q., Huete, A., 1995. A feedback based modification of the NDVI to minimize canopy background and atmospheric noise. IEEE Trans. Geosci. Remote Sens. 33 (2), 457–465.

Liu, J., Feng, Q., Gong, J., Zhou, J., Liang, J., Li, Y., 2018. Winter wheat mapping using a random forest classifier combined with multi-temporal and multi-sensor data. Int. J. Digital Earth 11 (8), 783–802. https://doi.org/10.1080/17538947.2017.1356388.

Liverpool-Tasie, L.S.O., Omonona, B.T., Sanou, A., Ogunleye, W.O., 2017. Is increasing inorganic fertilizer use for maize production in SSA a profitable proposition? Evidence from Nigeria. Food Policy 67, 41–51 Agriculture in Africa - telling myths from facts.

Loveland, T.R., Reed, B.C., Brown, J.F., Ohlen, D.O., Zhu, Z., Yang, L., Merchant, J.W., 2000. Development of a global land cover characteristics database and IGBP

DISCover from 1 km AVHRR data. Int. J. Remote Sens. 21 (6–7), 1303–1330.
McCarty, J., Neigh, C., Carroll, M., Wooten, M., 2017. Extracting smallholder cropped area in Tigray, Ethiopia with wall-to-wall sub-meter WorldView and moderate resolution Landsat 8 imagery. Remote Sens. Environ. 202, 142–151 Big remotely sensed data: tools, applications and experiences.
Mederski, H.J., Miller, M.E., Weaver, C., 1973. Accumulated heat units for classifying corn hybrid maturity. Agron. J. 65 (5), 743–747.
Monfreda, C., Ramankutty, N., Foley, J.A., 2008. Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. Glob. Biogeochem. Cycles 22 (1).
Nuarsa, I.W., Nishio, F., Hongo, C., Mahardika, I.G., 2012. Using variance analysis of multitemporal modis images for rice field mapping in Bali Province, Indonesia. Int. J. Remote Sens. 33 (17), 5402–5417.
Odenweller, J.B., Johnson, K.I., 1984. Crop identification using Landsat temporal-spectral profiles. Remote Sens. Environ. 14 (1), 39–54.
Ok, A.O., Akar, O., Gungor, O., 2012. Evaluation of random forest method for agricultural crop classification. Eur. J. Remote Sens. 45 (1), 421–432.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Portmann, F.T., Siebert, S., Döll, P., 2010. Mirca2000 global monthly irrigated and rainfed crop areas around the year 2000: a new high-resolution data set for agricultural and hydrological modeling. Glob. Biogeochem. Cycles 24 (1). https://doi.org/10.1029/2008GB003435.
Rondeaux, G., Steven, M., Baret, F., 1996. Optimization of soil-adjusted vegetation indices. Remote Sens. Environ. 55 (2), 95–107.
Roy, D., Wulder, M., Loveland, T.R., Woodcock, C.E., Allen, R.G., Anderson, M.C., Helder, D., Irons, J.R., Johnson, D.M., Kennedy, R., Scambos, T.A., Schaaf, C.B., Schott, J.R., Sheng, Y., Vermote, E.F., Belward, A.S., Bindschadler, R., Cohen, W.B., Gao, F., Hipple, J.D., Hostert, P., Huntington, J., Justice, C.O., Kilic, A., Kovalskyy, V., Lee, Z.P., Lymburner, L., Masek, J.G., McCorkel, J., Shuai, Y., Trezza, R., Vogelmann, J., Wynne, R.H., Zhu, Z., 2014. Landsat-8: science and product vision for terrestrial global change research. Remote Sens. Environ. 145, 154–172.
Shao, Y., Fan, X., Liu, H., Xiao, J., Ross, S., Brisco, B., Brown, R., Staples, G., 2001. Rice monitoring and production estimation using multitemporal radarsat. Remote Sens. Environ. 76 (3), 310–325.
Sharma, V., Irmak, S., Kilic, A., Sharma, V., Gilley, J.E., Meyer, G.E., Knezevic, S.Z., Marx, D., 2016. Quantification and mapping of surface residue cover for maize and soybean fields in south central Nebraska. Trans. ASABE 59 (3), 925.
Tan, P.-N., Steinbach, M., Kumar, V., 2005. Introduction to Data Mining (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
Tao, J., Shu, N., Wang, Y., Hu, Q., Zhang, Y., 2016. A study of a Gaussian mixture model for urban land-cover mapping based on VHR remote sensing imagery. Int. J. Remote Sens. 37 (1), 1–13.
Thornton, P., Thornton, M., Mayer, B., Wei, Y., Devarakonda, R., Vose, R., Cook, R., 2018. Daymet: daily surface weather data on a 1-km grid for North America, Version 3.
Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ. 8 (2), 127–150.
Turker, M., Ozdarici, A., 2011. Field-based crop classification using SPOT4, SPOT5, IKONOS and QuickBird imagery for agricultural areas: a comparison study. Int. J. Remote Sens. 32 (24), 9735–9768.
USDA National Agricultural Statistics Crop Progress and Condition. Available at https://www.nass.usda.gov/Charts_and_Maps/Crop_Progress_&_Condition/ (accessed 2018-03-14; verified 2018-03-14).
USDA National Agricultural Statistics Service Acreage. Published crop-specific acreage [Online]. Available at http://usda.mannlib.cornell.edu/usda/current/Acre/Acre-06-30-2017.pdf (accessed 2018-04-25; verified 2018-04-25).
USDA National Agricultural Statistics Service Cropland Data Layer. Published crop-specific data layer [Online]. Available at https://nassgeodata.gmu.edu/CropScape/ (accessed 2018-04-25; verified 2018-04-25).
USDA National Agricultural Statistics Service Quick Stats. 2018. Available at https://quickstats.nass.usda.gov/ (accessed 2018-04-25; verified 2018-04-25).
van Wart, J., van Bussel, L.G., Wolf, J., Licker, R., Grassini, P., Nelson, A., Boogaard, H., Gerber, J., Mueller, N.D., Claessens, L., van Ittersum, M.K., Cassman, K.G., 2013. Use of agro-climatic zones to upscale simulated crop yield potential. Field Crop Res. 143, 44–55 Crop yield gap analysis - rationale, methods and applications.
Waldner, F., Fritz, S., Gregorio, A.D., Defourny, P., 2015. Mapping priorities to focus cropland mapping activities: fitness assessment of existing global, regional and national cropland maps. Remote Sens. 7, 7959–7986.
Wu, B., Gommes, R., Zhang, M., Zeng, H., Yan, N., Zou, W., Zheng, Y., Zhang, N., Chang, S., Xing, Q., van Heijden, A., 2015. Global crop monitoring: a satellite-based hierarchical approach. Remote Sens. 7 (4), 3907–3933.
Xiong, J., Thenkabail, P.S., Gumma, M.K., Teluguntla, P., Poehnelt, J., Congalton, R.G., Yadav, K., Thau, D., 2017. Automated cropland mapping of continental Africa using Google Earth Engine cloud computing. ISPRS J. Photogramm. Remote Sens. 126, 225–244.
Yang, C., Everitt, J.H., Fletcher, R.S., Murden, D., 2007. Using high resolution QuickBird imagery for crop identification and area estimation. Geocarto Int. 22 (3), 219–233.
You, L., Wood, S., Wood-Sichra, U., Wu, W., 2014. Generating global crop distribution maps: from census to grid. Agric. Syst. 127, 53–60.
Zhong, L., Hu, L., Yu, L., Gong, P., Biging, G.S., 2016a. Automated mapping of soybean and corn using phenology. ISPRS J. Photogramm. Remote Sens. 119, 151–164.
Zhong, L., Yu, L., Li, X., Hu, L., Gong, P., 2016, 11. Rapid corn and soybean mapping in US corn belt and neighboring areas. Sci. Rep. 6, 36240.