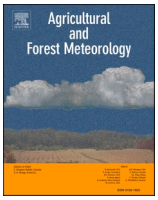
Agricultural and Forest Meteorology 300 (2021) 108317

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

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



Modelling wheat yield with antecedent information, satellite and climate data using machine learning methods in Mexico

Diego Gómez *, Pablo Salvador, Julia Sanz, José Luis Casanova

Remote Sensing Laboratory (LATUV), University of Valladolid, Paseo de Belen 11, 47011 Valladolid, Spain

ARTICLE INFO

Keywords:
Climate data
Food security
Machine learning
Satellite data
Wheat yield

ABSTRACT

Wheat is one of the most important cereal crops in the world, and its demand is expected to increase about 60% by 2050. Thus, appropriate and reliable yield forecasts are fundamental to ensure price stability and food security around the globe. In this study, we developed a Machine Learning (ML) approach to combine satellite and climate data with antecedent wheat yield information (YieldBaseLine) from 2004 – 2018, at municipal level, in Mexico. We compared the performance of four linear (generalized linear model –glm-, ridge regression –ridge-, lasso, partial least squares –pls-) and four non-linear algorithms (k-nearest neighbours –kknn-, support vector machine radial –svmR-, extreme gradient boosting -xgbTree- and random forest –rf) before harvest time. Additionally, we evaluated their performance using five different feature selection scenarios (No FS, FS = 0.9, FS = 0.75, FS = 0.9 and YieldBaseLine). The models were independently tested using two different approaches: random sampling and selective sampling. In the random sampling, the non-linear models performed generally better under the FS = 0.5 scenario, whereas the non-linear models were less sensitive to feature reduction. The results also evidenced the capacity of the YieldBaseLine predictor, combined with satellite and climate data, to address the inter-annual and spatial variability in the study area. The highest prediction accuracy was obtained by the rf method (No FS) with $R^2 = 0.84$. To further prove the model's operability in a simulated real-case scenario, we held out the last year records (2018) to test the models. The best performing model was again the rf ($R^2 = 0.81$). This study proposes a robust methodology to model crop yield (at large scale) and it may be used with operative purposes. Therefore, it can be of interest to decision and law makers, producers, authorities or the wheat industry. In addition, it can help to establish appropriate food security and trading policies. A similar approach can be applied to other regions or crops.

1. Introduction

Wheat, rice and maize are the most important cereal crops in the world (FAO, 2016). They constitute the largest source of calories in developing countries and approximately 30% of the total calories consumed in developed countries (Awika, 2011). While maize and rice are sensitive to water availability and low temperatures, wheat is more resilient to these constraints hence; it is grown in temperate and warmer regions (Awika, 2011). Some estimates suggest that wheat alone makes up to 20% of the food calories and daily proteins consumed by 4.5 billion people (FAO, 2012). World population is projected to be around 9.8 billion by 2050 (UN, 2017), therefore the wheat demand is expected to increase by 60% (FAO, 2012). In Mexico, wheat grain is the 7th most important crop in terms of production, with approximately 3.2 million tons harvested in 2019 (SIAP, 2019a). By federal states, the main producers are Sonora, Baja California, Sinaloa, Guanajuato and Michoacán, comprising 87 % of the total yield harvested in Mexico (SIAP, 2019b). Mexico's climatic diversity allow having 2 crop seasons depending on the sown period: autumn-winter and spring-summer (Escobar, 2014). Given the importance of wheat in Mexican economy, early crop yield forecasts are essential to assist decision makers and the wheat industry by planning the demand and ensuring an adequate supply.

Traditionally, crop yield models relied on climate variables such as solar radiation, precipitation or temperature, as well as in soil properties, to assess inter-annual variability in yield prediction (Stephens et al., 1989; Jiang and Thelen, 2004; Lobell and Field, 2007). In addition, specific crop information was generally added to assess crop status (nutrient deficiencies, diseases, biomass, etc.) throughout the season (Thenkabail, 2003). However, field data collection also constraints the

* Corresponding author.
E-mail address: diego@latuv.uva.es (D. Gómez).

https://doi.org/10.1016/j.agrformet.2020.108317
Received 1 October 2020; Received in revised form 22 December 2020; Accepted 26 December 2020
Available online 16 January 2021

extension and applicability of the models given its cost and time consumption. The use of satellite remote sensing has improved the accuracy of crop models that used only climate data, since they provide direct information about the crop growing status in a cost-effective manner (Jiang et al., 2003; Mahlein et al., 2012). In particular, remote sensing vegetation indices (VIs) such as Normalized Difference Vegetation Index (NDVI), Leaf Area Index (LAI) or Enhanced Vegetation Index (EVI) have been widely used to derive biophysical measurements with yield prediction purposes (Mkhabela et al., 2011; Dong et al., 2015; Skakun et al., 2017; Gómez et al., 2019). These indices exploit the vegetation spectral reflectance in the optical and near infra-red (NIR) parts of the spectrum. Nevertheless, some differences are observed across sensors and satellite platforms (Fleming, 2006; Franke et al., 2006; Roy et al., 2016). Nowadays, it is a common practice to combine climate, crop growth data and remote sensing VIs to undertake crop yield predictions at local, regional or global scale (Prasad, 2006; Rojas, 2007). Some international programs such as Monitoring Agricultural Resources (MARS -EC JRC), Global Information and Early Warning System (GIEWS - FAO), GEO Global Agricultural Monitoring Initiative (GEOGLAM) or China's global crop-monitoring system (CropWatch) intend to forecast crop yields at very large scale in an operative manner (Justice et al., 2015; Becker-Reshef, 2015). However, these methods need still more consistent calibration and validation protocols as well as improvements in terms of crop yield accuracy (Fritz et al., 2019). Many studies have shown the possibility to undertake wheat yield prediction in different locations at smaller scale (Alvarez, 2009; Pantazi et al., 2016; Cai et al., 2019; Ribeiro et al., 2019). However, these models require calibration with field-data measurements if they are intended to be used elsewhere (Thenkabail, 2003). Therefore, it is recommended to build site-specific models with representative data of the study area, and calibrate them with field data to address yield spatial and inter-annual variability for a specific location.

In recent years, machine learning (ML) methods have considerably advanced, providing cost-effective and accurate solutions to real-life problems. They can simultaneously find complex non-linear relationships in high-dimensional datasets (Crane-Driesch, 2018). Given the large volume of data provided by satellite remote sensing, ML approaches are becoming an indispensable tool in crop-yield modelling and precision agriculture (Chlingaryan et al., 2018). Previous literature used ML techniques as well as linear approaches to predict wheat yield using calibrated models at different scale. Romero et al. (2013) obtained rules for predicting durum wheat yield through different machine learning algorithms in Argentina at plant and pot level. Johnson et al. (2016) developed crop yield forecast models for the Canadian Prairies at larger scale (Census Agricultural Regions) with the highest skill score being 0.269. In United Kingdom, Pantazi et al. (2016) predicted wheat yield and classified field area into different yield potential zones in an experimental site of 22 ha (correct classification reached 91.3%). More recently, Cai et al. (2019) used and compared the performance of three ML algorithms (support vector machine, random forest, and neural network) and a linear regression model with LASSO penalty to predict wheat yield at large scale with an accuracy of R2~0.75 in Australia.

The aim of this work was to develop a methodology to predict wheat yield at large scale using ML methods in Mexico. We investigated the predictive capacity of four linear and four non-linear ML algorithms using five different feature selection scenarios. To build the models, satellite and climate data were combined with antecedent yield information, at municipal level, from 2004 to 2018. We evaluated the capacity of the YieldBaseLine predictor to reduce the inherent spatial and inter-annual variability of the data, as well as the possibility to implement the models with operative purposes.

2. Materials and methods

2.1. Study area

Our study area is Mexico. This country is located in the southern part of North America bordered by United States to the north, the Pacific Ocean to the west and south, Guatemala and Belize to the south-east and the Caribbean sea to the east (Fig. 1). It is the 13th largest (1.9 million km^2) and the 10th most populated (128.6 million inhabitants) country in the world. Administratively, it is divided in a federation of 32 states, which in turn are divided into municipalities (2458 in total).

The climate in Mexico is very diverse. According to Koppen classification (Kottek, 2006), the most representative climate types are hot semi-arid climate (BSh), hot desert climate (BWh), hot summer Mediterranean climate (Csa), tropical savanna climate (Aw) and cold semi-arid climate (BSk). The average of the accumulated annual precipitation goes from 44 to 4487 mm/year, and the average of the mean annual temperature from 2.0 to 28.7°C. Fig. 2 shows the wheat yield production by Mexican states in 2019.

2.2. Materials

2.2.1. In-situ data

Crop yield data was downloaded from Servicio de Información Agroalimentaria y Pesquera -SIAP- webpage (SIAP, 2019c). This dataset includes information such as wheat yield, type of crop cycle (Autum--Winter or Spring-Summer) and sown areas, at municipal level, from 2004 to 2018. Additionally, it provides information about the general conditions of the crop in terms of irrigation or rain-fed systems. Harvest date information was extracted at state level from SIAP (2019d). The Mexico land use map v.6.0 (1:25000) from the National Institute of Geography and Statistics –INEGI- (INEGI, 2017) was used to discern

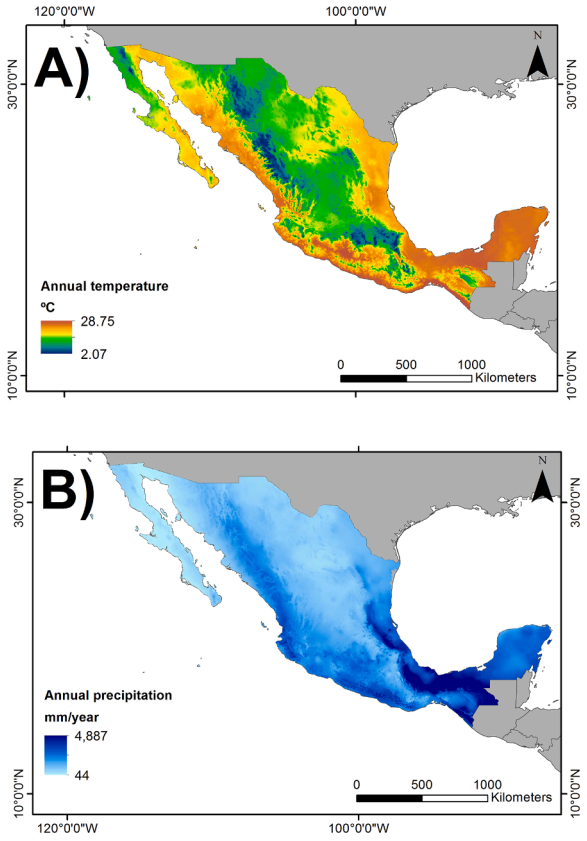


Fig. 1. A) Mean annual temperature (°C) and B) average of the accumulated annual precipitation (mm) (Fick and Hijmans, 2017).

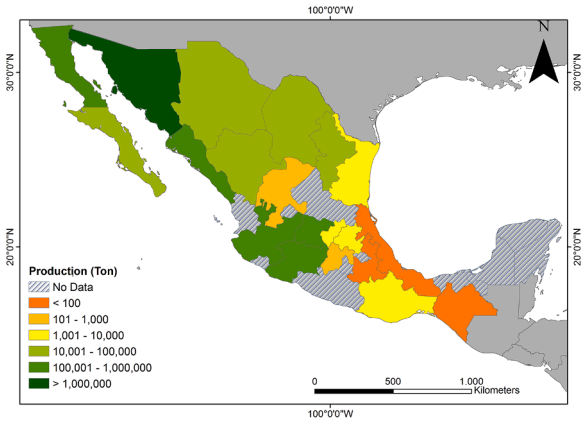


Fig. 2. Wheat yield production (Ton) by Mexican states in 2019. Source: http://infosiap.siap.gob.mx

cultivable/arable land and irrigation or rain-fed systems. We used the datasets provided by http://www.diva-gis.org/ webpage to delimit the states and municipal boundaries of Mexico.

2.2.2. Satellite and climate data

We used the NDVI data (Eq. (1)) provided by the MOD13Q1 v.6 product (LPDAAC-USGS, https://lpdaac.usgs.gov/products/mod13q1v006/) from the Moderate Resolution Imaging Spectroradiometer (MODIS). This sensor is onboard the Terra satellite and retrieves spectral data using 36 discrete bands (0.4 μm to 14.4 μm) with a temporal resolution of 1-2 days (NASA, 1999). The MOD13Q1 product has a spatial resolution of 250 m and a temporal resolution of 16 days. The temporal decorrelation between MODIS sensor and MOD13Q1 product is explained by the premise of selecting the best available pixel value using all the acquisitions, from the 16-day period, to reduce the cloud effects at pixel level. We downloaded the NDVI data (tiles: h09v06, h08v06, h08v07, h07v06, h09v07, h08v05) from the EarthData website (https://earthdata.nasa.gov/). Subsequently, we calculated a monthly composite image with the maximum value of NDVI. Thus, NDVI data can be coupled with the climate predictors by months of the year and at the same time, we reduce the cloud cover influence.

$$NDVI = (NIR - RED)/(NIR + RED) \quad (1)$$

where RED corresponds to the spectral reflectance in the red wavelength region and NIR (near infrared) to the reflection in the near-infrared spectrum.

We used the ERA5 re-analysis product developed by the European Centre for Medium-Range Weather Forecasts (ECMWF) and downloaded from the webpage: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means?tab=overview. It combines past observations with models to generate consistent time series of multiple climate variables (Copernicus Climate Change Service, 2019).

Table 1
Variables from the ERA5 (ECMWF) used in the study.

Variables	Abrev.	Unit
Evaporation	e	m of water equivalent
Leaf area index (high vegetation)	lai_hv	m^2/m^2
Leaf area index (low vegetation)	lai_lv	m^2/m^2
Potential evaporation	pev	m
Skin temperature	skt	° K
Surface net solar radiation	ssr	J/m^2
Temperature (2 m)	t2m	° K
Surface pressure	stl1	Pa
Total cloud cover	tcc	Times one
Total precipitation	tp	m
Volumetric soil water (layer 1)	swvl1	m^3/m^3

We selected 11 variables from the ERA5 monthly averaged data, on single levels, from 2004 to 2018 (Table 1). This dataset has a horizontal resolution of 0.25 degrees.

2.3. Methods

2.3.1. Data preparation

We used ENVI/IDL software (Exelis Visual Information Solutions, Boulder, Colorado) to extract the monthly mean values of NDVI and climate variables across all the municipalities of Mexico. As mentioned in section 2.2.1, we used an approximate harvest date based on SIAP (2019d) statistics at state level. In most states, harvest season spans from one to two months, independently of the cycle type (Autumn-Winter or Spring-Summer). Thus, we selected the two adjacent months whose percentage of harvest (in overall) were the highest, and take them as reference date. Subsequently, we extracted the NDVI and climate mean values (per month and municipality) for the antecedent 10 months, and also for those two months taken as reference of harvest. Note that models aimed to predict wheat yield before harvest months (at the end of ripening period) so that the 11^{th} and 12^{th} months were not included in the models. We inspected the NDVI curve presented in Fig. 3 (NDVI time series across all the municipalities and years using reference harvest months) to ensure that this approach can capture the wheat phenological phases in the study area (Magney et al., 2016; Boori et al., 2019). Based on the NDVI seasonal trends observed in Fig. 3, and given the phenological stages and seasonal trends presented by Magney et al. (2016), we comprised the NVDI and climate data as follows: germination and tillering (months: 1,2,3,4), stem extension (months: 5,6), heading (months: 7,8) and ripening (months: 9,10). Harvest months (11, 12) were not included in the models as variables since we aimed to predict wheat yield before the harvest reference date. Therefore, NDVI and climate data were averaged and aggregated accordingly for each of the established phases to be used as predictors. Note that these are temporal estimates of phenological phases, and some uncertainties are expected since the size of the states brings, inherently, some temporal variability within them. Table S1 summarizes the model predictors.

In addition, we used the information provided by SIAP (2019c) to add critical data to the models such as crop cycle (Autumn-Winter or Spring-Summer) and irrigation method (irrigation or rain-fed). The yield was taken as response variable (Ton/ha). Aiming to address any geographical bias of the yield, we created the YieldBaseLine variable. It comprises the average yield of the antecedent years in each municipality, with coincident irrigation methods and crop cycles. This is a modified version of the approach taken by (Salvador et al. (2020)), which proved to improve prediction accuracies and work better than models using the YieldBaseLine variable alone, or models not using this variable at all. Those samples/municipalities which had not antecedent yield data (e.g. first year of the time series) were deleted from the dataset during the filtering process until an antecedent yield value could be used.

We used R software (Core Team, 2017) to pre-process the data (center and scale) and build the models. The unfiltered dataset was composed by 51 predictors and 5823 samples. We filtered all those observations which had any missing value in any of the predictors (758) or had zero yield (137). The final dataset had 4928 samples and 51 predictors. Based on the crop cycle, there were 3283 samples for Autumn-Winter and 1645 samples for Spring-Summer. Regarding irrigation type, 2891 samples were irrigated and 2037 were under rain-fed conditions. The data distribution across years was the following: 2004 (317), 2005 (317), 2006 (325), 2007 (344), 2008 (337), 2009 (347), 2010 (347), 2011 (308), 2012 (314), 2013 (326), 2014 (335), 2015 (343), 2016 (333), 2017 (324) and 2018 (311).

2.3.2. Model building and prediction

We used the caret package (Kuhn, 2008) implemented in the R software for modelling purposes. As there is not a unique approach to

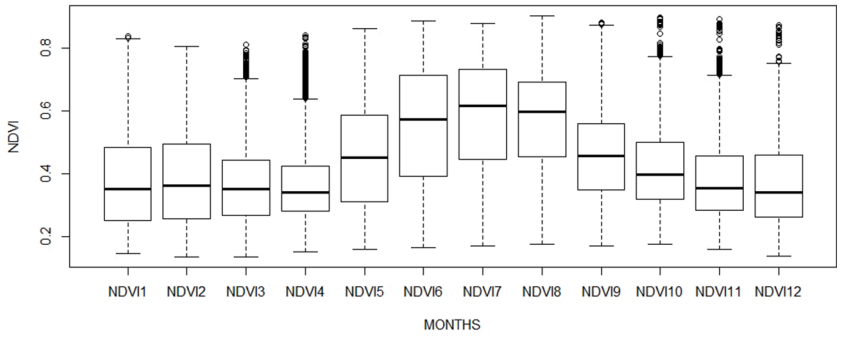


Fig. 3. Temporal evolution of the NDVI retrieved from the cultivated areas in the municipalities of Mexico over the study period (2004-2018). The x-axis represents the NDVI distribution over the growing cycle for each month before the harvest. NDVI11 and NDVI12 correspond to the months taken as harvest reference date.

solve machine learning problems, it is recommended to work iteratively with several statistical methods to obtain the best prediction performance (Agakov et al., 2006). In this context, we used four linear models (generalized linear model –glm- (Nelder and Wedderburn, 1972), ridge regression –ridge- (Zou and Hastie, 2012), lasso (Zou and Hastie, 2012), partial least squares –pls- (Wehrens and Mevik, 2007)) and four non-linear and more flexible models (k-nearest neighbours –kknn- (Schliep et al., 2016), support vector machine radial –svmR- (Scholkopf et al., 1997), extreme gradient boosting -xgbTree- (Chen et al., 2015) and random forest –rf- (Breiman, 2001)). There are several methods in the literature to identify the most important predictors in a dataset (Degenhardt et al., 2019). In this study, we used the findCorrelation function from caret package to apply different feature selection (FS) scenarios over the quantitative predictors. Four different scenarios were proposed to address multicollinearity in terms of Pearson's correlation coefficient: FS = 0.5, FS = 0.75, FS = 0.90 and No FS (no prior feature selection); and one additional scenario to compare the ability of a one-predictor model (YieldBaseLine) against the multivariate models. The quantitative input variables were centered and scaled using the pre-processing function of caret package. We randomly split the original dataset into training and validation set (80%) and hold-out set (20%). We used the bootstrapping resampling technique (25 repetitions) implemented in caret package to calibrate the models and find the best hyper-parameters. Optimized models were tested using the independent hold-out set (20%) not included during the training and validation phase. We used two metrics to evaluate model accuracies: the coefficient of determination (R^2) and root mean squared error (RMSE). The latter was converted to percent RMSE (%RMSE) by dividing the RMSE by the mean observed yield (3.56 ton/ha) across years and municipalities. The mean yield for irrigated wheat crops was 4.91 Ton/ha and for rain-fed crops was 2.08 Ton/ha. By crop-type seasonality, the mean yield was 2.20 Ton/ha for Spring-Summer, while 4.43 Ton/ha for Autumn-Winter.

2.3.4. Can our models be used in an operative manner? Simulation of a real-case scenario

We used the last-year of records in the dataset (2018) to assess the model's ability to estimate wheat yield in a real-case scenario. Thus, we simulate how the models would be used in reality in an operative manner. The models were trained and optimized using all data samples from 2004 to 2017 with a bootstrapping method (25 repetitions), and independently tested with 2018 data.

3. Results

3.1. Crop yield prediction with random splitting

We compared the model performance of four linear and four non-linear ML algorithms using different feature selection scenarios. Table S2 contains the selected predictors per scenario using the FindCorrelation function. Models were optimized by means of a bootstrapping method (25 repetitions) and tested using a randomly holdout dataset of 20% of the original data.

Fig. 4 and 5 show the prediction accuracy of each ML algorithm per scenario. Note that YieldBaseLine scenario -only one predictor- was just run for glm, kknn, svmRadial and rf; the other methods require more than one variable in caret. Most of the models presented RMSE values < 0.90 Ton/ha, which is less than 25% error compared to the mean yield (%RMSE). In general, the non-linear models (RMSE = 0.78 – 0.89 Ton/ha, %RMSE = 21.9% – 25.0%, R^2 = 0.79 – 0.84) performed better than the linear models (RMSE = 0.85 – 0.96 Ton/ha, %RMSE = 23.8% – 26.9%, R^2 = 0.75 – 0.81). By feature selection scenarios, the FS = 0.5 (9 predictors) selected the most informative variables to our models. Table S2 shows the variables included in each scenario. Models built with only the YieldBaseLine predictor achieved poorer performances for the tested ML algorithms, what demonstrates that our model configuration (adding NDVI and climate data) improves the simplest way of yield forecasting (the average of previous years). Table S3 shows the hyper-parameter configuration of the best models per feature selection scenario. The best performing algorithm was the rf and FS = 0.5 (RMSE = 0.78 Ton/ha, %RMSE = 21.9% R^2 = 0.84).

3.2. Simulation of a real-case scenario

To further assess the model's predictive capacity in a real-case scenario, we used the latest year of records (2018) to test the optimized models (Fig. 6). We used those predictors available before harvest (similar as in Section 3.1). The numerical results are provided in Table S4.

In general, linear and non-linear models performed similarly. The worst results were obtained by the pls (linear) and kknn (non-linear) models with RMSE > 0.90 Ton/ha (%RMSE > 25%). It is worth noting that most of the models under the YieldBaseLine scenario, performed worse than those including NDVI and climate data (with the exception of kknn).

The best model to predict wheat yield in 2018, before harvest

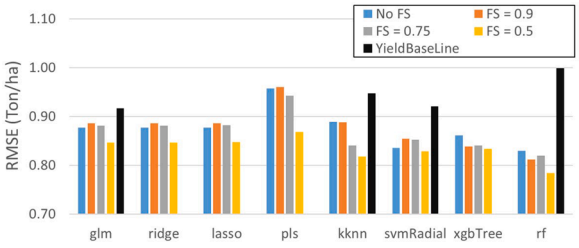


Fig. 4. RMSE values calculated from the hold-out dataset for linear (glm, ridge, lasso, pls) and non-linear (kknn, svmRadial, xgbTree and rf) regression models using five feature selection scenarios (No FS, FS = 0.90, FS = 0.75, FS = 0.5 and YieldBaseLine).

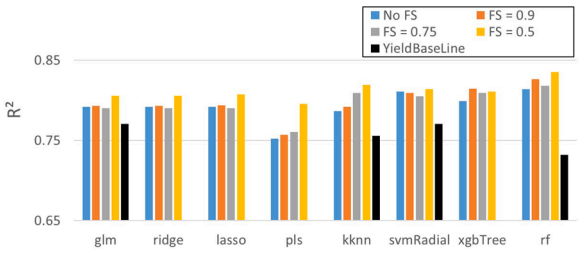


Fig. 5. R-squared values calculated from the hold-out dataset for linear (glm, ridge, lasso, pls) and non-linear (kknn, svmRadial, xgbTree and rf) regression models using five feature selection scenarios (No FS, FS = 0.90, FS = 0.75, FS = 0.5 and YieldBaseLine).

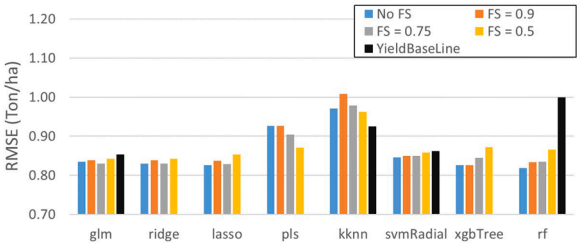


Fig. 6. Predictive performance calculated for the year 2018, at the end of ripening period, using five feature selection scenarios (No FS, FS = 0.90, FS = 0.75, FS = 0.5 and YieldBaseLine in terms of RMSE (Ton/ha).

months, was the rf under No FS scenario (RMSE = 0.82 Ton/ha, %RMSE = 23.0% R^2 = 0.81), and its best hyper-parameter was mtry = 26. The other models (glm, ridge, lasso and xgbTree) showed slightly higher predictive capacity under the No FS scenario. Table S5 shows the ten most important variables included in the best performing model (rf, no FS).

Fig. 7 jointly represents actual versus predicted yield (rf, No FS) in 2018. To better assess its model capabilities, we added a 1:1 line with comparison purposes. Additionally, each sample in the test set was labelled as rain-fed or irrigated (Fig. 7a) and its cycle type (Fig. 7b). The rf model showed great capability to predict wheat yield when actual yields were < 3.5 Ton/ or > 4.5 Ton/ha, showing higher uncertainties between those two values. Model performances were substantially better for those irrigated fields with actual yields close to or higher than the average yield for this type (4.91 Ton/ha). There was a consistent and good model performance for most of the rain-fed samples. Despite differences between the average of Spring-Summer yield (2.20 Ton/ha) and Autumn-Winter (4.43 Ton/ha) (Section 2.3.2.), Fig. 7b indicates no clear bias related to this variable. Note that higher uncertainties were observed in irrigated fields with unusually low yields, what can be

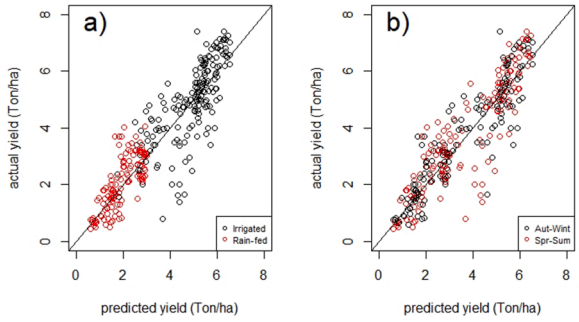


Fig. 7. Comparison between predicted and actual yields using rf model, scenario No FS, before harvest, in 2018. Sample points are labelled as: a) irrigation or rain-fed type and b) crop cycle, which can be Autumn-Winter (Aut-Wint) or Spring-Summer (Spr-Sum).

ascribed to agricultural malpractices or unusual/extreme events (e.g. pests) that may have affected certain wheat fields in some municipalities. Fig. 8 shows the spatial variability of predicted yield at municipal level using the rf model and No FS scenario.

4. Discussion

Crop yield forecasting can provide meaningful information to decision makers, authorities and producers for rapid decision-making during the growing season. Furthermore, crop yield estimates are also useful in relation to trade, development policies and humanitarian assistance linked to food security (JRC, 2018). In this context, we evaluated the predictive capacity of four linear and four non-linear ML algorithms under five different feature selection scenarios at municipal level in Mexico.

4.1. Crop yield prediction with random splitting

Our results show that the combination of NDVI and climate data can improve simpler (but still useful) models which only use previous year's yield with estimation purposes. These findings concur with previous works in the literature ((Salvador et al., 2020)). Across the proposed five feature selection scenarios, linear models obtained better results when less predictors were included (e.g. FS = 0.5), while some of the non-linear models (e.g. svmRadial or xgbTree) were less sensitive to feature reduction (Fig. 4 and 5). Many studies have addressed the importance of multi-collinearity among predictors and the different impact that they exert on linear and non-linear models (Bonate, 1999; Erkoç et al., 2010; Lavery et al., 2019). In overall, there were not large differences between linear and non-linear models in terms of model performance. Despite linear regularization approaches such as ridge and lasso intend to improve multiple linear regression methods (fitted with ordinary least squares) by reducing their model complexity (James et al., 2013), they did not outperform the glm models exerted in this study. Given that data dimensionality (1 to 51 predictors) is much lower than sample number (4928), multiple linear regressions can perform equally well (James et al., 2013). Pls models proved to be least accurate. In previous works, pls methods were effective when analyzing a large array of relatively correlated predictors and/or the sample size was not large enough when compared with the number of independent variables (Carrascal et al., 2009). Among the proposed non-linear models, kknn obtained poor performances; while svmRadial, xgbTree and rf obtained the lowest RMSE scores (0.78 – 0.85 Ton/ha). The highest accuracy was achieved by the rf models with RMSE= 0.78 - 0.83 Ton/ha, %RMSE = 21.9% – 23.3% and R^2 = 0.81 – 0.84.

Given the vast extension covered by most of Mexican states, there is a

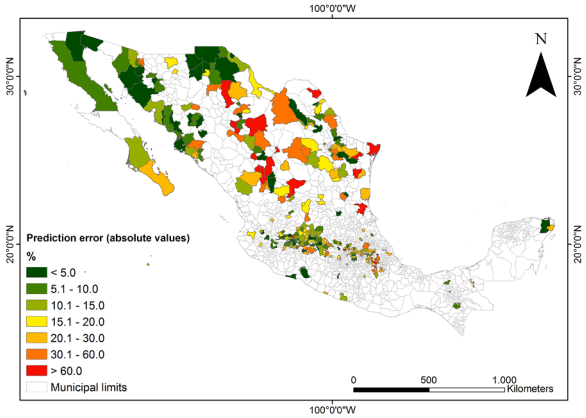


Fig. 8. Spatial variability of predicted wheat yield at municipal level using the rf model, No FS scenario, before harvest, in 2018. The units are the percentage of error expressed in absolute values using the actual yield as reference.

regional and climatic variability within states (Feng et al., 2010) that justifies the use of the two-month period harvest phase (in absence of more accurate harvest-date information). This time frame would meet the recommendations made by Hoefsloot et al., 2012 for optimal crop yield forecasting. Wheat yield predictions were undertaken, approximately, at the end of the ripening period; although some uncertainties introduced by the harvest date estimation are expected. Across scenarios, the best results were obtained under the FS = 0.5 scenario. It was proven that one-variable scenario (YieldBaseLine predictor) had lower predictive capacity when used alone. The best performing algorithm was rf under No FS scenario (RMSE = 0.78 Ton/ha, %RMSE = 21.9% R^2 = 0.84). This model has therefore shown a high potential to predict wheat yield before harvest in Mexico, as it has already demonstrated its capabilities in other regions around the world such as Canada, Australia or China (Cai et al., 2019; Cao, 2020; Mkhabela et al., 2011).

4.2. Simulation of a real-case scenario

Could these models be applied in an operative manner in Mexico? The results obtained in Section 3.2 show that the proposed models, and specially the rf model (No FS scenario), can predict with an %RMSE~ 23% before harvest months. These results are in accordance to those obtained in Section 3.1 when models were validated randomly across years. The YieldBaseLine alone proved to be a good estimator of crop yield. However, it does not comprise crop growth information or environmental conditions, so model results under this scenario were worse than those which used NDVI and climate data as well. Several studies confirm the capacity of satellite and climate data to add critical information in crop yield models and therefore, better explain the inter-annual variability of the yield (Mkhabela et al., 2011; Cai et al., 2019; Cao et al., 2020).

4.3. Comparison with similar studies and limitations

The results discussed in Sections 4.1 and 4.2 are in line with other efforts made to accurately forecast wheat yield throughout the world. Nevertheless, model comparisons need to be addressed with care since sampling design, study extension and temporal coverage play a fundamental role in explaining the spatial and temporal variability of the predictions (Leng and Hall, 2020) and therefore, the capacity of the models to predict well on unseen data with operative purposes (Thenkabail, 2003; Ye et al., 2007). Pantazzi et al. (2016) modelled spatial variability of wheat yield, at field scale, using satellite and soil data (R^2 = 0.70 – 0.91). In a similar vein, Hunt et al. (2019) obtained promising results to predict wheat yield with high spatial resolution data - Sentinel 2, 10m- (RMSE = 0.61 Ton/ha and R^2 = 0.91). However, those approaches are constrained by the short temporal coverage, extension of the study site and the knowledge of the exact location of the target fields. The latter is usually a severe limitation for field-based studies because agricultural data is often limited, if it exists at all (Sommer and Paxson, 2010; Grassini et al., 2015). Larger scale studies have also shown the benefits to use, in combination or not, satellite and climate data with wheat yield prediction purposes. Model results achieved by Mkhabela et al. (2011) using agro-climatic zones in the Canadian prairies ranged from R^2 = 0.47 – 0.80. Cai et al. (2019) developed a model to predict wheat yield at county-level across Australia, using 14 years' data (R^2 = 0.75). A similar approach has been recently developed by Cao et al., 2020 to predict wheat yield at county-level across China, using 15 years of samples (R^2 = 0.68-0.75). In this sense, our work addresses the inter-annual and spatial variability that affects some of the aforementioned studies (Pantazzi et al., 2016; Hunt et al., 2019), as well as concurs with larger-scale studies with similar if not higher accuracies (Mkhabela et al., 2011; Cai et al., 2019; Cao et al., 2020).

One of the major sources of uncertainty in large-scale studies (such as ours) is the lack of precise field geolocations (Hunt et al., 2019). This fact hinders the use of spectral Vis, as the reflectance of satellite imagery can be mixed with other crop types. This uncertainty can be diminished using land use/land cover maps, as well as other relevant spatial data whenever available (e.g. irrigated vs rain-fed areas). The more cultivated and extended the target crop is over an area, the more precise is expected to be the reflectance value acquired by the satellite sensor for the study. Climate data may also bring some sort of uncertainty to the models due to imprecision in data collection, combination of different data sources, or errors associated to produce continuous data spatially and temporally (Tebaldi and Lobell, 2008; Tao et al., 2009; Asseng et al., 2011; Boilley and Wald, 2015). The uncertainties associated to ERA5 variables decrease with time (e.g. lowest uncertainty in recent years), and this improvement can mainly be ascribed to the quantity and quality of the available observations (ECMWF, 2020).

In this study, we demonstrated the capacity of our models to predict wheat yield in Mexico before harvest time. We combined satellite data, climate data and the antecedent yield information at municipal level (YieldBaseLine) to address any temporal or geographical bias in the study area (e.g. soil characteristics, technical innovations in certain municipalities, etc.). Given the high predictive capacity shown by the YieldBaseLine predictor alone, it can be used in similar studies with other crop types to reduce any spatial or temporal bias in the data. Furthermore, we simulated a real-case scenario with 2018 data to prove the operative capacity of our models in Mexico. The methodology presented in this work could be applied to other regions or crops.

5. Conclusions

In this study, we proposed a new method to predict wheat yield that adjusts one of the simplest predictive models (YieldBaseLine) with satellite (NDVI) and climate data to account for the inter-annual and spatial variability of wheat yield in Mexico. We compared four linear (glm, ridge, lasso and pls) and four non-linear (kknn, svmRadial, xgbTree and rf) ML algorithms and five different feature selection scenarios. In general, the best scenario for the non-linear models was FS = 0.5, while the non-linear models were less sensitive to any feature reduction. Our results evidence the high capacity of the rf model (FS = 0.5) to predict wheat yield before harvest time (R^2 = 0.84). To prove model's operability, we also used the latest records in the dataset (year 2018) in a simulated real-case scenario. The best performing model was again the rf with R^2 = 0.81(No FS).

This study proposes a methodology to model crop yield at large scale, and it may be used with operative purposes. Therefore, it can be of interest to decision and law makers, producers, authorities or the wheat industry. In addition, it can help to establish appropriate food security and trading policies. Similar approaches could be applied to other regions and crops.

Author contributions

Conceptualization, D.G.; Methodology, D.G., P.S.; Formal Analysis, D.G.; Validation, P.S. and D.G.; Data Curation, D.G., J.S. and J.L.C., Writing—Original Draft Preparation, D.G., Writing—Review & Editing, D.G., P.S., J.L.C.; Supervision, J.S. and J.L.C.; Project Administration, J. S.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

All authors declare that they have no conflict of interest.

Acknowledgments

We would like to acknowledge the greatly contribution that the open-access policies of the following institutions have made to this work: LP DAAC (Land Processes Distributed Active Archive Center, NASA), European Centre for Medium-Range Weather Forecasts (ECMWF, European Comision) and Servicio de Información Agro-alimentaria y Pesquera (SIAP, Mexican government).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.agrformet.2020.108317.

References

Agakov, F., Bonilla, E., Cavazos, J., Franke, B., Fursin, G., O'Boyle, M.F., Williams, C.K, 2006. Using machine learning to focus iterative optimization. In: Proceedings of the International Symposium on Code Generation and Optimization, New York, NY, USA, 26–29 March 2006. IEEE Computer Society, Washington, DC, USA, pp. 295–305.

Alvarez, R., 2009. Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach. Eur. J. Agron. 30, 70–77. https://doi.org/10.1016/j.eja.2008.07.005.

Asseng, S., Foster, I.A.N., Turner, N.C., 2011. The impact of temperature variability on wheat yields. Glob. Change Biol. 17 (2), 997–1012.

Awika, J.M., 2011. Major cereal grains production and use around the world. Advances in Cereal Science: Implications to Food Processing and Health Promotion. American Chemical Society, pp. 1–13.

Becker-Reshef, I., 2015. GEOGLAM (GEO global agricultural monitoring) crop assessment tool. Ag data commons. 10.15482/USDA.ADC/1234202. Accessed 2020-05-23.

Boilley, A., Wald, L., 2015. Comparison between meteorological re-analyses from ERA-interim and MERRA and measurements of daily solar irradiation at surface. Renew. Energy 75, 135–143.

Bonate, P.L., 1999. The effect of collinearity on parameter estimates in nonlinear mixed effect models. Pharm. Res. 16, 709–717. https://doi.org/10.1023/A:1018828709196.

Boori, M.S., Choudhary, K., Paringer, R., Sharma, A.K., Kupriyanov, A., Corgne, S., 2019. Monitoring crop phenology using NDVI time series from sentinel 2 satellite data. In: 2019 5th International Conference on Frontiers of Signal Processing (ICFSP). IEEE, pp. 62–66.

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

Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., ..., Peng, B., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric. Forest Meteorol. 274, 144–159.

Cao, et al., 2020. Identifying the Contributions of Multi-Source Data for Winter Wheat Yield Prediction in China. Remote Sensing 12 (5), 750. https://doi.org/10.3390/rs12050750.

Carrascal, L.M., Galván, I., Gordo, O., 2009. Partial least squares regression as an alternative to current regression methods used in ecology. Oikos 118 (5), 681–690.

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y., 2015. Xgboost: extreme gradient boosting. R package version 0.4-2, 1-4.

Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. Comput. Electron. Agric. 151, 61–69.

Copernicus Climate Change Service, 2019. https://climate.copernicus.eu/climate-reanalysis (Accessed at 25-05-2020).

Degenhardt, F., Seifert, S., Szymczak, S., 2019. Evaluation of variable selection methods for random forests and omics data sets. Brief. Bioinf. 20 (2), 492–503.

Dong, J., Xiao, X., Wagle, P., Zhang, G., Zhou, Y., Jin, C., ..., Yan, H., 2015. Comparison of four EVI-based models for estimating gross primary production of maize and soybean croplands and tallgrass prairie under severe drought. Remote Sens. Environ. 162, 154–168.

ECMWF, 2020. https://confluence.ecmwf.int/display/CKB/ERA5%3A+uncertainty+estimation (Accessed at 03-06-2020).

Escobar, R., 2014. El cultivo de secano. Revista de Geografía Agrícola 52-53, 61–113. https://www.redalyc.org/pdf/757/75749284005.pdf (Accessed at 20-05-2020).

Erkoç, A., Tez, M., Akay, K.U., 2010. On multicollinearity in nonlinear regression models. Selçuk J. Appl. Math. (Special Issue: 65–72).

FAO, 2012. http://www.fao.org/docs/eims/upload/306175/Briefing%20Paper%20(3)-Wheat%20Initative%20-%20H%C3%A9l%C3%A8ne%20Lucas.pdf (Accessed at 04-06-2020).

FAO, 2016. http://www.fao.org/3/y3557e/y3557e08.htm (Accessed at 20-05-2020).

Feng, S., Krueger, A.B., Oppenheimer, M., 2010. Linkages among climate change, crop yields and Mexico–US cross-border migration. Proceed. Natl. Acad. Sci. 107 (32), 14257–14262.

Fleming, D.J., 2006. Effect of relative spectral response on multi-spectral measurements and NDVI from different remote sensing systems.

Franke, J., Heinzel, V., Menz, G., 2006. Assessment of ndvi-differences caused by sensor specific relative spectral response functions. In: 2006 IEEE International Symposium on Geoscience and Remote Sensing. IEEE, pp. 1138–1141.

Fritz, S., See, L., Bayas, J.C.L., Waldner, F., Jacques, D., Becker-Reshef, I., ..., Rembold, F., 2019. A comparison of global agricultural monitoring systems and current gaps. Agric. Syst. 168, 258–272.

Gómez, D., Salvador, P., Sanz, J., Casanova, J.L., 2019. Potato yield prediction using machine learning techniques and sentinel 2 data. Remote Sens. 11 (15), 1745.

Grassini, P., Lenny, G.J., van Bussel, J.W., Joost, W., Lieven, C., Haishun, Y., Hendrik, B., Groot, H., Ittersum, M., Cassman, K.G., 2015. How good is good enough? Data requirements for reliable crop yield simulations and yield-gap analysis. Field Crop. Res. 177, 49–63.

Hoefsloot, P., Ines, A.V., Dam, J.C.V., Duveiller, G., Kayitakire, F., Hansen, J., 2012. Combining crop models and remote sensing for yield prediction: concepts, applications and challenges for heterogeneous smallholder environments. In: Proceedings of the Report of CCFAS-JRC Workshop at Joint Research Centre, Ispra, Italy, 13–14 June. Joint Research Center Technical Report. Publications Office of the European Union, Luxembourg, 2012. Available online: http://publications.jrc.ec.europa.eu/repository/bitstream/JRC77375/lbna25643enn.pdf (accessed on 29 May 2019).

Hunt, M.L., Blackburn, G.A., Carrasco, L., Redhead, J.W., Rowland, C.S., 2019. High resolution wheat yield mapping using sentinel-2. Remote Sens. Environ. 233, 111410.

INEGI, 2017. Available online: https://www.inegi.org.mx/temas/usosuelo/default.html#Herramientas. (Accessed at 05-06-2019).

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, 112. springer, New York, pp. 3–7.

Joint Research Centre – European Comission, JRC, 2018. https://ec.europa.eu/jrc/en/research-topic/crop-yield-forecasting (Accessed at 01-06-2020).

Jiang, D., Wang, N.B., Yang, X.H., Wang, J.H., 2003. Study on the interaction between NDVI profile and the growing status of crops. Chin. Geograph. Sci. 13 (1), 62–65.

Jiang, P., Thelen, K.D., 2004. Effect of soil and topographic properties on crop yield in a north-central corn–soybean cropping system. Agron. J. 96 (1), 252–258.

Johnson, M.D., Hsieh, W.W., Cannon, A.J., Davidson, A., Bédard, F., 2016. Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. Agric. Forest Meteorol. 218, 74–84.

Justice, C; Becker-Reshef, I; McGaughey, K; Hansen, M; Whitcraft, A; Barker, B.; Humber, M.; Deshayes, M., 2015. "Enhancing agricultural Monitoring with EO-based information" http://www.apogeospatial.com/issues/AO_wi2015.pdf (Accessed at 04-06-2020).

Kuhn, M., 2008. Caret package. J. Stat. Softw. 28, 1–26. Available online: http://www.math.chalmers.se/Stat/Grundutb/GU/MSA.

Kottek, M, 2006. World map of the Köppen-Geiger climate classification updated. Meteorol. Z 15, 259–263.

Lavery, M.R., Acharya, P., Sivo, S.A., Xu, L., 2019. Number of predictors and multicollinearity: what are their effects on error and bias in regression? Commun. Stat.-Simul. Comput. 48 (1), 27–38.

Leng, G., Hall, J.W., 2020. Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. Environ. Res. Lett. 15 (4), 044027.

Lobell, D.B., Field, C.B., 2007. Global scale climate–crop yield relationships and the impacts of recent warming. Environ. Res. Lett. 2 (1), 014002.

Magney, T.S., Eitel, J.U., Huggins, D.R., Vierling, L.A., 2016. Proximal NDVI derived phenology improves in-season predictions of wheat quantity and quality. Agric. Forest Meteorol. 217, 46–60.

Mahlein, A.-K., Oerke, E.-C., Steiner, U., Dehne, H.W., 2012. Recent advances in sensing plant diseases for precision crop protection. Eur. J. Plant Pathol. 133, 197–209. https://doi.org/10.1007/s10658-011-9878-z.

Mkhabela, M.S., Bullock, P., Raj, S., Wang, S., Yang, Y., 2011. Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. Agric. Forest Meteorol. 151 (3), 385–393.

Nelder, J.A., Wedderburn, R.W., 1972. Generalized linear models. J. R. Stat. Soc. Ser. A (Gen.) 135, 370–384.

Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R.L., Mouazen, A.M., 2016. Wheat yield prediction using machine learning and advanced sensing techniques. Comput. Electron. Agric. 121, 57–65.

Core Team, R, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available online. https://www.R-project.org/ (accessed on 4 May 2019).

Prasad, et al., 2006. Crop yield estimation model for Iowa using remote sensing and surface parameters. International Journal of Applied Earth Observation and Geoinformation 8 (1), 26–33. https://doi.org/10.1016/j.jag.2005.06.002.

Ribeiro, A.F., Russo, A., Gouveia, C.M., Páscoa, P., 2019. Modelling drought-related yield losses in Iberia using remote sensing and multiscalar indices. Theo. Appl. Climatol. 136 (1-2), 203–220.

Rojas, O., 2007. Operational maize yield model development and validation based on remote sensing and agro-meteorological data in Kenya. Int. J. Remote Sens. 28 (17), 3775–3793.

Romero, J.R., Roncallo, P.F., Akkiraju, P.C., Ponzoni, I., Echenique, V.C., Carballido, J.A., 2013. Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires. Comput. Electron. Agric. 96, 173–179.

Roy, D.P., Kovalskyy, V., Zhang, H.K., Vermote, E.F., Yan, L., Kumar, S.S., Egorov, A., 2016. Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. Remote Sens. Environ. 185, 57–70.

Schliep, K., Hechenbichler, K., & Schliep, M. K., 2016. Package 'kknn'. http://ftp5.gwdg.de/pub/misc/cran/web/packages/kknn/kknn.pdf (Accessed at 28-05-2020).
Salvador, P., Gómez, D., Sanz, J., Casanova, José Luis, 2020. Estimation of Potato Yield Using Satellite Data at a Municipal Level: A Machine Learning Approach. ISPRS International Journal of Geo-Information 9 (6), 343. https://doi.org/10.3390/ijgi9060343.
Scholkopf, B., Sung, K.K., Burges, C.J., Girosi, F., Niyogi, P., Poggio, T., Vapnik, V., 1997. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. IEEE Trans. Signal Process. 45 (11), 2758–2765.
SIAP, 2019a. http://infosiap.siap.gob.mx:8080/agricola_siap_gobmx/AvanceNacionalSinPrograma.do (Accessed at 20-05-2020).
SIAP, 2019b. http://infosiap.siap.gob.mx:8080/agricola_siap_gobmx/ResumenProducto.do (Accessed at 20-05-2020).
SIAP, 2019c. http://infosiap.siap.gob.mx/gobmx/datosAbiertos.php (Accessed at 20-08-2019).
SIAP, 2019d. http://infosiap.siap.gob.mx/estacionalidad_gb/est_agricola-AA/index.php (Accessed at 20-08-2019).
Skakun, S., Franch, B., Vermote, E., Roger, J.C., Becker-Reshef, I., Justice, C., Kussul, N., 2017. Early season large-area winter crop mapping using MODIS NDVI data, growing degree days information and a Gaussian mixture model. Remote Sens. Environ. 195, 244–258.
Sommer, R., Paxson, V., 2010. Outside the closed world: On using machine learning for network intrusion detection. In: Proceedings of the 2010 IEEE Symposium on Security and Privacy, Berkeley/Oakland, CA, USA, 16–19, pp. 305–316.
Stephens, D.J., Lyons, T.J., Lamond, M.H., 1989. A simple model to forecast wheat yield in Western Australia. J. R. Soc. West. Aust. 71, 77–81.
Tao, F., Zhang, Z., Liu, J., Yokozawa, M., 2009. Modelling the impacts of weather and climate variability on crop productivity over a large area: a new super-ensemble-based probabilistic projection. Agric. For. Meteorol. 149 (8), 1266–1278.
Thenkabail, P.S., 2003. Biophysical and yield information for precision farming from near-real-time and historical Landsat TM images. Int. J. Remote Sens. 24 (14), 2879–2904.
United Nations - UN, 2017. https://www.un.org/development/desa/en/news/population/world-population-prospects-2017.html (Accessed at 04-06-2020).
Wehrens, R., & Mevik, B.H., 2007. The pls package: principal component and partial least squares regression in R.
Ye, X., Sakai, K., Manago, M., Asada, S.I., Sasao, A., 2007. Prediction of citrus yield from airborne hyperspectral imagery. Precis. Agric. 8 (3), 111–125.
Zou, H., & Hastie, T., 2012. Elastic-net for sparse estimation and sparse PCA. URL http://www.stat.umn.edu/~hzou. (Accessed at 28-05-2020).