Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography
Vol. 34, No. 4, 383-390, 2016
http://dx.doi.org/10.7848/ksgpc.2016.34.4.383

ISSN 1598-4850(Print)
ISSN 2288-260X(Online)
Original article

Machine Learning Approaches to Corn Yield Estimation Using Satellite Images and Climate Data: A Case of Iowa State

Kim, Nari[1] · Lee, Yang-Won[2]

Abstract

Remote sensing data has been widely used in the estimation of crop yields by employing statistical methods such as regression model. Machine learning, which is an efficient empirical method for classification and prediction, is another approach to crop yield estimation. This paper described the corn yield estimation in Iowa State using four machine learning approaches such as SVM (Support Vector Machine), RF (Random Forest), ERT (Extremely Randomized Trees) and DL (Deep Learning). Also, comparisons of the validation statistics among them were presented. To examine the seasonal sensitivities of the corn yields, three period groups were set up: (1) MJJAS (May to September), (2) JA (July and August) and (3) OC (optimal combination of month). In overall, the DL method showed the highest accuracies in terms of the correlation coefficient for the three period groups. The accuracies were relatively favorable in the OC group, which indicates the optimal combination of month can be significant in statistical modeling of crop yields. The differences between our predictions and USDA (United States Department of Agriculture) statistics were about 6-8 %, which shows the machine learning approaches can be a viable option for crop yield modeling. In particular, the DL showed more stable results by overcoming the overfitting problem of generic machine learning methods.

Keywords : Crop Yield, Machine Learning, Remote Sensing, Climate Data

1. Introduction

Monitoring crop yield is important for many agronomy issues such as farming management, food security and international crop trade. Because South Korea highly depends on imports of most major grains except for rice, reasonable estimations of crop yields are more required under recent conditions of climate changes and various disasters.

Remote sensing data has been widely used in the estimation of crop yields by employing statistical methods such as regression model. Prasad *et al*. (2006) conducted multivariate regression analyses to estimate corn and soybean yields in Iowa using MODIS (Moderate Resolution Imaging Spectroradiometer) NDVI (Normalized Difference Vegetation Index), climate factors and soil moisture. Ren *et al*. (2008) presented regression models for the estimation of winter wheat yields using MODIS NDVI and weather data in Shandong, China. Kim *et al*. (2014) estimated corn and soybean yields using several MODIS products and climatic variables for Midwestern United States (US) and represented prediction errors of about 10 %. Hong *et al*. (2015) built multiple regression models using MODIS NDVI and weather data to estimate rice yields in North Korea and showed the RMSE of 0.27 ton/ha. Most of the previous studies are based on the multivariate regression analysis using the relationship between crop yields and agro-environmental factors such as vegetation index, climate variables and soil properties.

Received 2016. 07. 22, Revised 2016. 08. 03, Accepted 2016. 08. 23
1) Division of Earth Environmental System Science, Pukyong National University (E-mail: kim.nari13@gmail.com)
2) Corresponding Author, Member, Department of Spatial Information Engineering, Pukyong National University (E-mail: modconfi@pknu.ac.kr)

Machine learning, which is an efficient empirical method for classification and prediction, is another approach to crop yield estimation. Jiang *et al*. (2004) adopted ANN (Artificial Neural Network) technique for estimation of winter wheat yields using AVHRR (Advanced Very High Resolution Radiometer) dataset, and the ANN model showed a higher accuracy than multivariate regression models. Jaikla *et al*. (2008) estimated rice yields using SVM (Support Vector Machine) and compared the result with the simulation of DSSAT (Decision Support System for Acrotechnology Transfer) model, which showed a similar performance. Kuwata and Shibasaki (2015) employed DL (Deep Learning) methods for estimation of corn yields for Illinois and presented that the DL contributed to higher accuracy than SVM. Despite the efficient predictability of machine learning techniques, the applications in crop yield estimation are relatively insufficient, and the comparative studies among various machine learning methods for crop yield estimation have not reported yet.

The objective of this study is to estimate crop yields by employing several major techniques for machine learning such as SVM, RF (Random Forest), ERT (Extremely Randomized Trees) and DL, and to present the comparisons of validation statistics among them. We used satellite images from MODIS and the climate reanalysis data created by PRISM (Parameter-Elevation Regressions on Independent Slopes Model) for the machine learning analyses. To improve the prediction accuracies according to phenology effects, we set up three types of data period: (1) May to September, (2) July and August and (3) an optimal combination of the months.

2. Data and Method

2.1 Study area

Iowa is a state in the Midwestern US and belongs to the Corn Belt (Fig. 1). Iowa produces approximately 18 % of the US corn yields, which is the highest ranking in the US (USDA, 2012). Out of the 99 counties of Iowa State, we selected 94 counties whose cropland exceeded 10 % of the county area. The study period is between 2004 and 2014 according to the data availability.

Fig. 1. Study area

2.2 Data

2.2.1 Remote sensing data

Satellite remote sensing data was acquired from NASA (National Aeronautics and Space Administration) and ESA (European Space Agency) CCI (Climate Change Initiative). The Terra/MODIS products by NASA such as NDVI, EVI (Enhanced Vegetation Index), LAI (Leaf Area Index), FPAR (Fraction of Photosynthetically Active Radiation), GPP (Gross Primary Production) and ET (Evapotranspiration) are closely related to crop yields. Also, SM (Soil Moisture) dataset was obtained from ESA CCI, which produces the most complete and consistent global soil moisture data on the grid of 0.25° using active and passive microwave sensors. Table 1 shows the summary of dataset used. Previous studies (Prasad *et al*., 2006; Na *et al*., 2014; Kim *et al*., 2014) presented these variables were associated with the corn yield.

2.2.2 Climate data

The PRISM Climate Group (http://www.prism.oregonstate.edu/) provides daily and monthly reanalysis of seven climate elements in the US: precipitation (PPT), maximum temperature (Tmax), minimum temperature (Tmin), mean temperature (Tmean), mean dew point temperature (TDmean), minimum vapor pressure deficit (VPDmin) and maximum vapor pressure deficit (VPDmax).

We used monthly data for PPT, Tmax, Tmin, Tmean at the 4-km resolution.

2.2.3 Crop yield data

As a reference dataset, county-level yield statistics of corn were obtained from the NASS (National Agricultural Statistics Service) of USDA (United States Department of Agriculture) (http://quickstats.nass.usda.gov). The unit of corn yield (bushels per acre) was converted to ton per hectare for convenience sake.

Table 1. Summary of dataset used in this study

Data		Spatial Resolution	Temporal Resolution	Source
Remote Sensing	NDVI	1 km	Monthly	NASA EarthData
	EVI			
	LAI		8-day	
	FPAR			
	GPP			
	ET		Monthly	The University of Montana
	Land Cover	500 m	Yearly	NASA EarthData
	SM	0.25°	Daily	ESA CCI
Climate	PPT	4 km	Monthly	PRISM Climate Group
	Tmax			
	Tmin			
	Tmean			
Yield	Corn	County	Yearly	USDA

2.2.4 Data processing

Because cropland areas for each county should be first determined, we extracted the pixels which were recorded as cropland (land cover ID = 12) throughout the period of 2004-2014 from the MODIS land cover data. Fig. 2 shows that the distribution of the cropland pixels is similar to the pattern of major counties for corn production in Iowa. For these cropland pixels, we constructed a database including satellite images and climate variables. Crop yield statistics were the values accumulated by county, so the satellite and climate data need to be averaged at the county level. We employed the zonal operation to summarize the pixel values for a given county.

Various environmental factors related to crop yields can have different sensitivities to growing seasons. Hence, we derived 13 cases for month combination such as MJJAS (from May to September), each individual month between May and September (May, Jun, Jul, Aug and Sep), two successive months (MJ, JJ, JA and AS), and three successive months (MJJ, JJA and JAS) for calculation of the correlation coefficients (Table 2). From these combinations, we selected three period groups: (1) MJJAS for the whole growing season, (2) JA as the group having mostly highest correlation coefficients and (3) OC for the optimal combination of the periods in terms of the correlation coefficient (shaded in gray in Table 2). In order to estimate the corn yield in the 94 counties in Iowa, we built a matchup database consisting of 11 input variables from satellite images (NDVI, EVI, LAI, FPAR, GPP, ET and SM) and climate dataset (PPT, Tmin, Tmax and Tmean) for the three period groups between 2004 and 2012.

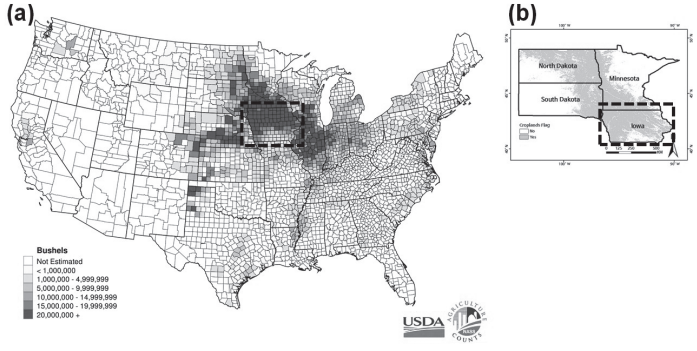

Fig. 2. (a) Corn yields by county and (b) cropland pixels derived from MODIS land cover data (Iowa State in the dashed line)

Table 2. Correlation coefficients of the variables against corn yields, 2004-2014

	MJJAS	May	June	July	Aug.	Sep.	MJ	JJ	JA	AS	MJJ	JJA	JAS
NDVI	0.589	-0.436	0.272	0.637	0.784	0.547	-0.171	0.449	0.805	0.683	0.098	0.663	0.769
EVI	0.684	-0.364	0.278	0.634	0.794	0.602	-0.058	0.490	0.802	0.738	0.254	0.709	0.798
LAI	0.650	-0.417	-0.162	0.551	0.746	0.407	-0.290	0.330	0.721	0.709	0.209	0.625	0.763
FPAR	0.313	-0.411	-0.213	0.517	0.756	0.313	-0.336	0.088	0.743	0.595	-0.126	0.424	0.758
GPP	0.439	-0.401	0.076	0.543	0.554	0.342	-0.173	0.473	0.558	0.505	0.302	0.527	0.542
ET	0.471	-0.149	-0.078	0.187	0.690	0.389	-0.125	0.228	0.697	0.644	0.147	0.535	0.703
SM	0.355	-0.017	0.094	0.465	0.431	0.309	0.051	0.320	0.489	0.388	0.246	0.395	0.456
PPT	0.113	-0.073	-0.011	0.277	0.067	0.047	-0.049	0.131	0.216	0.081	0.087	0.145	0.195
Tmax	-0.575	-0.350	-0.448	-0.567	-0.542	-0.136	-0.453	-0.577	-0.597	-0.426	-0.563	-0.595	-0.574
Tmin	-0.423	-0.342	-0.347	-0.489	-0.166	0.046	-0.434	-0.513	-0.414	-0.085	-0.544	-0.430	-0.352
Tmean	-0.544	-0.357	-0.433	-0.546	-0.374	-0.057	-0.467	-0.573	-0.538	-0.295	-0.576	-0.544	-0.523

2.3 Methods

2.3.1 Support vector machine

SVM is a powerful technique for general classification which can minimize the classification error of existing machine learning techniques (Vapnik, 1998). For estimation or prediction, regression methods are combined with each classified group. SVM finds the optimal separating classifier between the two classes by maximizing the margin between support vectors using the kernel functions such as linear, Gaussian RBF (Radial Basis Function), polynomial and hyperbolic tangent (Cortes and Vapnik, 1995; Karatzoglou *et al*., 2006). The Gaussian RBF were used in our experiment.

2.3.2 Random forest

The RF, which is an improved version of CART (Classification and Regression Trees), is an ensemble method using bootstrap aggregating (Breiman, 2001). RF makes decision trees by extracting random samples from the training data and predicts results through the vote for classification or averaging of the regression using a large number of trees (Ali *et al*., 2012). In our experiment, the number of trees were 500, and the number of variables used for splitting nodes were set to n/3 (n = number of input variables). In addition, the out-of-bag error was used as the criterion of model suitability.

2.3.3 Extremely randomized trees

ERT is an ensemble classifier method using unpruned decision trees. ERT is different from the other tree-based ensemble methods such as RF, in that it divides nodes by randomly choosing cut-points and that it uses the complete learning sample (no bootstrap copying) to grow the trees (Geurts *et al*., 2006). Such randomization is based on the bias-variance analysis like the Friedman test (Friedman, 1997). Randomization increases bias and variance of individual trees, but they can be attenuated by averaging over a sufficiently large ensemble of trees. In our experiment, the number of trees and the number of variables used for splitting nodes were set to the same as those of RF.

2.3.4 Deep learning

DL is a machine learning method similar to ANN but is capable of processing the complicated, huge input data by learning tasks by using feed-forward multi-layer network (Ali *et al*., 2015). Training process of DL usually consists of pre-training and fine-tuning. Pre-training is the phase of data processing by using unsupervised learning for improving the generalization error of trained deep architectures. Fine-tuning by supervised learning is performed to improve the classification error (Erhan *et al*., 2010). Our experiment used a 200×200 multi-layer network.

2.3.5 Validation

The leave-one-year-out cross-validation, also known as the Jackknife, was conducted to examine the accuracies of the corn yield estimation by machine learning methods. We calculated the mean bias, MAE (Mean Absolute Error), RMSE (Root-Mean-Square Error), MAPE (Mean Absolute Percentage Error) and the correlation coefficient (r) between the observed and predicted yields during the period of 2004-2014.

3. Results and Discussion

We implemented the machine learning methods (SVM, RF, ERT and DL) using R libraries (https://www.r-project.org/). We first estimated the corn yields using the MJJAS dataset for the whole growing season, and the results were compared with the USDA yield statistics. The leave-one-year-out cross-validation produced 11 sets of validation results for each year between 2004 and 2014. Table 3 shows the averages of the 11-year validation results in terms of the mean bias, MAE, MAPE, RMSE and r. Fig. 3 shows the scatter plots of the predicted corn yields against USDA statistics between 2004 and 2014. According to the results, DL achieved the highest accuracy with the correlation coefficient of 0.776 and the RMSE of 0.844 ton/ha, although three methods (RF, ERT and DL) presented similar accuracies. In particular, RF and ERT showed very similar results with the correlation coefficients of 0.651 and 0.654, respectively, and the RMSE were 0.879 and 0.891 ton/ha, respectively. This is because the two approaches are based on regression trees even if their randomization strategies for tree splitting are somewhat different. The SVM showed the lowest accuracy with the correlation coefficient of 0.560 and the RMSE of 0.959 ton/ha.

Tables 4 and 5 show the 11-year averaged statistics for JA and OC, respectively. When comparing the results of the three period groups (MJJAS, JA and OC), the correlation coefficients for SVM were almost the same (MJJAS=0.590, JA=0.575, OC=0.606), but the RMSE of OC (0.852 ton/ha) were somewhat improved than those of MJJAS (0.959 ton/ha) and JA (0.936 ton/ha). As for RF and ERT, the correlation coefficients (JA=0.774 and 0.774, OC=0.772 and 0.785, respectively) and the RMSE (JA=0.803 and 0.802 ton/ha, OC=0.767 and 0.756 ton/ha, respectively) were similar for both JA and OC, showing improved results than those of the MJJAS. Hence, it is notable that the seasonal sensitivities of corn yields were well captured by the RF and ERT methods. The DL method produced the highest accuracies for the three period groups in terms of the correlation coefficients

Table 3. Validation statistics for the period group MJJAS (May to September)

	Mean bias (ton/ha)	MAE (ton/ha)	RMSE (ton/ha)	MAPE (%)	r
SVM	0.112	0.730	0.959	8.1	0.590
RF	0.063	0.666	0.879	7.3	0.651
ERT	0.091	0.674	0.891	7.4	0.654
DL	-0.031	0.657	0.844	6.9	0.776

Table 4. Validation statistics for the period group JA (July and August)

	Mean bias (ton/ha)	MAE (ton/ha)	RMSE (ton/ha)	MAPE (%)	r
SVM	0.085	0.721	0.936	8.0	0.575
RF	0.019	0.616	0.803	6.6	0.774
ERT	0.023	0.616	0.802	6.6	0.774
DL	-0.169	0.709	0.901	7.5	0.796

Table 5. Validation statistics for the period group OC (optimal combination of month)

	Mean bias (ton/ha)	MAE (ton/ha)	RMSE (ton/ha)	MAPE (%)	r
SVM	0.072	0.650	0.852	7.3	0.606
RF	0.002	0.057	0.767	6.3	0.772
ERT	0.015	0.568	0.756	6.1	0.785
DL	-0.059	0.608	0.787	6.5	0.800

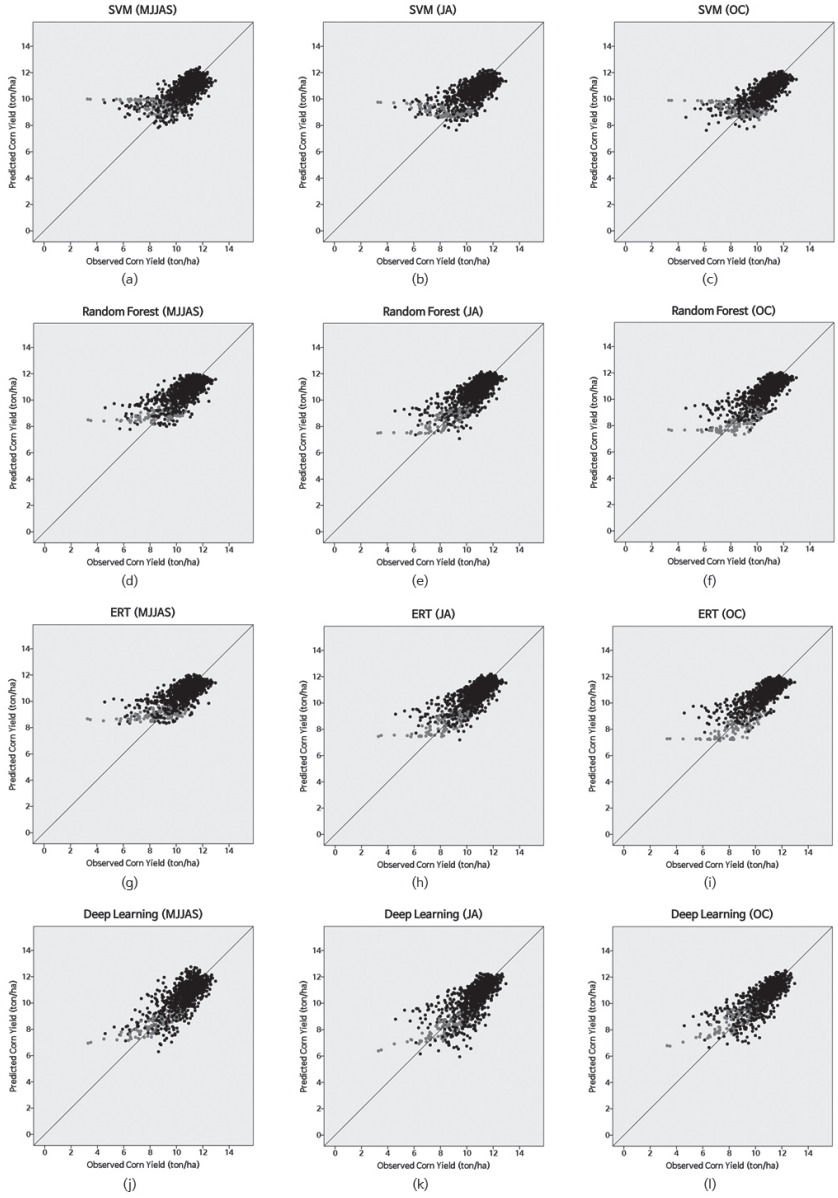

Fig. 3. Scatter plots for observed vs. predicted corn yields, 2004-2014 (red dots: 2012, black dots: all years except for 2012)

(MJJAS=0.776, JA=0.796 and OC=0.800, respectively).

Moreover, the DL presented more stable results in the scatter plots while the other three methods had a tendency of overfitting. Machine learning techniques such as SVM, RF and ERT can have an overfitting problem, which occurs when a model is very complex with many parameters and shows a poor predictive performance by overreacting to minor fluctuations in dataset. The red dots in Fig. 3 were the cases of 2012, in which an extreme drought occurred in the Midwestern US. The machine learning models for prediction of 2012 (that is, the models built using the data of the years except for 2012, for the Jackknife) were too trained for non-drought years (except for 2012), so that they could not predict the corn yield under conditions of abrupt drought. However, the DL method can overcome the overfitting problem by a pre-training process based on unsupervised learning (Erhan *et al*., 2010). Fig. 3(j), 3(k) and 3(l) for the DL method shows that the red dots for 2012 are more closely located around the 1:1 line.

4. Conclusions

This paper described the estimation of corn yields in Iowa State using four machine learning techniques such as SVM, RF, ERT and DL, and presented the comparisons of the validation statistics among them. We set up the three period groups (MJJAS, JA and OC) to examine the seasonal sensitivities of the corn yields. In overall, the DL method showed the highest accuracies in terms of the correlation coefficient for all the period groups. The accuracies were relatively favorable in the OC group, which indicates an optimal combination of month can be influential in statistical modeling of crop yields. The differences between our predictions and the USDA statistics were about 6-8 %, which shows the machine learning approaches can be a viable option for crop yield modeling. In particular, the DL showed more stable results by overcoming the overfitting problem of generic machine learning methods. To utilize temporal characteristics of crop yields, time-series machine learning techniques such as RNN (Recurrent Neurual Network) are challengeable as a future work. A sensitivity test to examine the contribution of climate change to the crop yields by including or excluding the climate variables can be another future work.

Acknowledgment

This work was supported by the Research Grant of Pukyong National University (2015).

References

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., and Notarnicol, C. (2015), Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data, *Remote Sensing*, Vol. 7, No. 12, pp. 16398-16421.

Ali, J., Khan, R., Ahmad, N., and Maqsood, I. (2012), Random forests and decision trees, *International Journal of Computer Science Issues*, Vol. 9, No. 5, pp. 272-278.

Breiman, L. (2001), Random forests, *Machine Learning*, Vol. 45, No. 1, pp. 5-32.

Cortes, C. and Vapnik, V. (1995), Support-vector network, *Machine Learning*, Vol. 20, No. 3, pp. 273-297.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., and Vincent, P. (2010), Why does unsupervised pre-training help deep learning?, *Journal of Machine Learning Research*, Vol. 11, pp. 625-660.

Friedman, J.H. (1997), On bias, variance, 0/1-loss, and the curse-of-dimensionality, *Data Mining and Knowledge Discovery*, Vol. 1, pp. 55-77.

Geurts, P., Ernst, D., and Wehenkel, L. (2006), Extremely randomized trees, *Machine Learning*, Vol. 63, No. 1, pp. 3-42.

Hong, S.Y., Na, S.I., Lee, K.D., Kim, Y.S., and Baek, S.C. (2015), A study on estimating rice yield in DPRK using MODIS NDVI and rainfall data, *Korean Journal of Remote Sensing*, Vol. 31, No. 5, pp. 441-448. (in Korean with English abstract)

Jaikla, R., Auephanwiriyakul, S., and Jintrawet, A. (2008), Rice yield prediction using a support vector regression method, *Proceedings of Electrical Engineering/ Electronics, Computer, Telecommunications and*

Information Technology 2008, 14-17 May, Krabi, Thailand, pp. 908-913.

Jiang, D., Yango, X., Clinton, N., and Wang, N. (2004), An artificial neural network model for estimating crop yields using remotely sensed information, *International Journal of Remote Sensing*, Vol. 25, No. 9, pp. 1723-1732.

Karatzoglou, A., Meyer, D., and Hornik, K. (2006), Support vector machines in R, *Journal of Statistical Software*, Vol. 15, No. 9. pp. 1-28.

Kim, N., Cho, J., Shibasaki, R., and Lee, Y.W. (2014), Estimation of corn and soybean yields of the US Midwest using satellite imagery and climate dataset, *Journal of Climate Research*, Vol. 9, No. 4, pp. 315-329. (in Korean with English abstract)

Kuwata, K. and Shibasaki, R. (2015), Estimating crop yields with deep learning and remotely sensed data, *Proceedings of 2015 IEEE International Geoscience and Remote Sensing Symposium*, 26-31 July, Milan, Italy, pp. 858-861.

Na, S., Hong, S., Kim, Y., and Lee, K. (2014), Estimation of corn and soybean yields based on MODIS data and CASA model in Iowa and Illinois, USA, *Korean Journal of Soil Science and Fertilizer*, Vol. 47, No. 2, pp. 92-99. (in Korean with English abstract)

Prasad, A.K., Chai, L., Singh, R.P., and Kafatos, M. (2006), Crop yield estimation model for Iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation*, Vol. 8, pp. 26-33.

Ren, J.Q., Chen, Z.X., Zhou, Q.B., and Tang, H.J. (2008), Regional yield estimation for winter wheat with MODIS-NDVI data in Shandong, China. *International Journal of Applied Earth Observation and Geoinformation*, Vol. 10, pp. 403-413.

USDA (2012), Census of agriculture, *United States Department of Agriculture,* https://www.agcensus.usda.gov/ (last date accessed: 17 August 2016).

Vapnik, V. (1998), *Statistical Learning Theory*, Wiley, New York, NY.