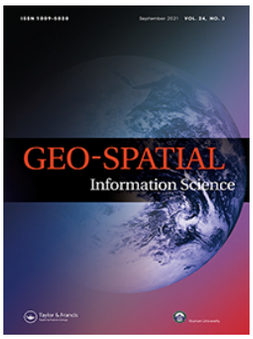


Geo-spatial Information Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgsi20

Integration of maximum crop response with machine learning regression model to timely estimate crop yield

Qiming Zhou & Ali Ismaeel

To cite this article: Qiming Zhou & Ali Ismaeel (2021) Integration of maximum crop response with machine learning regression model to timely estimate crop yield, Geo-spatial Information Science, 24:3, 474-483, DOI: 10.1080/10095020.2021.1957723

To link to this article: https://doi.org/10.1080/10095020.2021.1957723

Published online: 06 Aug 2021.

GEO-SPATIAL INFORMATION SCIENCE
2021, VOL. 24, NO. 3, 474–483
https://doi.org/10.1080/10095020.2021.1957723

Taylor & Francis
Taylor & Francis Group

Integration of maximum crop response with machine learning regression model to timely estimate crop yield

Qiming Zhou and Ali Ismaeel

Department of Geography, Hong Kong Baptist University, Hong Kong, China

ABSTRACT

Timely and reliable estimation of regional crop yield is a vital component of food security assessment, especially in developing regions. The traditional crop forecasting methods need ample time and labor to collect and process field data to release official yield reports. Satellite remote sensing data is considered a cost-effective and accurate way of predicting crop yield at pixel-level. In this study, maximum Enhanced Vegetation Index (EVI) during the crop-growing season was integrated with Machine Learning Regression (MLR) models to estimate wheat and rice yields in Pakistan's Punjab province. Five MLR models were compared using a fivefold cross-validation method for their predictive accuracy. The study results revealed that the regression model based on the Gaussian process outperformed over other models. The best performing model attained coefficient of determination (R^2), Root Mean Square Error (RMSE, t/ha), and Mean Absolute Error (MAE, t/ha) of 0.75, 0.281, and 0.236 for wheat; 0.68, 0.112, and 0.091 for rice, respectively. The proposed method made it feasible to predict wheat and rice 6–8 weeks before the harvest. The early prediction of crop yield and its spatial distribution in the region can help formulate efficient agricultural policies for sustainable social, environmental, and economic progress.

ARTICLE HISTORY
Received 14 December 2020
Accepted 15 July 2021

KEYWORDS
Machine learning; remote sensing; crop yield; timely forecast

1. Introduction

The current world population of 7.6 billion is expected to increase by up to 9.8 billion by 2050, with most population growth in developing countries of Asia and Africa (United Nations 2019). Future dietary requirements of these developing nations will require a regular increase in agriculture production by maintaining a stable agroecosystem. A changing climate is also a profound threat to the food security of developing regions. Agriculture policies play a vital role in achieving productivity growth and raising an agrarian society's overall economic status (Simoncini et al. 2019). An effective agriculture policy relies on timely and accurate crop yield information to better manage supply and demand to ensure food security in the region (Maya Gopal and Bhargavi 2019). Additionally, a thorough picture of crop yield status helps control market swings that can be extremely disruptive in regions with an agriculture-based economy (Giannakis and Bruggeman 2015).

Crop yield monitoring in developing countries is mainly based on two types of sampling surveys. The early crop yield prediction is based on subjective surveys, like taking opinions from growers and the field officers' visual judgment. The later crop yield estimation is done using objective surveys such as a whole plot harvest or crop cut measurement from sample fields (Craig and Atkinson 2013). The sampling sites for a region are selected through a systematic random sampling scheme, and the small data sets often do not reflect the complete status of the seasonal croplands as samples are only collected from accessible fields. This labor-intensive crop monitoring system can be an adequate provision for overall agriculture management, but the crop yield estimation through this mechanism is time-consuming with large associated uncertainties. Moreover, final estimates of crop yields at a national scale are finalized after months of a crop harvest, making it challenging to take timely decisions of import and export to ensure food security and economic growth.

Satellite remote sensing is a cost-effective tool for real-time monitoring and crop status assessment (Fritz et al. 2019). Earth-orbiting satellites capture essential information about vegetation conditions over large areas with frequent revisits. Medium to coarse spatial resolution data from different satellites (e.g. Sentinel, SPOT, Landsat, and MODIS) are freely available for analyzing vegetation conditions at the regional and global scales (Jianxi Huang et al. 2019). Many recent studies have used remote sensing derived biophysical parameters in Crop Growth Models (CGM) to estimate crop yields. These biophysical parameters include leaf area index, soil moisture, the fraction of absorbed photosynthetically active radiation, evapotranspiration, and above-ground biomass. The use of these data in CGM has helped the researchers to evaluate different crop management

CONTACT Qiming Zhou qiming@hkbu.edu.hk

strategies for improving the crop yield at the regional and global scales (Jin et al. 2018). However, the need for local calibration makes it difficult to apply the CGM in developing countries where data scarcity often presents a prohibitive obstacle. The lack of spatial data to capture the heterogeneity of land surface also makes the CGM application misleading at the local scale.

Statistical modeling seems to be a more reasonable approach for data scarce regions. A regression model can be built between reported crop yield and remote sensing derived Vegetation Index (VI). Although regression models are considered region-specific and time-dependent, they can effectively fulfill spatiotemporal yield gaps by mapping regional crop yield at a fine scale within the study period (Lobell 2013). Previous studies like Jingfeng Huang et al. (2013) used Normalized Difference Vegetation Index (NDVI) derived from NOAA AVHRR to develop a stepwise regression model to estimate crop yield in major rice grown provinces of China. Petersen (2018) developed a multivariate regression model using indices derived from MODIS to predict the real-time yield of corn, soybean, and sorghum for the continent of Africa. The proposed methods in these studies do not need to map crop cover areas and utilize the monthly anomalies of indices to measures relative vegetation health.

More recently, Liu et al. (2019) compared the results of yield estimation on aggregated croplands and masks of specific crops in Canada. A better correlation was found between MODIS two-band Enhanced Vegetation Index (EVI2) at the peak growth stage of the crop and its national yield using crop-specific masks. A multi-linear regression model was developed to estimate the crop yields of winter wheat, corn, and soybean. Machine Learning(ML) methods in solving complex non-linear problems have also been utilized in crop yield prediction using remote sensing data. Johnson et al. (2016) employed these methods to predict yield canola, barley, and spring wheat using NDVI and EVI data. Two non-linear models (Bayesian Neural Networks and model-based recursive partitioning) were used to estimate crop yield in Canadian Prairies. In addition, recent studies are utilizing Machine Learning Regression (MLR) approaches with remote sensing data to tackle multiple cropland issues in a geo-spatial modeling environment (Prins and Van Niekerk 2020; Ujoh, Igbawua, and Ogidi Paul 2019; Mfuka, Byamukama, and Zhang 2020).

The advances in machine learning algorithms and increasing availability of multitemporal remote sensing data have largely helped to deal with data non-linearity and multi-dimensionality, which often create difficulties in applying regression models. Taking the advantages, this research aims (1) to develop a method using remote sensing data and machine learning regression models to estimate crop yield of wheat and rice in a large crop-production region; (2) to compare the performances of different regression models to identify the most suitable one for the yield prediction of each crop; and (3) to analyze the spatial variation of crop yield across the study region and to evaluate the proposed prediction model for pre-harvest yield forecasting. This study will help managers formulate efficient agricultural policies to improve crop production in low-performing regions to benefit society.

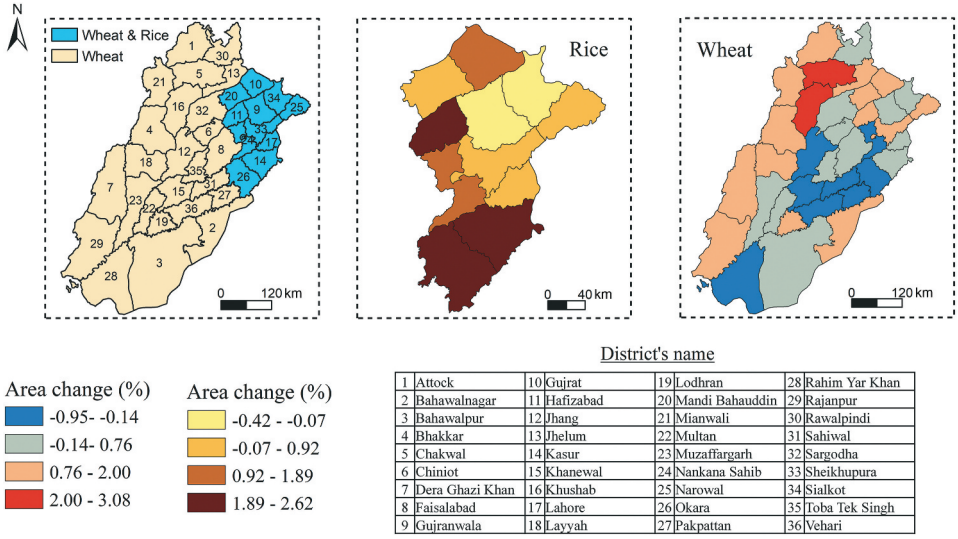
2. Material and methods

2.1. Study area

The study area for this research is the Punjab province of Pakistan, which has the largest share (73%) of the country's total cropped area (Figure 1). Wheat and rice, along with cotton and sugarcane, are the cash crops of this region. Punjab contributes 76% of wheat and 56% of rice in total national production (Dempewolf et al. 2014; Rehman et al. 2017). Wheat is the main crop of the Rabi season (November to April of the next year). Rice is the main crop in the Kharif season (May to October). The period of study was from November 2002 to October 2018. All districts that had wheat and rice cultivated area greater than 50 kHa were used to estimate the crop yield. The change (%) of wheat and rice cultivation area based on reported statistics in selected districts since 2003 is shown in Figure 1. No more than a 3% increase and a 1% decrease were observed for each selected district. A total of 16 years of historic EVI and reported crop statistics of both crops were used for this research.

2.2. Reported crop statistics

Under the directorate of agriculture, the Crop Reporting Service (CRS) is responsible for producing district level crop statistics for the Government of Punjab. The CRS Punjab has 1038 crop reporters that collect field data from 1240 sample villages. These sample villages account for 5% of Punjab's total villages and are selected for 5 years through stratified random sampling. The CRS generates three survey reports for each major crop. The first survey is a visual inspection of field reporters in sample villages to estimate the crop cultivated area after completing each crop's sowing period. The second inspection is done in the middle of crop season to prepare a list of all fields of a particular crop in the sample village, known as a frame of concerned crop. It also helps forecast crop yield based on grower's opinion, availability of input products, weather conditions, and field officers' expert judgment.



District's name

1	Attock	10	Gujrat	19	Lodhran	28	Rahim Yar Khan
2	Bahawalnagar	11	Hafizabad	20	Mandi Bahauddin	29	Rajanpur
3	Bahawalpur	12	Jhang	21	Mianwali	30	Rawalpindi
4	Bhakkar	13	Jhelum	22	Multan	31	Sahiwal
5	Chakwal	14	Kasur	23	Muzaffargarh	32	Sargodha
6	Chiniot	15	Khanewal	24	Nankana Sahib	33	Sheikhupura
7	Dera Ghazi Khan	16	Khushab	25	Narowal	34	Sialkot
8	Faisalabad	17	Lahore	26	Okara	35	Toba Tek Singh
9	Gujranwala	18	Layyah	27	Pakpattan	36	Vehari

Figure 1. The district-level variation of change (%) in wheat and rice cultivated area since 2003.

The third survey is taken during the harvest season by measuring crop yield in three experimental plots of 20 × 15 ft. These plots are selected randomly within each sample village using the crop frame. The sample data is interpreted and aggregated at the district by the statistical experts of CRS. The final crop production estimates are generated using crop areas generated through a complete census at the district level. This complete enumeration of crop area in each district is carried out by field officers of the Department of Revenue, Pakistan. CRS estimates the final crop production by multiplying district level measured yield values with the final area values. For this study, district-level crop yield data of wheat and rice was collected from CRS for the study period.

2.3. Remote sensing data

Two satellites Terra and Aqua carries a sensor Moderate Resolution Imaging Spectroradiometer (MODIS) that record the reflection from the earth in 36 different spectral bands. Two MODIS products MOD13Q1 and MYD13Q1 from Terra and Aqua, respectively, contain EVI data that was used in this study. These products have a spatial resolution of 250 m, and their composite temporal resolution is 8 days. A total of 46 images complete one calendar year. Based on the study area's crop calendar, images from Nov 2003 to Oct 2018 were used for this analysis. Three tiles (h24v05, h23v05, and h24v06) covers the entire study area.

2.4. Overall methodology

The flow diagram of the procedure adopted for this research is shown in Figure 2. An eight-day temporal MODIS-EVI at 250 m spatial resolution was first stacked by seasons for rice and wheat crops. A six-month EVI stack from November to April next year was prepared for the wheat crop. For the rice crop, a six-month EVI stack from May to October was prepared. From EVI stacks of rice and wheat, the maximum seasonal EVI (EVI_{max}) was extracted for each crop season. The Day of the Year (DOY) for EVI_{max} was used to evaluate the model's pre-harvest prediction ability. As no significant areal change in rice and wheat croplands in the selected districts, an existing crop cover data set was used to extract the EVI_{max} values. To match the reported crop yield data, the district average EVI_{max} was calculated for input to MLR models. The best performing MLR model for each crop was then selected to simulate the crop yield using the EVI_{max} value.

2.5. Machine learning regression models

2.5.1 Linear Regression (LR)

The linear regression models have been used in many previous yield estimation studies to indicate a linear or exponential relationship to single or multiple variables (Hou et al. 2019; Gaso, Berger, and Ciganda 2019). The slope and intercept are the two parameters that describe the linear relationship between the dependent and independent variables. To accommodate multiple dependent variables in LR modeling, a stepwise regression analysis is often used to eliminate non-significant variables in the linear expression. The other machine learning methods are more efficient in handling predictor variables' nonlinearity than simple LR models (Chlingaryan, Sukkarieh, and Whelan 2018).

2.5.2 Support Vector Regression (SVR)

In ML methods, a Support Vector Machine (SVM) is a training-based discriminative model that offers various distinctive edges in handling complex multidimensional

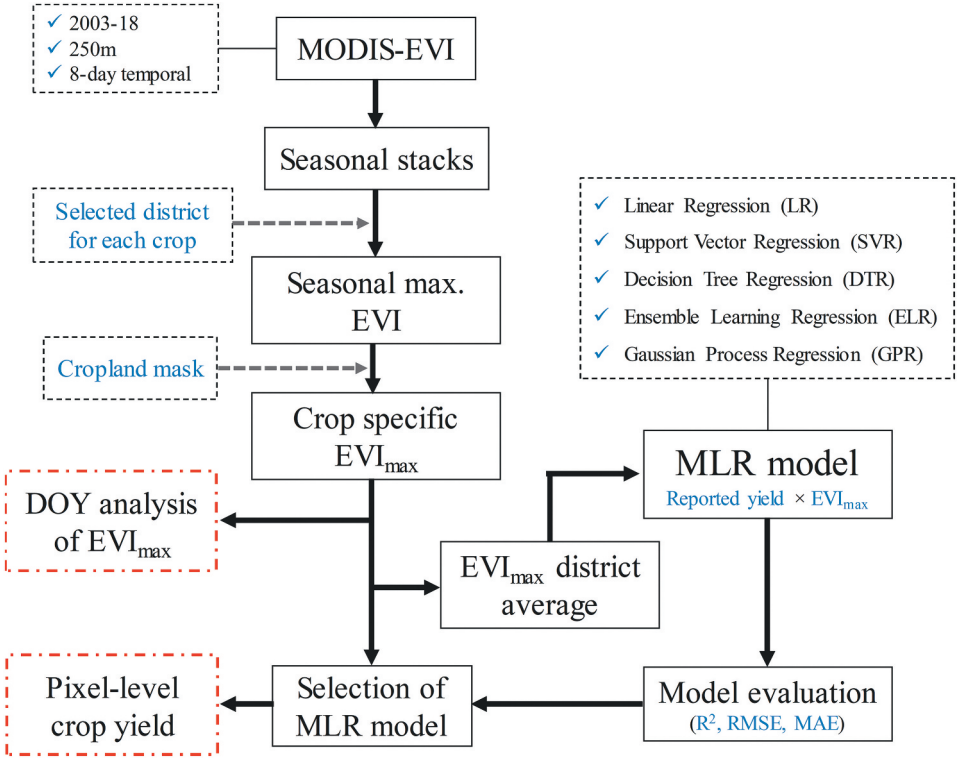


Figure 2. The flow diagram of the procedure adopted for this research.

data using hyperplane (Schölkopf and Smola 2001). Apart from being used as a classification method, SVM can also be used as a regression model with a few minor differences. The support vector regression model maintains all the main elements that distinguish the algorithm, i.e. maximal margin. Different kernel functions such as linear and Radial Basis Function (RBF) enable the SVR to operate in higher dimensions. For linear SVR, the function used to predict values $f(x)$ depending on the support vectors can be mathematically expresses as:

$$f(x) = \sum_{i=1}^{N} (a_i - a_i^*).\langle x_i, x\rangle + b \qquad (1)$$

Where x_i is the number of training dataset in N observations, a_i and a_i^* are non-negative real numbers known as Lagrange multipliers, and b is the intercept.

2.5.3 Decision Tree Regression (DTR)

A decision tree regression model is a non-parametric supervised ML method that develops a hierarchical tree structure by learning from training data. The DTR model divides the data into smaller subsets using decision nodes and leaf nodes. A decision node has further sub-classes, while the leaf node has an associated decision value (M. Xu et al. 2005). A deep tree structure leads to complex decision rules and is prone to overfitting. Alternately, to mitigate the problem of overfitting, multiple coarse decision tree structures are combined in Ensemble Learning Regression (ELR) for prediction. There are two commonly used ensemble methods. One method is to group decision trees is in parallel order called Bagging, and the other is in sequential order called Boosting. Bagging is used to reduce the variance in estimation by creating multiple random samples from the original dataset to train multiple models individually. The final predictions are determined by combining the predictions from all individual models. Boosting is an iterative method that alters the weight of an observation based on the last prediction. The main goal of boosting is to achieve higher accuracy (Zhou 2012). The mathematical expression of DTR is similar to LR model and nodes are defined using standard deviation and variance formulas.

2.5.4 Gaussian Process Regression (GPR)

The Gaussian process regression is another non-parametric kernel-based probabilistic model that makes a significant ML application impression. A well-trained GPR model not only predicts the response variable but also quantifies the uncertainty associated with it. The random fluctuations and the coefficients are estimated from the training data. A GPR is a function space model and supposes that the covariance between any two random variables is a multivariate Gaussian. A multivariate Gaussian is defined by its mean and covariance (kernel) function. The kernel function defines the spatial or temporal similarities between two random variables. There are multiple kernel functions (e.g. rational quadratic and

marten 5/2) available to choose from that capture the smoothness of the response variables. As GPR model is probabilistic model, an instance of response y can be modeled as:

$$P(y_i|f(x_i),x_i)\sim N\left(y_i|h(x_i)^T\beta+f(x_i),\sigma^2\right) \quad (2)$$

Where $P(y_i|f(x_i),x_i)$ is the density of the sample, σ^2 is the noise in the density, $f(x_i)$ is a latent variable developed for each observation x_i, h is explicit basis function, and β is a coefficient vector. More details on different parameters of GPR model can be found in Rasmussen (2003). The present study has compared LR, SVR, DTR, ELR (both bagged and boosted), and GPR models to predict crop yield in selected districts. The Bayesian optimization with 50 iterations was used to fine-tune hyperparameters in all MLR models.

2.6. Model evaluation

All regression models used fivefold cross-validation to indicate model performance in yield estimation. In the fivefold cross-validation, for each MLR model, the whole data is randomly divided into training and testing datasets for five simulations. The average value of evaluating coefficients from five simulations depict each model's efficacy. The comparison of all regression models led to the selection of the best predictive model for each crop. The best selected, trained model was used to simulate pixel-level crop yield information for the entire study period. The coefficient of determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) were used to evaluate each model. The formulas for R^2, RMSE, and MAE for one simulation are given below:

$$R=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}} \quad (3)$$

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(x_i-y_i)^2}{n}} \quad (4)$$

$$MAE=\frac{\sum_{i=1}^{n}|x_i-y_i|}{n} \quad (5)$$

Where x_i and y_i are the individual reported and estimated crop yield during the whole study period (2003–2018); $\bar{x}$ and $\bar{y}$ are the average values of reported and estimated yields in that period, respectively, and n is the total sample size for each crop. The higher value of R^2 and lower value of RMSE and MAE indicated a better performing model.

3. Results

3.1. Evaluation of MLR models

The comparison of five MLR models in predicting crop yield of rice and wheat using EVI_{max} is shown in Table 1. A fivefold cross-validation was used to evaluate each model. The GPR model outperformed in predicting yield for both crops. For wheat crop, it achieved R^2 of 0.75 with RMSE and MAE of 0.281 (t/ha) and 0.236 (t/ha), respectively. In comparison, the performances of other MLR models are lower but still acceptable for wheat yield prediction with $R^2 > 0.60$. For example, the LR was the least efficient model but still achieved 0.64, 0.351 (t/ha), and 0.290 (t/ha) for R^2, RMSE, and MAE, respectively. For rice crop, GPR model attained $R^2 = 0.68$, RMSE = 0.112 (t/ha), and MAE = 0.091 (t/ha). The DTR model was the least accurate in predicting rice yield with MAE = 0.154 (t/ha). For both crops, the SVR model performed the second best with small margins compared to GPR for predicting wheat and rice yield. At the district-level, the scatter plot between reported crop yield and simulated crop using the GPR model is shown in Figure 3 for both wheat and rice.

3.2. Spatial variation of crop yield

A district-level trained GPR model was further used to simulate the pixel-level yield information from 2003 to 2018. Figure 4 shows the spatial variation of average wheat yield at the pixel-level and district-level simulated using the GPR model during the study period. The results indicated that for the past 16 years an average annual yield of 2.60 (t/ha) with a standard deviation of ±0.48 (t/ha) was produced in the study area. The maximum wheat yield was observed in the Gujranwala district with an average of 3.19 ± 0.23 (t/ha). On average, the arid districts of Punjab province (Chakwal, Rawalpindi, and Attock) have the lowest yield of wheat (1.52 ± 0.27 t/ha) throughout the study period. The maximum yield variation of 0.41 (t/ha) was observed in the Rajanpur district, with an average yield of 2.53 (t/ha) during the study period. In southern Punjab, Khanewal, and Lodhran have the

Table 1. The evaluation statics of five MLR models in predicting wheat and rice yield using EVI_{max} value.

Crop	MLR models	R^2	RMSE	MAE
Wheat	LR	0.64	0.351	0.290
	DTR	0.67	0.313	0.261
	SVR	0.73	0.290	0.242
	GPR	0.75	0.281	0.236
	ELR	0.69	0.302	0.254
Rice	LR	0.55	0.139	0.146
	DTR	0.51	0.141	0.154
	SVR	0.65	0.128	0.102
	GPR	0.68	0.112	0.091
	ELR	0.51	0.155	0.120

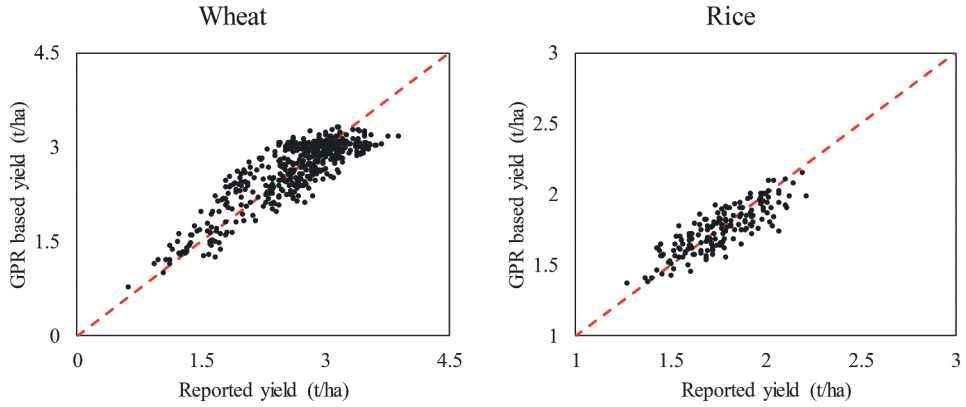


Figure 3. A district-level scatter plot comparison between reported and simulated yield of rice and wheat using the GPR model.

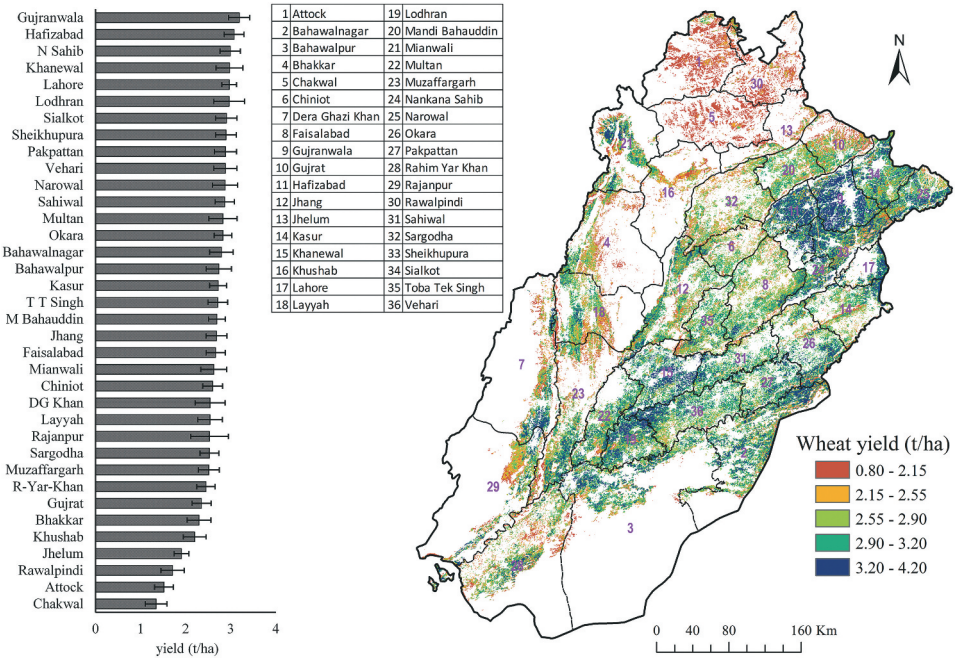


Figure 4. The pixel-level and district-level spatial variation of average wheat yield simulated using the GPR model during the study period.

highest wheat yield with an average of 2.97 ± 0.31 (t/ha). Central Punjab has an average yield of 2.67 ± 0.21 (t/ha).

The spatial variation of an average rice yield from 2003 to 2018 simulated by the GPR model is shown in Figure 5. The selected districts had an average yield of 1.86 ± 0.09 (t/ha) in the past 16 years. The maximum rice yield was estimated in the Gujranwala district (with an average of 2.06 ± 0.11 (t/ha)), while the Gujarat district had the least seasonal yield (with an average of 1.76 ± 0.07 (t/ha)). In Punjab province, the varieties of cultivated rice are divided into three main categories: a) basmati rice, b) long grain/non-basmati rice, and c) coarse/medium rice. The basmati and long grain/non-basmati have a similar yield percentage but the coarse/medium rice has a higher yield. The selected rice districts of this study are mostly cultivated under basmati or long grain/non-basmati variety. The coarser rice is mostly sown in the southern part of Punjab and Sindh province of Pakistan and is not considered in this study.

3.3. Timely yield estimation

The rice and wheat crops reach to the EVI_{max} value in the middle of their growth cycle during the greening stage. The day of the year to reach EVI_{max} was studied for each crop during the study period to assess the effectiveness of the proposed method on timely yield prediction. The median value of DOY of EVI_{max} and its spatial variation during the study period for both crops is shown in Figure 6. More than 75% of the area with wheat crop shows EVI_{max} in the two weeks of mid-February. The harvesting period of the wheat crop is from mid-April to mid-May. This implies

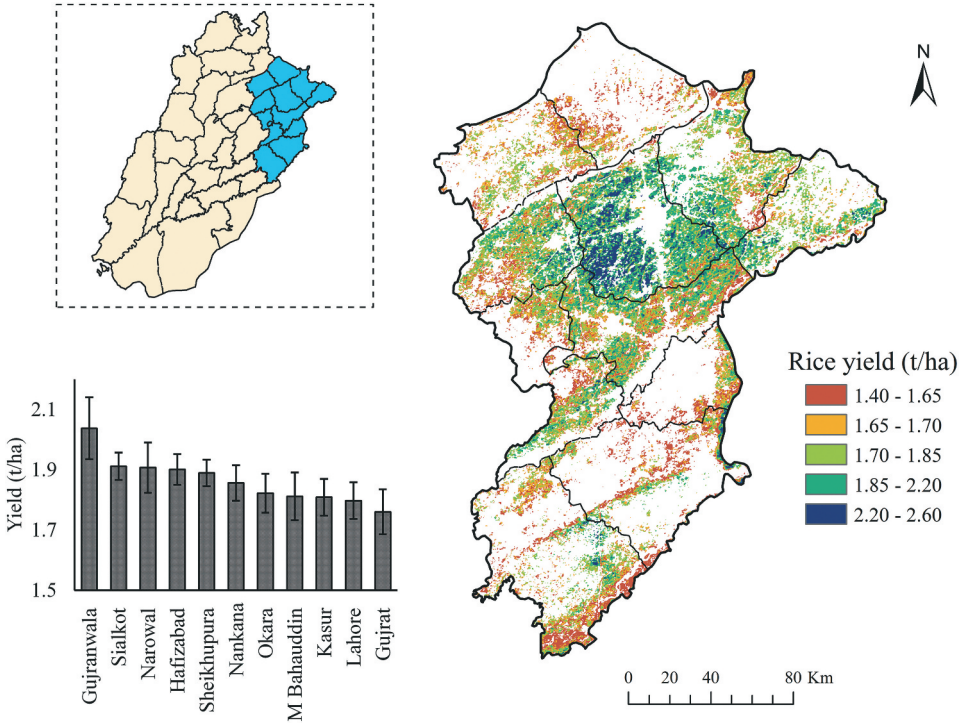
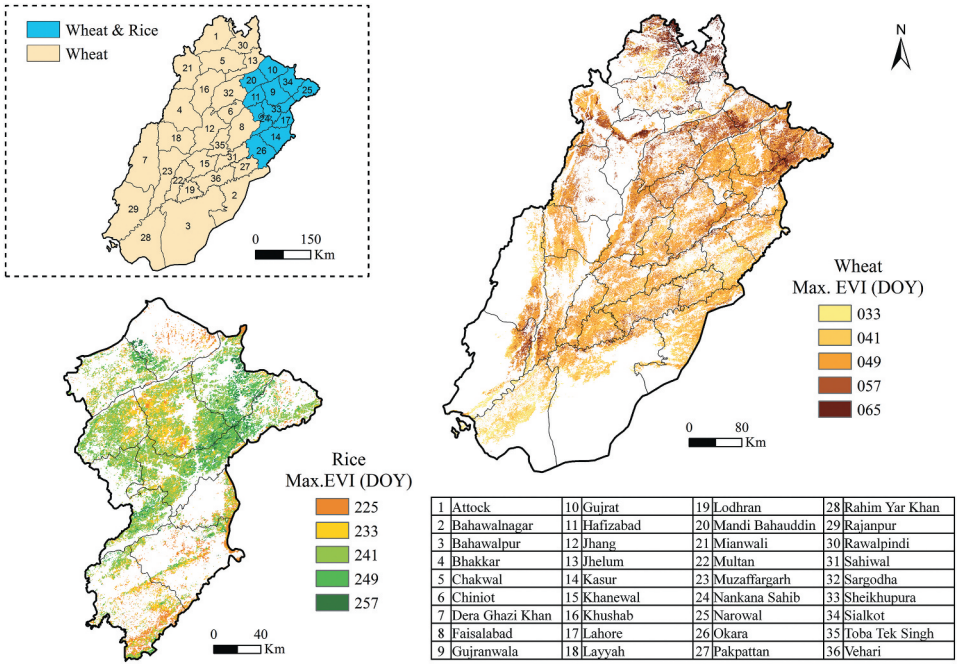


Figure 5. The spatial variation of an average rice yield from 2003–2018 simulated by the GPR model.



1	Attock	10	Gujrat	19	Lodhran	28	Rahim Yar Khan
2	Bahawalnagar	11	Hafizabad	20	Mandi Bahauddin	29	Rajanpur
3	Bahawalpur	12	Jhang	21	Mianwali	30	Rawalpindi
4	Bhakkar	13	Jhelum	22	Multan	31	Sahiwal
5	Chakwal	14	Kasur	23	Muzaffargarh	32	Sargodha
6	Chiniot	15	Khanewal	24	Nankana Sahib	33	Sheikhupura
7	Dera Ghazi Khan	16	Khushab	25	Narowal	34	Sialkot
8	Faisalabad	17	Lahore	26	Okara	35	Toba Tek Singh
9	Gujranwala	18	Layyah	27	Pakpattan	36	Vehari

Figure 6. The median value of the day of the year for EVI_{max} occurrence and its spatial variation during the study period for both crops.

that the wheat yield can be predicted using the EVI_{max} with GPR model 7–8 weeks before the harvest.

Similarly, for the rice region, more than 70% of the cultivated area shows EVI_{max} in the last two weeks of August. The rice harvesting season is from mid-October to late-November. The prediction of rice yield can be achieved 6–7 weeks before the start of the harvest. The spatial variation of EVI_{max} can also provide a general trend of crop harvest. In general, the wheat is harvested in the southern area earlier than in the northern area. The spatial trend of EVI_{max} can be analyzed according to various climatic conditions,

cropping sequence, and crop varieties cultivated throughout the region. However, this association analysis is beyond the scope of this study.

4. Discussion

The crop yield estimation is of paramount importance in recent times, especially for developing countries where food production is under high pressure. Satellite remote sensing data provides regular updates about crop conditions throughout its growth cycle. Many past studies have analyzed crop yield relation with remote sensing derived parameters at different crop growth stages. Most studies pointed out that the VIs and biophysical parameters around the crop's peak season are positively correlated with final yields for various crops (Dempewolf et al. 2014; Bolton and Friedl 2013; Sakamoto, Gitelson, and Arkebauer 2013; Ren et al. 2008). However, developing regression based on a single day VI during peak growth duration for a vast area is problematic as crop sowing time changes from one field to another. Climatic conditions throughout the region also played a crucial role in deciding the start of the crop season. To compensate this, these studies used the spatiotemporally aggregated information of crop phenology to estimate crop yield. Most past studies have also overlooked the use of cropland masks for aggregating remote sensing VI values to regress against reported crop yield that can be misleading in crop yield estimation and its spatial variation.

In this study, a crop mask was used to select VI values of a specific area. The use of the season's maximum EVI compensated for the variation in the crop's sowing timing and simplified the use of suitable crop phenology information for regression purposes. The seasonal EVI_{max} was then used to train various MLR models, and their prediction ability was compared to select the one with the best performance. This EVI_{max} based approach, however, cannot cater to the impact of extreme weather events after the crop's greenness stage as these events can greatly affect the crop yield during the critical stage of ripening before the harvest. Abbas and Mayo (2020) studied the effect of rainfall and temperature on rice production in Punjab, Pakistan, and reported a negative impact of rainfall during the rice ripening stage and its production. Crop lodging is a major yield-reducing factor that is connected to extreme weather events. The crop lodging during the ripening stage can seriously damage the production of cereal crops (Niu et al. 2016).

Nevertheless, this study has proven that the use of EVI_{max} with the GPR model can provide a reasonable estimate of crop productivity 7 weeks before the harvest. The DOY analysis based on MODIS has identified the spatial patterns of EVI_{max} occurrence throughout the region. Further studies using fine resolution remote sensing data based on identified DOY can improve the spatial resolution and quality of yield prediction. The timely information on the productivity of primary cereal crops helps to ensure regional food security and regulates market prices in regions with an agro-based economy. The spatial mapping of yield levels can also help policymakers improve the crop productivity of underperforming areas.

Furthermore, the pixel-level information of crop productivity levels can help to derive actual spatial diversity of crop management practices in the region. These management practices include the actual levels of irrigation and farm chemicals (fertilizers and pesticides) being applied in the region. Steduto et al. (2012) have expensively reported the direct nexus between crop water used and its yield for multiple crops including wheat and rice. Similarly, studies (Xu et al. 2019; Havlin and Heiniger 2020) have reported positive impact of soil fertility on crop yield. Therefore, by extrapolating the results of study, one can indirectly estimate the actual levels of crop management practices in the region that is a valuable information for a data-scarce developing regions like Pakistan. The information of DOY for EVI_{max} can be further explored to derive indirectly the crop calendar events (like crop sowing and harvest time) at regional scale. Such local-scale information of diverse crop management practices are a valuable input to robustly assess the impacts of changing climate on regional crop productivity.

5. Conclusions

This paper reports a study to develop an efficient crop yield forecasting model with a case study in a diversified agriculture region of Punjab, Pakistan. The crop-specific EVI_{max} from MODIS was used to train and test the predictive ability of five machine learning regression models. The GPR model was identified as the best prediction model for the yields of rice and wheat crops. The analysis of EVI_{max} revealed that the proposed method is capable of estimating wheat yield 7–8 weeks and rice yield 6–7 weeks before the harvest. Further studies using fine spatial resolution remote sensing datasets and incorporating the spatial information of EVI_{max} occurrence derived from this study will enhance yield results. The findings and method developed in this study would help to forecast regional crop productivity and produce timely crop yield predictions well before the harvest, which can help better management of crop market, food security, and rural development.

Data availability statement

The input data used in this research can be accessed freely from online sources. Remote sensing data used in this study can be downloaded via various web portals of NASA, e.g.

https://earthdata.nasa.gov/. The district-level reported crop yield data are available at the Crop Reporting Service (CRS), Punjab's website (http://crs.agripunjab.gov.pk/reports). All the derived data are included in the manuscript. The machine learning regression models were trained and tested using MATLAB's Regression Learner app. The MATLAB script to estimate pixel-level crop yield and DOY for EVI_{max} can be made available upon request to corresponding author.

Funding

The research is supported by the Natural Science Foundation of China (NSFC) General Research (Grant number 41971386) and Hong Kong Research Grant Council (RGC) General Research Fund (Grant number 12301820). The work is a part of PhD research funded by Hong Kong PhD Fellowship Scheme (HKPFS);Natural Science Foundation of China (NSFC) General Program (Grant number 41971386); Hong Kong Research Grant Council (RGC) General Research Fund (Grant number 12301820).

Notes on contributors

Ali Ismaeel has a PhD in geography and master's degree in remote sensing and GIS with bachelor's degree of agricultural engineering. His research broadly covers ecological modeling using remote sensing and GIS approaches with prime focus on landuse/landcover mapping, vegetation trend analysis, hydrological modeling and cropland ecosystem modeling.

Qiming Zhou is a Professor of Geography, Associate Dean of Faculty of Social Sciences (Research) and Founding Director of Centre for Geo-Computation Studies at Hong Kong Baptist University. His research interests cover a broad area of geo-spatial information science, particularly in geo-computation and remote sensing applications. He has been actively engaged in research such as digital terrain analysis, climate change and its impacts on regional and global ecosystems, landuse and land cover change detection, and GIS and remote sensing applications to urban, environment and natural resource management.

ORCID

Qiming Zhou http://orcid.org/0000-0003-0934-0602

References

Abbas, S., and Z.A. Mayo. 2020. "Impact of Temperature and Rainfall on Rice Production in Punjab, Pakistan." *Environment, Development and Sustainability* 23 (2): 1706–1728. doi:10.1007/s10668-020-00647-8.

Bolton, D. K., and M.A. Friedl. 2013. "Forecasting Crop Yield Using Remotely Sensed Vegetation Indices and Crop Phenology Metrics." *Agricultural and Forest Meteorology* 173: 74–84. doi:10.1016/j.agrformet.2013.01.007.

Chlingaryan, A., S. Sukkarieh, and B. Whelan. 2018. "Machine Learning Approaches for Crop Yield Prediction and Nitrogen Status Estimation in Precision Agriculture: A Review." *Computers and Electronics in Agriculture* 151: 61–69. doi:10.1016/j.compag.2018.05.012.

Craig, M., and D. Atkinson. 2013. "A Literature Review of Crop Area Estimation", Report by UN-FAO, Rome, Italy. http://www.fao.org/fileadmin/templates/ess/documents/meetings_and_workshops/GS_SAC_2013/Improving_methods_for_crops_estimates/Crop_Area_Estimation_Lit_review.pdf

Dempewolf, J., B. Adusei, I. Becker-Reshef, M. Hansen, P. Potapov, A. Khan, and B. Barker. 2014. "Wheat Yield Forecasting for Punjab Province from Vegetation Index Time Series and Historic Crop Statistics." *Remote Sensing* 6 (10): 9653–9675. doi:10.3390/rs6109653.

Fritz, S., L. See, J. C. Laso Bayas, F. Waldner, D. Jacques, I. Becker-Reshef, A. Whitcraft, et al. 2019. "A Comparison of Global Agricultural Monitoring Systems and Current Gaps." *Agricultural Systems* 168 :258–272. doi:10.1016/j.agsy.2018.05.010.

Gaso, D. V., A. G. Berger, and V. S. Ciganda. 2019. "Predicting Wheat Grain Yield and Spatial Variability at Field Scale Using a Simple Regression or a Crop Model in Conjunction with Landsat Images." *Computers and Electronics in Agriculture* 159: 75–83. doi:10.1016/j.compag.2019.02.026.

Giannakis, E., and A. Bruggeman. 2015. "The Highly Variable Economic Performance of European Agriculture." *Land Use Policy* 45: 26–35. doi:10.1016/j.landusepol.2014.12.009.

Havlin, J., and R. Heiniger. 2020. "Soil Fertility Management for Better Crop Production." *Agronomy* 10 (9): 1349. doi:10.3390/agronomy10091349.

Hou, M. J., F. Tian, T. Zhang, and M. S. Huang. 2019. "Evaluation of Canopy Temperature Depression, Transpiration, and Canopy Greenness in Relation to Yield of Soybean at Reproductive Stage Based on Remote Sensing Imagery." *Agricultural Water Management* 222: 182–192. doi:10.1016/j.agwat.2019.06.005.

Huang, J. F., X. Z. Wang, X. X. Li, H. Tian, and Zhuokun Pan. 2013. "Remotely Sensed Rice Yield Prediction Using Multi-Temporal NDVI Data Derived from NOAA's-AVHRR." *PLoS ONE* 8 (8): e70816. doi:10.1371/journal.pone.0070816.

Huang, J. X., J. L. Gómez-Dans, H. Huang, H. Y. Ma, Q. L. Wu, P. E. Lewis, S. L. Liang, et al. 2019. "Assimilation of Remote Sensing into Crop Growth Models: Current Status and Perspectives." *Agricultural and Forest Meteorology* 276 :107609. doi:10.1016/j.agrformet.2019.06.008.

Jin, X. L., L. Kumar, Z. H. Li, H. K. Feng, X. G. Xu, G. J. Yang, and J. H. Wang. 2018. "A Review of Data Assimilation of Remote Sensing and Crop Models." *European Journal of Agronomy* 92: 141–152. doi:10.1016/j.eja.2017.11.002.

Johnson, M. D., W. W. Hsieh, A. J. Cannon, A. Davidson, and F. Bédard. 2016. "Crop Yield Forecasting on the Canadian Prairies by Remotely Sensed Vegetation Indices and Machine Learning Methods." *Agricultural and Forest Meteorology* 218: 74–84. doi:10.1016/j.agrformet.2015.11.003.

Liu, J. G., J. L. Shang, B. D. Qian, T. Huffman, Y. S. Zhang, T. F. Dong, Q. Jing, and T. Martin. 2019. "Crop Yield Estimation Using Time-Series MODIS Data and the Effects of Cropland Masks in Ontario, Canada." *Remote Sensing* 11 (20): 2419. doi:10.3390/rs11202419.

Lobell, D. B. 2013. "The Use of Satellite Data for Crop Yield Gap Analysis." *Field Crops Research* 143: 56–64. doi:10.1016/j.fcr.2012.08.008.

Maya, G. P. S., and R. Bhargavi. 2019. "A Novel Approach for Efficient Crop Yield Prediction." *Computers and Electronics in Agriculture*. doi:10.1016/j.compag.2019.104968.

Mfuka, C., E. Byamukama, and X. Y. Zhang. 2020. "Spatiotemporal Characteristics of White Mold and Impacts on Yield in Soybean Fields in South Dakota." *Geo-Spatial Information Science* 23 (2): 182–193. doi:10.1080/10095020.2020.1712265.

Niu, L.Y., S. W. Feng, W. H. Ding, G. Li, and A. Zhang. 2016. "Influence of Speed and Rainfall on Large-Scale Wheat Lodging from 2007 to 2014 in China." *PLoS ONE* 11 (7): e0157677. doi:10.1371/journal.pone.0157677.

Petersen, L. K. 2018. "Real-Time Prediction of Crop Yields from MODIS Relative Vegetation Health: A Continent-Wide Analysis of Africa." *Remote Sensing* 10 (11): 1726. doi:10.3390/rs10111726.

Prins, A., and A. V. Niekerk. 2020. "Crop Type Mapping Using LiDAR, Sentinel-2 and Aerial Imagery with Machine Learning Algorithms." *Geo-Spatial Information Science* 24 (2): 215–227. doi:10.1080/10095020.2020.1782776.

Rasmussen, C. E. 2003. "Gaussian Processes in Machine Learning." In *Summer School on Machine Learning*, 63–71. doi:10.1007/978-3-540-28650-9_4.

Rehman, A., J. D. Luan, A. A. Chandio, M. Shabbir, and I. Hussain. 2017. "Economic Outlook of Rice Crops in Pakistan: A Time Series Analysis (1970–2015)." *Financial Innovation* 3 (1): 1–9. doi:10.1186/s40854-017-0063-z.

Ren, J. Q., Z. X. Chen, Q. B. Zhou, and H. J. Tang. 2008. "Regional Yield Estimation for Winter Wheat with MODIS-NDVI Data in Shandong, China." *International Journal of Applied Earth Observation and Geoinformation* 10 (4): 403–413. doi:10.1016/j.jag.2007.11.003.

Sakamoto, T., A. A. Gitelson, and T. J. Arkebauer. 2013. "MODIS-Based Corn Grain Yield Estimation Model Incorporating Crop Phenology Information." *Remote Sensing of Environment* 131: 215–231. doi:10.1016/j.rse.2012.12.017.

Schölkopf, B., and A. J. Smola. 2001. "Support Vector Machines." In *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, 187–188. MIT Press. https://ieeexplore.ieee.org/document/6282654

Simoncini, R., I. Ring, C. Sandström, C. Albert, U. Kasymov, and R. Arlettaz. 2019. "Constraints and Opportunities for Mainstreaming Biodiversity and Ecosystem Services in the EU's Common Agricultural Policy: Insights from the IPBES Assessment for Europe and Central Asia." *Land Use Policy* 88: 104099. doi:10.1016/j.landusepol.2019.104099.

Steduto, P., T. C. Hsiao, E. Fereres, and D. Raes. 2012. "Crop Yield Response to Water", FAO Irrigation and Drainage Paper No. 66., ISBN: 9789251072745.

Ujoh, F., T. Igbawua, and M. O. Paul. 2019. "Suitabili ty Mapping for Rice Cultivation in Benue State, Nigeria Using Satellite Data." *Geo-Spatial Information Science* 22 (4): 332–344. doi:10.1080/10095020.2019.1637075.

UN. 2019. "World Population Prospects 2019: Ten Key Findings." Accessed 24 June 2019. https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/files/documents/2020/Jan/un_2019_wpp_methodology.pdf

Xu, M., P. Watanachaturaporn, P. K. Varshney, and M. K. Arora. 2005. "Decision Tree Regression for Soft Classification of Remote Sensing Data." *Remote Sensing of Environment* 97 (3): 322–336. doi:10.1016/j.rse.2005.05.008.

Xu, X. P., P. He, M. F. Pampolino, S. J. Qiu, S. C. Zhao, and W. Zhou. 2019. "Spatial Variation of Yield Response and Fertilizer Requirements on Regional Scale for Irrigated Rice in China." *Scientific Reports* 9 (1): 1–8. doi:10.1038/s41598-019-40367-2.

Zhou, Z. H. 2012. *Ensemble Methods: Foundations and Algorithms*. Florida: CRC Press. doi:10.1201/b12207.