Computers and Electronics in Agriculture 194 (2022) 106704

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Assessing machine leaning algorithms on crop yield forecasts using functional covariates derived from remotely sensed data

Luca Sartore [a,b,*], Arthur N. Rosales [b], David M. Johnson [b], Clifford H. Spiegelman [c,b,1]

[a] 1750 K Street NW Suite 1100, Washington, DC 20006, United States
[b] 1400 Independence Avenue SW, Washington, DC 20250, United States
[c] 155 Ireland St, College Station, TX 77843, United States

ARTICLE INFO

Keywords:
Remote sensing
Functional covariates
Nonparametric regression
Feature extraction
Multidimensional scaling

2010 MSC:
00–01
62P12
92D99

ABSTRACT

Machine learning methods are increasingly used in analyzing remotely sensed data and studying different aspects of agricultural production. In particular, several of these flexible models are widely adopted to predict regional crop yield during or after the growing season. However, most existing models cannot be applied when dealing with functional covariates. In this paper, an approach based on multidimensional scaling is proposed to generate a set of artificial covariates from empirical density functions of different phenomena captured within specific administrative boundaries through satellites. In contrast to traditional aggregation methods, this approach is designed to reduce the loss of information associated with the use of summary statistics as covariates. The proposed methodology is applied to NASA remote sensing data, combined with information from surveys and USDA's end-of-season county estimates, to study the prediction accuracy of different crop-yield models for three major crops in North Dakota.

1. Introduction

The United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) is mandated by the United States (US) Congress to publish official statistics for major crops produced in the US at the county, state, and national administrative levels. These statistics play significant roles in several aspects of the US economy, ranging from market regulations to business decisions pertaining to marketing and risk management. Therefore, accuracy and timeliness are both critical objectives in NASS estimates and forecasts.

Crop forecasting has evolved over the decades as theoretical approaches improved, data collection intensified, and technology unlocked new analytical and computational capabilities. Irwin (1938) distinguished between subjective yield forecasts based only on expert opinions and objective methods relying solely on meteorological information. Cochran (1938) further discussed agricultural meteorology when noting the challenges between early-season yield predictions and post-harvest estimates. It was also noted that the prediction task is more difficult due to unstable weather dynamics and other factors that have a substantial effect on the actual yield. Yield forecasts of this time were largely derived using linear models. While still widely used for their simplicity of interpretation, these models can be limited as they exclude any important nonlinear effects. Additionally, objective-yield measurements used to derive forecasts for large geographic areas were sparsely sampled due to technology limitations. With the subsequent introduction of satellite technologies, Walker and Sigman (1982) compared two estimators that combined remote sensing information with survey data to provide estimates for administrative areas having few or no sampling units. However, these two estimators were originally developed under unrealistic assumptions of normality to quantify cultivated areas rather than yields.

Matis et al. (1985) developed a crop simulation model based on Markov chains using soil moisture data to forecast crop yield distributions. Although they compared their approach to a regression method, the precision of the yield forecasts was not evaluated. Other families of simulation models were considered by Doraiswamy et al. (2005), who computed county-level yield predictions based on data acquired by the National Aeronautics and Space Administration (NASA) Moderate Resolution Imaging Spectroradiometer (MODIS; Justice et al., 2002) instruments. However, their model must be calibrated when applied to

* Corresponding author at: 1750 K Street NW Suite 1100, Washington, DC 20006, United States.
E-mail address: lsartore@niss.org (L. Sartore).
URL: https://www.niss.org/people/luca-sartore (L. Sartore).
[1] Deceased on June 14, 2020.

https://doi.org/10.1016/j.compag.2022.106704
Received 1 July 2021; Received in revised form 16 December 2021; Accepted 8 January 2022
Available online 4 February 2022

different areas. To avoid calibration, the distribution of agricultural yields was assessed by Wang et al. (2012) through a Bayesian model that achieves accurate forecasts for the speculative region associated with a crop using only multiple repeated surveys. The extension provided by Nandram et al. (2014) allows one to forecast state-level yields in the United States. Cruze et al. (2019) and Erciulescu et al. (2020) focused on crop production estimates computed at the county level using Bayesian inference methods. However, these Bayesian models are all linear and do not consider relevant crop phenology information. Furthermore, Bayesian inference suffers from several drawbacks despite its popularity. Notably, it is computationally demanding, and for this reason, it restricts the analysis to quite rigid models. The effects of using linear models for yield prediction were discussed by Rasmussen (1997), who highlighted the irregular distribution of residuals and noted that this and other issues could be related to model misspecification due to unknown nonlinear relationships characterizing the data.

Young (2019) provided a review of current methods being used to produce official statistics for large geographical areas by integrating survey data, administrative records and other sources of information. Early attempts to integrate several sources of information during the growing season for forecasting agricultural yields at different administrative levels of aggregation were made by Chipanshi et al. (2015), who discussed the role of their methods to complement early season survey estimates. Some private firms in the agronomic industry are capable of producing yield forecasts at much finer resolution than the official US county estimates. This is accomplished by using big data obtained from remotely sensed images and precision-agriculture technologies. Due to the large amount of useful information currently available, machine learning algorithms have been extensively used in recent years to forecast crop-yields (Elavarasan et al., 2018). For example, Cai et al. (2017) and Zhao et al. (2017) studied a variety of semi-parametric regression models to predict crop-yields at the end of the harvesting season. They linked NASS official statistics (yield estimates and crop conditions) at the county level with a variety of other data sources acquired during the growing season. Predictive approaches in machine learning use flexible regression models that can account for nonlinear effects and produce more accurate predictions for the unknown values of a response variable. Classification and Regression Trees (CART; Breiman et al., 1984), Random Forests (RF; Breiman, 2001), Support Vector Machines (SVM; Cortes and Vapnik, 1995), Artificial Neural Network (ANN; McCulloch and Pitts, 1943) and Cubist (Quinlan, 1992; Quinlan, 1993) are among the most common models that improve prediction accuracy. However, all these models were developed under the assumption that all input variables are organized into numerical structures (mostly vectors or matrices). Several extensions of generalized linear regression models have been proposed to combine numerical and functional predictors (Moyeed and Diggle, 1994; Zeger and Diggle, 1994; Hoover et al., 1998; Wu et al., 1998; Lin and Ying, 2001; James and Hastie, 2001; James, 2002), but these methods only considered functional data that vary over time.

The use of deep neural network (DNN) models as proposed in Jiang et al. (2020) focused on Long-Short Term Memory (LSTM) networks to better capture nonlinear relationships among sequential patterns over time and agricultural yields. Convolutional Neural Networks (CNN; You et al., 2017) have been also applied to explore spatial phenomena, while the combination between CNN and LSTM networks (Sun et al., 2019) has recently provided an enhanced framework to process tensor-valued training data. However, despite the popularity of deep neural network (DNN) models, these suffer from several drawbacks (Ertel, 2018). For instance, the validation and training of a DNN are quite challenging tasks and have profound consequences in developing and finalizing a reliable predictive model. Furthermore, the extracted features are based on a sequence of numerical transformations designed to improve the prediction accuracy without accounting for other statistical properties that could be beneficial in describing useful relationships between the input data and the output.

This paper introduces an innovative approach for extracting numerical features from remotely sensed data to better inform crop-yield predictions and to assess the associated levels of uncertainty. The aim of this approach is to reduce the loss of information due to traditional data-aggregation strategies used to summarize remote sensing data. The proposed methodology provides a set of artificial covariates that is created by extracting most of the information from the empirical density functions of real-life phenomena estimated at the county level. This allows relevant features of empirical densities to be used as input variables in standard machine learning algorithms. County-level predictions of crop yields (and correspondent measures of uncertainty) are successively obtained with the fitted models by accounting also for state-level expert opinion and survey data summarized at the county level.

This article is organized as follows. Section 2 provides an overview of the current remote sensing methodology used at NASS to assist the Agricultural Statistics Board in the production of official estimates. Section 3 describes an innovative approach for extracting numerical features from remotely sensed images, linking those features to historical official statistics and survey data, and predicting yields for well-defined administrative areas. A case study is provided in Section 4 to establish the performance of machine learning algorithms in predicting the yields of three major crops (corn, soybeans, and wheat) in North Dakota (see red area in Fig. 1). Concluding remarks are given in Section 5.

2. Overview of current remote sensing methodology

The current remote sensing methodology in use at NASS is primarily based on the use of vegetation indices. Historically, linear relationships between yield and stress-degree-days were found by Hatfield (1983), who focused on defining the reproductive period of a crop's life cycle using spectral and thermal data. However, these data relationships were not always maintained due to differences in agricultural practices in different regions. This issue was addressed by a quadratic model developed by Hayes and Decker (1996). Their model made use of weekly data to provide yield forecasts at least two months prior to harvest, and they discussed the role of weather assessment in the forecasting process. These two approaches were successively extended by Johnson (2014), who developed a procedure to predict county-level yields that are independent from NASS survey data (see Fig. 2). Historical data from satellite imagery and county-level NASS official statistics are combined to build a data set for training and validating a machine learning model. Afterwards, the new remotely sensed data collected within the growing season are processed to generate a separate dataset that serves to provide out-of-sample yield predictions after the growing season has ended.

Currently, the imagery data that NASS uses for remote sensing yield estimation is acquired by NASA MODIS instruments (Johnson, 2016), which are found on the polar orbiting Terra and Aqua earth observation satellites (Barnes et al., 2003). This two-satellite constellation views the entire Earth's surface every one to two days and collects data in 36 spectral bands. The wavelengths that these bands cover range from short bandwidths for visible light to longer wavelength thermal bands. For assessing vegetation density, remote sensing researchers rely on the most pertinent visible light and the longer near-infrared (near-IR) bands. This is due to the nature of plants themselves, where chlorophyll strongly absorbs visible light for use in photosynthesis, and leaf cell structure strongly reflects near-IR light. Additionally, canopy temperature is an important factor in crop growth monitoring. Therefore, MODIS reflectance and thermal bands have been widely used for tracking and assessing crop growth.

The red and near-IR bands (respectively denoted by ϱ_{Red} and ϱ_{NIR}) are combined into the Normalized Difference Vegetation Index (NDVI; Tucker, 1979), i.e.

$$\text{NDVI} = \frac{\varrho_{\text{NIR}} - \varrho_{\text{Red}}}{\varrho_{\text{NIR}} + \varrho_{\text{Red}}}.$$

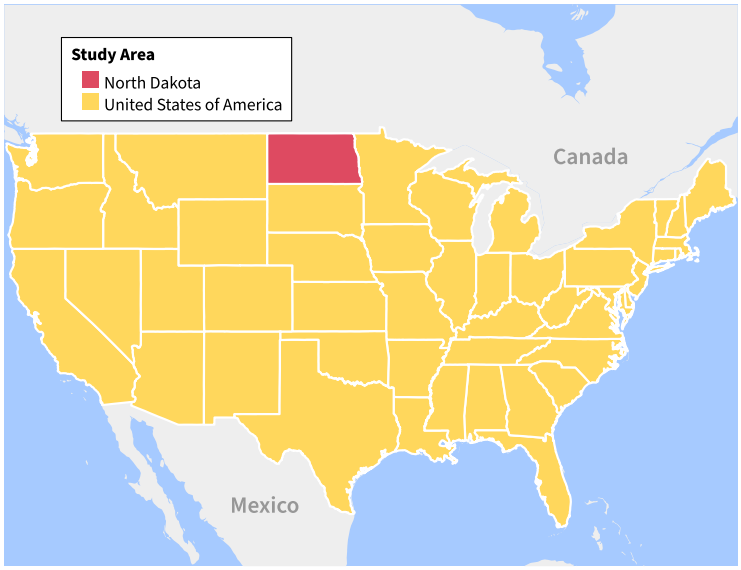

Fig. 1. The state of North Dakota highlighted with respect to the other states in the conterminous US.

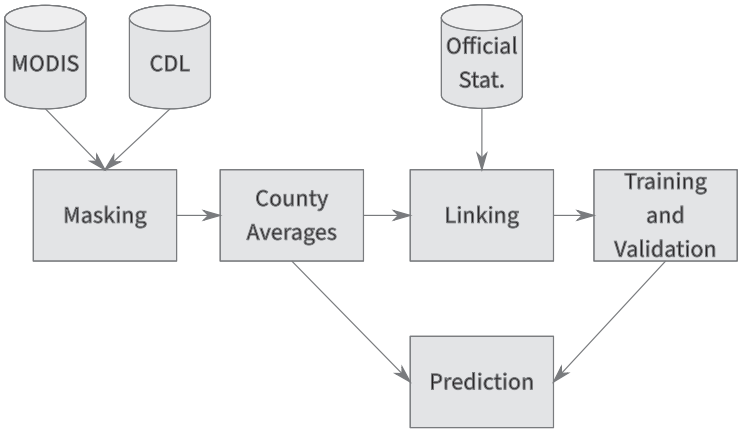

Fig. 2. Graphical representation of the current approach to predict agricultural yield at the county level using summary statistics from remote sensing data.

The evolution of this index value and the variations of the thermal bands during the growing season summarize most of the information needed to properly predict crop yield (Johnson, 2014). The red, near-IR, and thermal band data are collected by MODIS instruments and are globally available as 8-day composite high-level-product raster-image tiles. Those tiles, which cover the area of interest, are spatially mosaiced together as a series of composite images that span a well-defined period between early April and late October. The resulting raster images are then clipped by a specific crop based on the NASS Cropland Data Layer (CDL; Boryan et al., 2011), which is a GIS layer identifying the location of 112 crops during a growing season. The result is useful masks that remove spurious signals from other crops or forested areas. Once the remotely sensed variables have been cleaned by the masking process, they are summarized at the county level by taking averages.

Averaging the pixel values over the fields covered by the crop of interest is computationally advantageous. In fact, a well-organized dataset can be easily constructed by linking official statistics published by NASS at the county level with the average dynamic of important remotely sensed predictors. Furthermore, this approach allows one to train and validate a model to predict county level yield; however, the dynamics learned at a higher aggregation (e.g. at the county) level can later be used to predict yield at the pixel level. On the other hand, averages are not robust and may produce unreliable values due to the possible presence of outliers in the data. Furthermore, averages do not fully summarize any observed stochastic process at the county level causing possible losses of information by ignoring field-level and/or pixel-level variability. In the next section, a new approach to link remote sensing data to both official statistics and survey summaries is proposed to address the loss of information.

3. Use of empirical densities in predicting yields

The temporal evolution of the predictor variables is characterized by dynamic variations of a multivariate distribution describing the behavior of a stochastic process observed over the crop fields of interest. When summarizing remotely sensed data from pixel-level values to higher levels of aggregation, the use of statistics beyond the mean (such as variance, skewness, kurtosis and other higher moments) can fully describe the underlying unknown distribution of the observed stochastic process. However, this approach is not robust to outliers when the underlying model assumptions are either violated, ignored, or misspecified (Pearson, 2005, Section 2.2.1). Furthermore, different spatial resolutions of the images used for producing values at the desired level of aggregation can adversely affect model results (De Wit, 2005; Gao et al., 2018). To address these issues, a dynamic temporal multivariate distribution can be empirically estimated, but this solution is not feasible

since standard kernel density estimators become computationally prohibitive when processing more than two spectral bands (Hastie et al., 2009, Chapter 6). Therefore, the empirical estimation of marginal densities becomes a computationally viable option to summarize and study the evolution of remotely sensed variables at different spatial resolutions. The use of histograms as covariates instead of summary statistics was initially proposed by You et al. (2017) to better inform their prediction model. In particular, relevant features are extracted through a DNN consisting of several convolutional layers that provide a nonlinear mapping from histograms to yield.

In contrast to a convolutional DNN approach, a multidimensional scaling (MDS) methodology is proposed to generate a set of artificial covariates that are later used in several predictive algorithms (similarly to Cuadras and Arenas, 1990; Boj et al., 2016). Furthermore, MDS is the foundation of distance-based regression methods that are also well-suited for minimizing the loss of information caused by the transformation of county-level density functions into data vectors (see Fig. 3).

MDS allows one to summarize relevant statistical properties of the marginal distributions of interest, while the effect of outliers is mitigated through smoothed histograms (Zambom and Ronaldo, 2013). These smoothed histograms are produced using kernel density estimates evaluated at the center of each histogram bin. In practice, an empirical density function can be formulated as

$$\widehat{f}_i(x) \propto \sum_{j=1}^{m} \widehat{h}_i(x_j) \exp\left\{ -\frac{(x-x_j)^2}{2\phi} \right\}, \tag{1}$$

where the index i denotes the variable of interest (e.g. a spectral band), x takes values in the support of the i-th variable, m represents the total number of histogram bins, and the j-th Gaussian kernel function is centered in x_j (i.e. the middle point of the j-th histogram bin), scaled through the bandwidth ϕ usually selected through Silverman's (1986) rule-of-thumb (or other methods as described in Sheather and Jones, 1991; Wand and Jones, 1995), and weighted by the estimated height of its histogram bar, $\widehat{h}_i(x_j)$, for any $j = 1, \ldots, m$.

To obtain numerical data that approximate the evolution of the original functional predictors, MDS is applied week-by-week to extract an optimal number of dimensions from a distance matrix obtained by pairwise comparisons between county-level densities of a remotely sensed variable acquired during the growing season. The Jensen-Shannon (JS) distance,

$$\mathrm{JS}(f,g) = \sqrt{\int_{\mathbb{R}} \frac{f(x)}{2} \log \frac{2f(x)}{f(x)+g(x)} \mathrm{d}x + \int_{\mathbb{R}} \frac{g(x)}{2} \log \frac{2g(x)}{f(x)+g(x)} \mathrm{d}x},$$

provides a proper metric[2] to measure the discrepancy between two generic density functions, saying f and g. This distance always returns a value in the interval $[0,1]$, and it is more stable than other symmetrized divergences since it satisfies the triangular inequality (Nielsen, 2010). Moreover, the JS distance allows the MDS algorithm to map the smoothed histograms into a multidimensional Euclidean space. Thus, the data vectors generated by the MDS algorithm can be used in any standard prediction model, since they numerically summarize the most relevant features of county-level densities (such as within-field variations, asymmetry, and other information beyond the first three moments).

New datasets can be built by combining MDS results (as covariates) and NASS official statistics (as responses) at the county level. However, NASS estimates are publicly available for the previous years and only for the counties that satisfy NASS publication standards. Furthermore, predicting a within-season crop-type mask is helpful for cleaning spectral signals at the field level to generate a reliable set of artificial covariates for a specific crop planted during the current year. When survey data are available, they can improve the model by providing a direct and independent source of information on the average yield at the state or county level. Therefore, the survey summaries (usually weighted average, coefficients of variation, and sample size) can enter the model along with the other covariates allowing the chosen model to automatically adjust the prediction for the typical survey errors (Lessler and Kalsbeek, 1992; Biemer and Lyberg, 2003). County-level crop-yield predictions can be produced for the current year when a flexible regression model has been fully trained and validated on past available data. A cross-validation method for time-dependent data was proposed by Burman et al. (1994), but the performance evaluation of out-of-sample predictions is a better practice when making final decisions on the family of models to use (Tashman, 2000).

To assess the uncertainty of the prediction error at the county level, the location-scale-shape paradigm (Rigby and Stasinopoulos, 2005) provides an established statistical methodology that allows modeling of the second moment of agricultural yields at the county level. However, the current implementations of most machine learning algorithms do not allow this framework to be fully explored. Thus, a simplified two-step approach can be considered instead (Gasser et al., 1986; Fan and Yao, 1998). First, a regression model predicting county-level yields is trained and validated. Then, its residuals are squared and used as the response variable of a second regression model providing a measure of uncertainty based on the same set of covariates used to predict the yield.

4. Case study in North Dakota

In 2018, North Dakota (ND) was one of the largest producers of wheat (mostly spring and durum), soybeans, canola, flax seeds, peas, oats, and honey in the US. Corn is usually the third major commodity crop cultivated in ND. Over the last couple of decades, there has been a shift to growing more soybeans and corn. Spring wheat is cultivated almost uniformly over the entire state, while soybeans and corn are respectively the dominant cultivations in the eastern and southern sides of the state. Using the RGB standard (Süsstrunk et al., 1999; Stone, 2005; Stone, 2016), Fig. 4 illustrates the historic crop extents in ND aggregated from 2008 to 2019 for corn (in red), soybeans (in green), and wheat (in blue). Swaths of orange in the southeast of the state correspond to areas that were frequently planted to either corn or soybeans, and swaths of purple seen in the south-central portion of the state represent areas that were frequently planted to either corn or wheat.

Due to its northern mid-continental geographical position, ND experiences cold winters, so the onset of fieldwork often lags behind the rest of the US. Delays also affect harvesting dates, which can impact yields. Usually, drier and warmer conditions in May allow farmers to make good planting progress once fieldwork is possible. Experts on the Agricultural Statistics Board (ASB) consider these events, along with survey summaries and remotely sensed values, when setting the official statistics.

Historical official yield estimates[3] for corn, soybeans, and wheat for each of the 53 counties in ND from 2008 to 2019 are considered in this study. This range of years coincides with the availability of the NASS CDLs having full national coverage (Boryan et al., 2011). CDLs are used to screen the remotely sensed spectral amplitudes of the red, near-IR, diurnal and nocturnal thermal bands. To establish proper support boundaries of the histograms, the minimum and maximum values of each band are computed using all pixels acquired over ND during the 2008–2019 time window. Once the remotely sensed values have been cleaned by masking out spurious signals in Google Earth Engine (Gorelick et al., 2017), the filtered data of each 8-day week from January 1st

[2] A "proper metric" satisfies all conditions required by definition to induce a specific class of topological spaces (e.g. see Engelking, 1989, Section 4.1).

[3] Historical estimates are retrieved from https://www.nass.usda.gov/Quick_Stats/ (accessed on November 24, 2020).

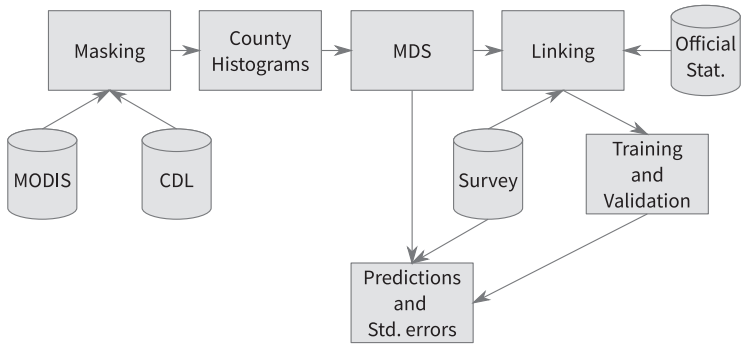

Fig. 3. Graphical representation of the algorithm to predict agricultural yields at the county level using density functions estimated from remotely sensed data as histograms.

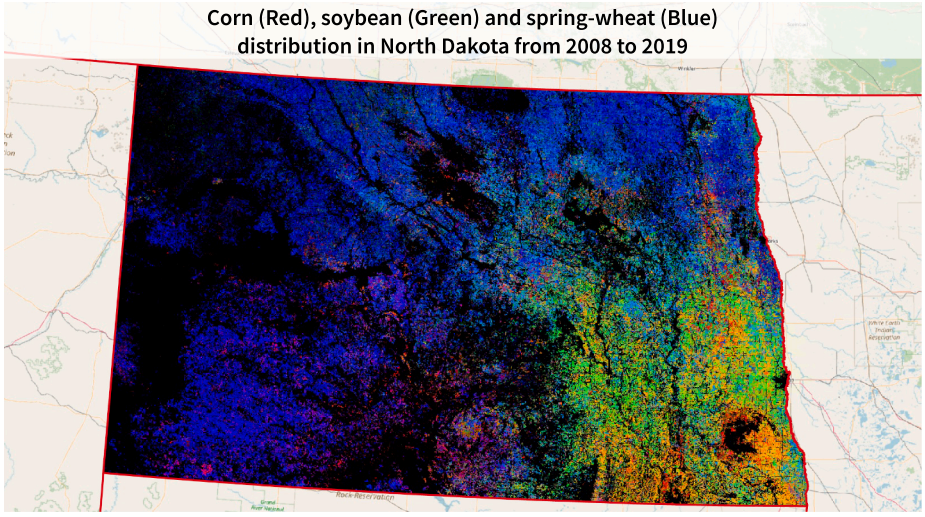

Fig. 4. Spatial distribution of the cultivation frequency for corn, soybean, and wheat fields in North Dakota between 2008 and 2019 using the RGB-color standard. Swaths of orange in the southeast correspond to areas frequently planted to either corn or soybeans. Swaths of purple seen in the south-central part of the state represent areas frequently planted to either corn or wheat.

to October 31st are summarized at the county level by computing 38 standard histograms (Pearson, 1895) of the four selected bands using 256 bins for all ND counties in any given year. Smoothed histograms (Fig. 5) are successively obtained as in Eq. (1) to better approximate the marginal density of the spectral amplitudes over the fields planted with corn, soybeans, and wheat. The 96,672 smoothed histograms per crop are successively used to produce numerical covariates for each county through the MDS algorithm. The MDS has been found to perform well on the root mean squared error (RMSE) of the predicted yields when extracting on average five features from each weekly distance matrix. The artificial values produced via MDS are then linked to historical NASS official statistics and other numerical variables including county-level survey summaries, and longitude/latitude coordinates of county centroids. Before training, validating, and testing several machine learning models, all 768 input variables are standardized to have zero mean and unit variance. Counties with missing historical yields are ignored in the analysis. Eventual missing data from the survey summaries are replaced with zeroes, which are consistent with the expectation of the standardized input variables.

Among several machine learning algorithms, the following are considered:

- Extreme gradient boosting (`xgbLinear`; Chen et al., 2020).
- CART without bagging (`rpart2`; Therneau and Atkinson, 2019) and with bagging (`treebag`; Lin and Li, 2014, Chapter 11.4).
- Conditional inference trees (`ctree2`; Hothorn et al., 2006).
- Evolutionary trees (`evtree`; Grubinger et al., 2014).
- Tree-based ensemble modelling (`nodeHarvest`; Meinshausen, 2010).
- RF (`rf`; Liaw and Wiener, 2002) and conditional RF (`cforest`; Hothorn et al., 2006; Strobl et al., 2007; Strobl et al., 2008).
- Cubist (`cubist`; Quinlan, 1992; Quinlan, 1993; Kuhn and Johnson, 2013, this model uses committees for boosting linear models fitted on an input dataset partitioned by regression trees, and it also handles residual dependencies using local smoothing via *k*-nearest neighbors).
- Relevance vector machines with a linear kernel (`rvmLinear`; Tipping, 2000; Camps-Valls et al., 2007) and polynomial kernels (`rvmPoly`; Fei et al., 2013).
- Model averaged ANN (`avNNet`; Ripley, 2007) and ANN with feature extraction (`pcaNNet`; Ravi and Pramodh, 2008).
- LSTM networks (`LSTM`; Hochreiter and Schmidhuber, 1997) as implemented by Tian et al. (2021).

Table 1 shows the aforementioned models, and their correspondent training and validation settings considered for this case study. All models, with the exception of the LSTM networks, have been implemented using the R-package `caret` and related package dependencies. The LSTM models have been implemented using the python library Keras with a Tensorflow backend.

Each of the models in Table 1 has its own strengths and weaknesses depending on both their mathematical formulation and fitting

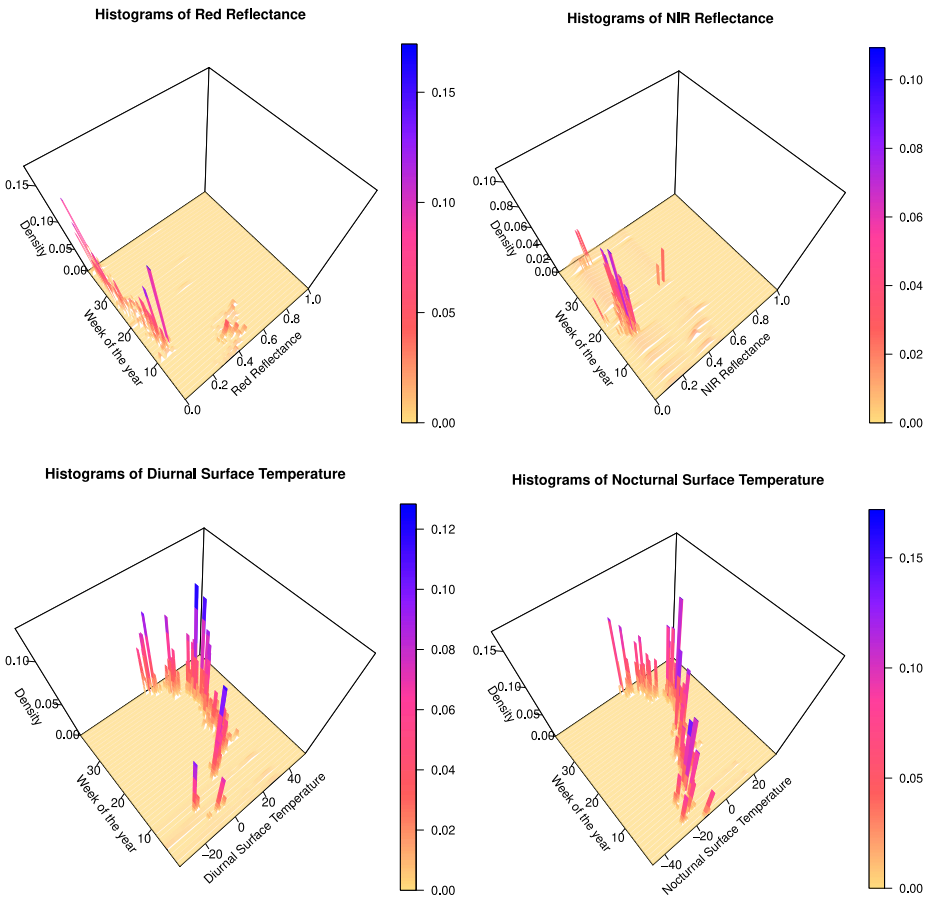

Fig. 5. Smoothed weekly histograms obtained for red (top-left), near-IR (top-right), and the two thermal bands (on the bottom) over the corn fields of Barnes County (ND) in 2008.

algorithm. Different models are more robust with respect to outliers, shifting domains (mainly due to the use of the realizations from non-stationary random processes as covariates; Ben-David et al., 2010; Bobu et al., 2018), or model misspecification (either due to the exclusion of important covariates or improper choice of model family). To select the best performing model in terms of robustness and prediction accuracy, the minimax criterion (Wald, 1949; Savage, 1951; Robbins, 1964; Berger, 2013) was applied. This criterion summarizes the performance of a single model in terms of the largest RMSE, expressed in bushels per acre – bu/ac) among all years modeled (2012 and 2019), and the model with the smallest maximum RMSE is selected. The RMSE is computed using the model-based predictions and the official NASS yield values of the counties in ND.

Before fitting the chosen regression models, the data are split into three parts (Hastie et al., 2009, Chapter 7): the statistical units (i.e., the county-level data) belonging to the training set (used to estimate the parameters), the validation set (used to mitigate and/or avoid overfitting) are randomly selected by partitioning the data, and the remaining statistical units populate the test set (used to evaluate the extrapolation performance of the model on new data points). In particular, training and 10-fold cross-validation (Molinaro et al., 2005; Kuhn and Johnson, 2013) are performed sequentially on the data considered as information from the previous years. Training and 10-fold cross-validation are then replicated 10 times (due to computational limitations) on randomized partitions of the data to assess the distribution of the RMSE for guiding model selection (Kim, 2009). On the other hand, prediction accuracy is evaluated using only the data from the following year (considered as current). Table 2 summarizes the data subsets and the numbers of statistical units.

To compare the predictive performances of each model family, the maximum RMSE computed between 2012 and 2019 is shown in Table 3. This table also separates the results for each commodity showing the worst performance during training, validation, and testing (or extrapolation). Most models have produced better results on training (with the exception of `avNNet` for soybeans, which performed better during validation). Based on validation results, `cubist` represents the best model choice for predicting yields, even if `xgbLinear`, `cforest`, `ctree2`, `evtree`, `nodeHarvest`, `rf`, `rpart2` and `treebag` have produced comparable results with validation RMSEs lower than 16.43 bu/ac for corn, 3.39 bu/ac for soybeans, and 3.312 bu/ac for wheat. In particular, `cubist` has been the most robust model on the testing sets for corn (with RMSE of 10.3 bu/ac) and soybeans (with RMSE of 5.35 bu/ac). However, `xgbLinear`, `cforest`, `ctree2`, `rf` and `treebag` have outperformed `cubist` in robustness producing RMSE lower than 4.24 bu/ac on the testing sets for wheat, where `xgbLinear` attained the minimax RMSE at 3.46 bu/ac.

To test unexplained autocorrelations among the model residuals, Moran's I test (Moran, 1950; Kelejian and Prucha, 2001) has been computed by relating spatio-temporal information and the differences between the official yields and fitted values produced by Cubist using all data from the training and validation sets. Spatio-temporal information among data points has been encoded into a distance matrix computed by summing the temporal lags (in years) and the geodesic distances (in kilometers) between each pair of county centroids (Karney, 2013). The results in Table 4 show a significant absence of unexplained autocorrelation; therefore, the residuals can be considered as independent with the exception of those produced for soybeans in 2018 and 2019.

Table 5 presents the performances of the model families used to assess the uncertainty of county-level yields. Based on validation results, `avNNet` represents the best model choice for computing the uncertainty

Table 1
Machine-learning algorithms and validation settings used in the study.

Algorithm	Setting Notes
Extreme gradient boosting	Fitted using combinations of boosting iterations, learning rates, and elastic-net penalties. The combinations include boosting iterations ranging from 5 to 200 with step in increments of 5, learning rates in the set {0.1, 0.01, 0.001}, and elastic-net penalty using either 0, 1, or 2 as values for LASSO and ridge weighting.
CART	Trained and validated with and without bagging, where the maximum tree-depth was allowed to range from 5 to 15 nodes.
Conditional inference trees	Used to test conditional splits with at 90%, 95% and 99% significance levels.
Evolutionary trees	Trained with genetic algorithms that randomly decrease the model complexity using a penalty assuming values 0.25, 0.75, and 1.
Tree-based ensemble modelling	Set to consider two-ways interactions, as well as both deterministic and stochastic splitting criteria.
RF/ Conditional RF	Both were validated considering either 10, 100, 1,000 and 10,000 regression trees.
Cubist	Validated using 5, 10, 25, 50, and 100 boosting committees, combined with a limited amount of nearest-neighbors ranging from 0 to 9.
Relevance vector machines	Fitted using linear kernel and second-order polynomial kernels.
Model averaged ANN/ ANN with feature extraction	Both were constructed using a single hidden layer consisting of 2, 5, and 7 nodes linearly combined to the output layer. Fitted using 0.1, 0.01 and 0.001 decay values.
LSTM networks	The structure of the layers proposed by Tian et al. (2021) has been identically replicated. The default settings of the python library Keras for the Adam aglorithm (Kingma and Ba, 2014) have been used to fit the model.

Table 2
Summary of data partitions (training, validation and test sets).

Years used for training and validation	Year used for testing
(191 tr. and 21 valid. units) 2008, 2009, …, 2011	(53 units) 2012
(239 tr. and 26 valid. units) 2008, 2009, …, 2012	(53 units) 2013
(286 tr. and 32 valid. units) 2008, 2009, …, 2013	(53 units) 2014
(334 tr. and 37 valid. units) 2008, 2009, …, 2014	(53 units) 2015
(382 tr. and 42 valid. units) 2008, 2009, …, 2015	(53 units) 2016
(429 tr. and 48 valid. units) 2008, 2009, …, 2016	(53 units) 2017
(477 tr. and 53 valid. units) 2008, 2009, …, 2017	(53 units) 2018
(525 tr. and 58 valid. units) 2008, 2009, …, 2018	(53 units) 2019

of yields, and only `pcaNNet` has produced similar results with RMSEs lower than 45.3 squared-bushels per squared-acre (bu^2/ac^2) for corn, 0.5 bu^2/ac^2 for soybeans, and 0.65 bu^2/ac^2 for wheat. Most model families have provided similar results on the testing sets producing minimax RMSE between 513.12 and 539.44 bu^2/ac^2 for corn, between 113.85 and 115.01 bu^2/ac^2 for soybeans, and between 24.60 and 36.76 bu^2/ac^2 for wheat. The most robust models evaluated on the testing sets are `treebag` for corn, and `LSTM` for soybeans and wheat.

Standard errors (SE) are computed using the square root of the positive part of the outputs produced by the `avNNet` models. The distribution of model-based SEs is then summarized via the median and the maximum values calculated at the state level (see Table 6). Through the years, the distribution of the county-level SEs has been quite stable for corn; however, the variability of soybean and wheat yields increased in the recent years of the study window. Between 2012 and 2019, the SEs of corn yields did not exceed 9.94 bu/ac, the maximum SE of soybean yields at the county level steadily increased from 0.68 bu/ac to 3.93 bu/ac, and the largest SEs of wheat yields fluctuated between 1.6 bu/ac and

Table 4
Results of the autocorrelation test based on Moran's I statistics computed with the residuals produced by the Cubist model (bold p-values highlight unexplained residual autocorrelation).

	Year	Observed	Expected	Std.Dev.	P-value
Corn	2012	−0.067	−0.0047	0.038	0.098
	2013	−0.036	−0.0038	0.029	0.267
	2014	−0.036	−0.0032	0.024	0.175
	2015	−0.039	−0.0027	0.020	0.080
	2016	−0.017	−0.0024	0.018	0.404
	2017	−0.028	−0.0021	0.016	0.091
	2018	−0.027	−0.0019	0.014	0.070
	2019	−0.026	−0.0017	0.013	0.055
Soybeans	2012	0.015	−0.0057	0.041	0.615
	2013	−0.024	−0.0044	0.031	0.514
	2014	−0.036	−0.0036	0.025	0.198
	2015	−0.033	−0.0030	0.021	0.141
	2016	−0.035	−0.0026	0.017	0.057
	2017	−0.004	−0.0023	0.016	0.933
	2018	0.040	−0.0020	0.014	**0.004**
	2019	0.060	−0.0018	0.013	**0.000**
Wheat	2012	−0.014	−0.0047	0.037	0.807
	2013	−0.024	−0.0038	0.029	0.482
	2014	−0.017	−0.0032	0.024	0.550
	2015	−0.006	−0.0027	0.020	0.885
	2016	−0.020	−0.0024	0.018	0.311
	2017	−0.025	−0.0021	0.016	0.136
	2018	−0.011	−0.0019	0.014	0.499
	2019	0.009	−0.0017	0.013	0.387

Table 3
Maximum RMSEs for the yields predicted between 2012 and 2019. The selected model is based on the validation results (in bold), while the best extrapolation performances are highlighted in italic.

Model	Results for corn			Results for soybeans			Results for wheat		
	Training	Validation	Testing	Training	Validation	Testing	Training	Validation	Testing
`avNNet`	18.541	23.184	46.374	4.879	4.266	17.557	1.843	7.295	27.147
`cforest`	8.176	12.708	17.569	1.982	2.727	6.241	1.307	1.920	3.707
`ctree2`	14.751	15.905	21.032	2.903	2.975	6.989	1.865	2.640	4.205
`cubist`	4.352	**11.306**	*10.304*	1.112	**2.477**	*5.351*	0.758	**1.064**	4.239
`evtree`	9.466	15.874	21.662	2.525	3.237	6.831	1.715	2.458	4.444
`LSTM`	41.440	39.022	56.864	9.868	9.514	12.028	15.792	16.957	19.754
`nodeHarvest`	13.516	15.538	22.810	2.964	3.043	7.727	3.249	3.312	5.594
`pcaNNet`	22.219	24.596	32.058	5.032	6.480	11.209	6.387	8.927	11.624
`rf`	5.201	13.230	17.303	1.099	2.701	6.235	0.723	1.890	3.685
`rpart2`	10.193	16.430	18.958	2.711	3.388	7.320	2.531	3.034	4.550
`rvmLinear`	17.672	63.888	>1000.000	6.255	26.790	339.557	7.176	24.011	510.301
`rvmPoly`	8.441	24.060	296.440	2.407	5.772	83.716	2.004	7.887	124.297
`treebag`	8.404	13.883	18.402	2.333	2.853	6.407	2.037	2.385	4.185
`xgbLinear`	0.045	12.522	16.320	0.089	2.889	7.067	0.055	1.425	*3.462*

Table 5
Maximum RMSEs for the variances of yields computed between 2012 and 2019. The selected model is based on the validation results (in bold), while the best extrapolation performances are highlighted in italic.

Model	Results for corn			Results for soybeans			Results for wheat		
	Training	Validation	Testing	Training	Validation	Testing	Training	Validation	Testing
avNNet	49.785	**40.199**	514.987	7.842	**0.176**	114.451	0.142	**0.522**	36.759
cforest	42.493	48.533	516.892	7.570	5.339	114.803	2.083	1.943	29.052
ctree2	57.727	51.992	520.953	7.795	6.077	114.806	2.385	2.119	28.861
cubist	31.217	50.033	515.418	4.670	6.668	114.806	1.299	2.171	28.999
evtree	45.723	58.120	520.127	7.670	6.479	114.806	2.105	2.470	28.854
LSTM	78.880	108.229	516.234	5.513	2.475	*113.848*	3.039	1.249	*24.595*
nodeHarvest	51.481	49.374	517.774	8.418	5.344	114.804	2.210	2.012	29.052
pcaNNet	49.906	45.254	518.790	8.170	0.498	115.009	0.116	0.649	29.345
rf	26.136	51.751	516.823	4.265	5.648	114.802	1.266	2.051	28.829
rvmLinear	40.664	61.355	539.441	5.818	8.558	114.815	1.787	2.584	30.771
rvmPoly	0.010	57.156	521.947	0.857	6.913	114.816	0.007	2.383	29.315
treebag	42.378	49.921	*513.414*	7.532	5.755	114.805	2.064	2.080	28.856
xgbLinear	28.546	56.533	521.153	3.831	8.175	114.783	1.286	2.279	29.049

Table 6
Summary statistics (median and maximum values) of the county-level standard errors (bushels per acre) produced by the neural-network-averaging models.

Year	Std. err. for corn		Std. err. for soybeans		Std. err. for wheat	
	Median	Maximum	Median	Maximum	Median	Maximum
2012	2.441	9.941	0.039	0.678	0.214	2.187
2013	2.468	8.133	0.064	1.276	0.157	1.598
2014	1.752	8.629	0.058	1.233	0.116	2.037
2015	2.788	7.287	0.064	1.839	0.151	1.930
2016	2.258	7.696	0.080	2.150	0.164	3.311
2017	2.505	7.901	0.983	3.446	0.195	3.760
2018	3.220	8.736	0.890	3.565	0.238	5.959
2019	3.131	8.022	0.928	3.925	0.254	6.531

6.53 bu/ac and had an upward trend.

The avNNet method for assessing yield uncertainties is then evaluated by testing for unexplained residual autocorrelation based on the same spatio-temporal relationships considered for the previous Moran's I tests. Based on the results of Table 7, the residuals of the uncertainty models for wheat are assumed to be independent. However, the last three models of soybeans and four models (among eight) for corn have produced residuals with unexplained autocorrelations that can be addressed by including spatio-temporal dynamics in the chosen models.

Table 7
Results of the autocorrelation test based on Moran's I statistics computed with the residuals produced by the avNNet model for standard errors (bold p-values highlight unexplained residual autocorrelation).

	Year	Observed	Expected	Std.Dev.	P-value
Corn	2012	0.105	−0.0047	0.036	**0.002**
	2013	0.051	−0.0038	0.028	0.052
	2014	0.042	−0.0032	0.023	**0.048**
	2015	0.060	−0.0027	0.020	**0.002**
	2016	0.024	−0.0024	0.016	0.093
	2017	0.024	−0.0021	0.013	0.051
	2018	0.012	−0.0019	0.013	0.276
	2019	0.029	−0.0017	0.012	**0.008**
Soybeans	2012	−0.006	−0.0057	0.038	0.997
	2013	−0.016	−0.0044	0.028	0.687
	2014	−0.018	−0.0036	0.023	0.538
	2015	−0.013	−0.0030	0.018	0.589
	2016	0.008	−0.0026	0.015	0.490
	2017	0.017	−0.0023	0.009	**0.027**
	2018	0.070	−0.0020	0.008	**0.000**
	2019	0.030	−0.0018	0.005	**0.000**
Wheat	2012	−0.009	−0.0047	0.013	0.740
	2013	−0.015	−0.0038	0.020	0.577
	2014	−0.016	−0.0032	0.008	0.083
	2015	−0.001	−0.0027	0.011	0.864
	2016	−0.013	−0.0024	0.012	0.349
	2017	−0.015	−0.0021	0.009	0.155
	2018	−0.001	−0.0019	0.009	0.950
	2019	−0.003	−0.0017	0.004	0.765

The predictions produced by the proposed methodology are then compared with those produced by using only survey data, and a standard remote sensing approach (as explained in Section 2). Tables 8 and 9, respectively, report the RMSE and the mean absolute percentage error (MAPE) from the official statistics (Chen and Yang, 2004). From a first analysis of these tables, the current remote sensing methodology is less precise than the proposed predictive approach. Although the RMSEs used to compare the three methods are quite stable across time, the survey results produced for corn in 2012, 2015 and 2019, and for soybeans (with the exception of 2016 and 2017) had larger RMSEs than those produced by the proposed methodology. However, the survey achieves better RMSEs for wheat than the approaches using remote sensing data. Furthermore, a substantial improvement of soybean results is noted when the average MAPE of the survey passed from 12.87% (before 2016) to 4.71% (after 2016). Nonetheless, the same improvements in MAPEs are not observed when using the other two methods, which remained quite stable across time. Using the minimax approach, the proposed methodology would have been selected for corn and soybeans with a minimax MAPE of 6.38% and 7.08% respectively, and a pure survey approach would have been selected instead for wheat with a minimax MAPE of 8.31%.

5. Conclusion

Remotely sensed data provide a detailed description of the evolution of several physical phenomena happening on the Earth surface. Several traditional methods, such as radial smoothing or the conditional average within administrative boundaries, provide enough information to summarize the complexity of these data by accounting only for the first moment of the phenomena observed. Due to their simplicity, these methods are appealing when linking official statistics and other variables obtained by summarizing remotely sensed data. However, empirical densities provide a full description of a stochastic process to be used as an input in a modelling framework. From a statistical point of view, the estimation of empirical densities suffers computational drawbacks due to high-dimensionality, optimal search of matrix-valued bandwidths, nonlinear transformations, and data sample-size (which mostly depends on the spatio-temporal resolutions of satellite imagery).

To address these problems, inference on the physical phenomena of interest is performed on the marginal distribution of each variable by conditioning on specific areas within administrative boundaries. The dynamical evolution through time of the estimated distributions can then be viewed as functional data. These are known to provide more information than scalar covariates, thus helping to improve model selection, and prediction analyses (Ramsay and Silverman, 2007).

Table 8
Comparison between survey only (SO), current remote sensing (RS) and proposed methodology (PM) using the RMSE (bu/ac). The values in italics correspond the maximum by column, while the values highlighted in bold correspond to the minimax solution by crop.

	Max. RMSE for corn			Max. RMSE for soybeans			Max. RMSE for wheat		
Year	SO	RS	PM	SO	RS	PM	SO	RS	PM
2012	7.78	22.08	6.28	*8.10*	7.54	2.98	1.69	6.02	3.62
2013	2.60	*26.68*	5.59	7.80	5.69	1.77	0.70	4.79	1.34
2014	4.62	13.80	5.65	6.32	4.74	1.87	1.76	3.91	3.02
2015	12.66	13.73	**10.30**	5.28	4.79	1.89	1.13	5.04	1.63
2016	2.88	16.33	4.52	3.56	*8.34*	**5.35**	1.10	6.12	1.57
2017	6.53	20.48	6.87	3.13	5.72	3.15	**3.84**	*8.92*	*4.24*
2018	7.18	21.51	8.56	4.11	6.06	3.34	1.51	4.59	1.52
2019	*14.72*	21.74	9.58	7.03	6.83	2.31	1.63	4.70	2.00

Table 9
Comparison between survey only (SO), current remote sensing (RS) and proposed methodology (PM) using the MAPE (%). The values in italics correspond the maximum by column, while the values highlighted in bold correspond to the minimax solution by crop.

	Max. MAPE for corn			Max. MAPE for soybeans			Max. MAPE for wheat		
Year	SO	RS	PM	SO	RS	PM	SO	RS	PM
2012	2.77	*24.56*	4.93	*14.66*	*19.48*	**7.08**	3.72	10.26	6.98
2013	1.95	21.65	4.30	14.27	17.17	4.68	1.48	8.87	2.48
2014	2.90	11.05	3.80	11.41	13.59	5.53	2.95	7.44	5.45
2015	6.25	12.27	5.88	11.15	16.76	4.79	0.57	9.08	3.23
2016	1.87	11.57	2.52	1.15	12.71	2.08	1.73	11.29	2.66
2017	5.57	24.40	**6.38**	6.69	16.33	6.15	**8.31**	*25.72*	*8.42*
2018	5.44	15.50	6.27	3.88	17.24	3.52	2.88	8.32	2.78
2019	*7.69*	14.76	4.56	7.12	17.27	3.49	3.09	8.23	3.93

However, due to their nature, most of the traditional methods do not capture the full complexity expressed by these data. For a given metric distance between two functions, MDS is an algorithmic solution that can map functional data into a vector-valued space. This process generates a desired number of numerical features that summarize most of the information embedded into the estimated empirical distributions. These features are successively linked to survey summaries and official statistics to enhance the analyses as shown in Table 8 and 9.

This approach has been shown to be capable of generating artificial covariates that can be effective when used in predictive models for crop yield predictions at the county level. However, when enriching the model by including higher moments of the distributions inferred at the county level, the proposed methodology cannot downscale crop yield predictions to finer spatial resolutions (such as at the pixel level). Depending on the availability of satellite imagery, finer resolution remotely sensed data can be linked to precision agriculture data for more detailed analyses.

Future research will focus on studying the robustness of this approach to the domain-shift problem. In particular, model fitness will be evaluated by accounting for ongoing advancements in both survey methods and data acquisition technologies based on satellites. Further modelling extensions may include temporal dynamics of the residuals. These often refine the model results by addressing unexplained spatio-temporal autocorrelations (e.g. see Zhang et al., 2018). Furthermore, the proposed methodology can be tested on its ability to simultaneously process a variety of county-level information retrieved from all states in the conterminous US.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The findings and conclusions in this paper are those of the authors and should not be construed to represent any official USDA, or US Government determination or policy. This research was supported by the intramural research program of the USDA, NASS. Resources provided by the SCINet project of USDA Agricultural Research Service, ARS project number `0500-00093-001-00-D` were used.

References

Barnes, W.L., Xiong, X., Salomonson, V.V., 2003. Status of terra MODIS and aqua MODIS. Adv. Space Res. 32, 2099–2106.
Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W., 2010. A theory of learning from different domains. Mach. Learn. 79, 151–175.
Berger, J.O., 2013. Statistical decision theory and Bayesian analysis. Springer Science & Business Media.
Biemer, P., Lyberg, L., 2003. Introduction to Survey Quality. Wiley, New York.
Bobu, A., Tzeng, E., Hoffman, J., Darrell, T., 2018. Adapting to continuously shifting domains. URL https://openreview.net/forum?id=BJsBjPJvf.
Boj, E., Caballé, A., Delicado, P., Esteve, A., Fortiana, J., 2016. Global and local distance-based generalized linear models. TEST 25, 170–195.
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 International 26, 341–358.
Breiman, L., 2001. Random forests. Machine learning 45, 5–32.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees. The Wadsworth statistics/probability series. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.
Burman, P., Chow, E., Nolan, D., 1994. A cross-validatory method for dependent data. Biometrika 81, 351–358.
Cai, Y., Moore, K., Pellegrini, A., Elhaddad, A., Lessel, J., Townsend, C., Solak, H., Semret, N., 2017. Crop yield predictions-high resolution statistical model for intra-season forecasts applied to corn in the us. In: 2017 Fall Meeting. URL https://www.gro-intelligence.com/yield-model-pdf/us-corn.
Camps-Valls, G., Martínez-Ramón, M., Rojo-Alvarez, J.L., Muñoz-Marí, J., 2007. Nonlinear system identification with composite relevance vector machines. IEEE Signal Process. Lett. 14, 279–282.
Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., 2020. xgboost: Extreme Gradient Boosting. URL https://CRAN.R-project.org/package=xgboost r package version 1.2.0.1.
Chen, Z., Yang, Y., 2004. Assessing forecast accuracy measures. Preprint Series 2010, 2004–2010.
Chipanshi, A., Zhang, Y., Kouadio, L., Newlands, N., Davidson, A., Hill, H., Warren, R., Qian, B., Daneshfar, B., Bedard, F., et al., 2015. Evaluation of the integrated canadian crop yield forecaster (iccyf) model for in-season prediction of crop yield across the canadian agricultural landscape. Agric. For. Meteorol. 206, 137–150.
Cochran, W., 1938. Discussion: crop estimation and its relation to agricultural meteorology. Suppl. J.R. Stat. Soc. 5, 12–20.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20, 273–297.

Cruze, N.B., Erciulescu, A.L., Nandram, B., Barboza, W.J., Young, L.J., et al., 2019. Producing official county-level agricultural estimates in the united states: Needs and challenges. Statistical Science 34, 301–316.
Cuadras, C., Arenas, C., 1990. A distance-based regression model for prediction with mixed data. Communications in Statistics A - Theory and Methods 19, 2261–2279.
De Wit, A.d., Boogaard, H., Van Diepen, C., 2005. Spatial resolution of precipitation and radiation: the effect on regional crop yield forecasts. Agric. For. Meteorol. 135, 156–168.
Doraiswamy, P.C., Sinclair, T.R., Hollinger, S., Akhmedov, B., Stern, A., Prueger, J., 2005. Application of MODIS derived parameters for regional crop yield assessment. Remote sensing of environment 97, 192–202.
Elavarasan, D., Vincent, D.R., Sharma, V., Zomaya, A.Y., Srinivasan, K., 2018. Forecasting yield by integrating agrarian factors and machine learning models: A survey. Computers and Electronics in Agriculture 155, 257–282.
Engelking, R., 1989. General topology (Revised and completed ed.). Heldermann Verlag.
Erciulescu, A.L., Cruze, N.B., Nandram, B., 2020. Statistical challenges in combining survey and auxiliary data to produce official statistics. Journal of Official Statistics 36, 63–88.
Ertel, W., 2018. Introduction to artificial intelligence. Springer.
Fan, J., Yao, Q., 1998. Efficient estimation of conditional variance functions in stochastic regression. Biometrika 85, 645–660.
Fei, H., Jinwu, X., Min, L., Jianhong, Y., 2013. Product quality modelling and prediction based on wavelet relevance vector machines. Chemometrics and Intelligent Laboratory Systems 121, 33–41.
Gao, F., Anderson, M., Daughtry, C., Johnson, D., 2018. Assessing the variability of corn and soybean yields in central iowa using high spatiotemporal resolution multi-satellite imagery. Remote Sensing 10, 1489.
Gasser, T., Sroka, L., Jennen-Steinmetz, C., 1986. Residual variance and residual pattern in nonlinear regression. Biometrika 73, 625–633.
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.
Grubinger, T., Zeileis, A., Pfeiffer, K.-P., 2014. evtree: Evolutionary learning of globally optimal classification and regression trees in R. J. Stat. Softw. 61, 1–29 http://www.jstatsoft.org/v61/i01/.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer Series in Statistics, Second ed. Springer, New York Inc., New York, NY, USA.
Hatfield, J., 1983. Remote sensing estimators of potential and actual crop yield. Remote Sens. Environ. 13, 301–311.
Hayes, M., Decker, W., 1996. Using noaa avhrr data to estimate maize production in the united states corn belt. Remote Sensing 17, 3189–3200.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
Hoover, D.R., Rice, J.A., Wu, C.O., Yang, L.-P., 1998. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika 85, 809–822.
Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., Van Der Laan, M.J., 2006. Survival ensembles. Biostatistics 7, 355–373.
Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics 15, 651–674.
Irwin, J.O., 1938. Crop estimation and its relation to agricultural meteorology. Supplement to the Journal of the Royal Statistical Society 5, 1–45.
James, G.M., 2002. Generalized linear models with functional predictors. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, 411–432.
James, G.M., Hastie, T.J., 2001. Functional linear discriminant analysis for irregularly sampled curves. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63, 533–550.
Jiang, H., Hu, H., Zhong, R., Xu, J., Xu, J., Huang, J., Wang, S., Ying, Y., Lin, T., 2020. A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the us corn belt at the county level. Global change biology 26, 1754–1766.
Johnson, D., 2014. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the united states. Remote Sens. Environ. 141, 116–128.
Johnson, D.M., 2016. A comprehensive assessment of the correlations between field crop yields and commonly used MODIS products. Int. J. Appl. Earth Obs. Geoinf. 52, 65–81.
Justice, C., Townshend, J., Vermote, E., Masuoka, E., Wolfe, R., Saleous, N., Roy, D., Morisette, J., 2002. An overview of MODIS Land data processing and product status. Remote sensing of Environment 83, 3–15.
Karney, C.F., 2013. Algorithms for geodesics. J. Geodesy 87, 43–55.
Kelejian, H.H., Prucha, I.R., 2001. On the asymptotic distribution of the moran i test statistic with applications. Journal of Econometrics 104, 219–257.
Kim, J.-H., 2009. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational statistics & data analysis 53, 3735–3745.
Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, https://arxiv.org/abs/1412.6980.
Kuhn, M., Johnson, K., 2013. Applied predictive modeling, volume 26. Springer.
Lessler, J., Kalsbeek, W., 1992. Nonsampling Error in Surveys. Wiley, New York.
Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2, 18–22 https://CRAN.R-project.org/doc/Rnews/.
Lin, D., Ying, Z., 2001. Semiparametric and nonparametric regression analysis of longitudinal data. Journal of the American Statistical Association 96, 103–126.
Lin, H., Li, M., 2014. Introduction to Data Science.
Matis, J., Saito, T., Grant, W., Iwig, W., Ritchie, J., 1985. A markov chain approach to crop yield forecasting. Agricultural systems 18, 171–187.
McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 5, 115–133.
Meinshausen, N., 2010. Node harvest. The Annals of Applied Statistics, pp. 2049–2072.
Molinaro, A.M., Simon, R., Pfeiffer, R.M., 2005. Prediction error estimation: a comparison of resampling methods. Bioinformatics 21, 3301–3307.
Moran, P.A., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.
Moyeed, R., Diggle, P.J., 1994. Rates of convergence in semi-parametric modelling of longitudinal data. Australian Journal of Statistics 36, 75–93.
Nandram, B., Berg, E., Barboza, W., 2014. A hierarchical bayesian model for forecasting state-level corn yield. Environmental and ecological statistics 21, 507–530.
Nielsen, F., 2010. A family of statistical symmetric divergences based on jensen's inequality. arXiv preprint arXiv:1009.4004.
Pearson, K., 1895. X. contributions to the mathematical theory of evolution.—ii. skew variation in homogeneous material. Philosophical Transactions of the Royal Society of London. A, 343–414.
Pearson, R.K., 2005. Mining imperfect data: Dealing with contamination and incomplete records. SIAM.
Quinlan, J.R., 1992. Learning with continuous classes. Hobart, Australia, pp. 343–348.
Quinlan, J.R., 1993. Combining instance-based and model-based learning. In: Proceedings of the tenth international conference on machine learning, pp. 236–243.
Ramsay, J.O., Silverman, B.W., 2007. Applied functional data analysis: methods and case studies. Springer.
Rasmussen, M.S., 1997. Operational yield forecast using avhrr ndvi data: reduction of environmental and inter-annual variability. Int. J. Remote Sens. 18, 1059–1077.
Ravi, V., Pramodh, C., 2008. Threshold accepting trained principal component neural network and feature subset selection: Application to bankruptcy prediction in banks. Applied Soft Computing 8, 1539–1548.
Rigby, R.A., Stasinopoulos, D.M., 2005. Generalized additive models for location, scale and shape. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) 54, 507–554.
Ripley, B.D., 2007. Pattern recognition and neural networks. Cambridge University Press.
Robbins, H., 1964. The empirical bayes approach to statistical decision problems. Ann. Math. Stat. 35, 1–20.
Savage, L.J., 1951. The theory of statistical decision. Journal of the American Statistical association 46, 55–67.
Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society, Series B 53, 683–690.
Silverman, B.W., 1986. Monographs on statistics and applied probability. In: Density estimation for statistics and data analysis, p. 26.
Stone, M., 2016. A field guide to digital color. CRC Press.
Stone, M.C., 2005. Representing colors as three numbers [color graphics]. IEEE Comput. Graphics Appl. 25, 78–85.
Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC bioinformatics 9, 307.
Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics 8, 25.
Sun, J., Di, L., Sun, Z., Shen, Y., Lai, Z., 2019. County-level soybean yield prediction using deep cnn-lstm model. Sensors 19, 4363.
Süsstrunk, S., Buckley, R., Swen, S., 1999. Standard RGB color spaces. In: Color and Imaging Conference. Society for Imaging Science and Technology volume 1999, pp. 127–134.
Tashman, L.J., 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. International journal of forecasting 16, 437–450.
Therneau, T., Atkinson, B., 2019. rpart: Recursive Partitioning and Regression Trees. https://CRAN.R-project.org/package=rpart r package version 4.1-15.
Tian, H., Wang, P., Tansey, K., Zhang, J., Zhang, S., Li, H., 2021. An lstm neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the guanzhong plain, pr china. Agric. For. Meteorol. 85 https://doi.org/10.1016/j.agrformet.2021.108629.
Tipping, M.E., 2000. The relevance vector machine. In Advances in neural information processing systems, pp. 652–658.
Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote sensing of Environment 8, 127–150.
Wald, A., 1949. Statistical decision functions. The Annals of Mathematical Statistics, pp. 165–205.
Walker, G., Sigman, R., 1982. The use of landsat for county estimates of crop areas: evaluation of the huddleston-ray and the battese-fuller estimators. SRS staff report (USA). no. AGES 820909.
Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman and Hall, London.
Wang, J.C., Holan, S.H., Nandram, B., Barboza, W., Toto, C., Anderson, E., 2012. A bayesian approach to estimating agricultural yield based on multiple repeated surveys. Journal of agricultural, biological, and environmental statistics 17, 84–106.
Wu, C.O., Chiang, C.-T., Hoover, D.R., 1998. Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. Journal of the American statistical Association 93, 1388–1402.
You, J., Li, X., Low, M., Lobell, D., Ermon, S., 2017. Deep gaussian process for crop yield prediction based on remote sensing data. In: Thirty-First AAAI conference on artificial intelligence, pp. 4559–4565.
Young, L., 2019. Agricultural crop forecasting for large geographical areas. Annual Review of Statistics and Its Application 6, 173–196.
Zambom, A.Z., Ronaldo, D., 2013. A review of kernel density estimation with applications to econometrics. International Econometric Review 5, 20–42.

Zeger, S.L., Diggle, P.J., 1994. Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. Biometrics 689–699.
Zhang, J., Zheng, Y., Qi, D., Li, R., Yi, X., Li, T., 2018. Predicting citywide crowd flows using deep spatio-temporal residual networks. Artif. Intell. 259, 147–166.
Zhao, Y., Cai, Y., Lessel, J., Toure, A., Semret, N., 2017. Crop yield predictions-high resolution statistical model for intra-season forecasts applied to soybeans in the united states. In: 2017 Fall Meeting. URL https://www.gro-intelligence.com/yield-model-pdf/us-soybean.