Crop yield estimation model for Iowa using remote sensing and surface parameters

Anup K. Prasad a, Lim Chai b, Ramesh P. Singh a,b,*, Menas Kafatos b

a Department of Civil Engineering, Indian Institute of Technology, Kanpur 208016, India
b Center for Earth Observing and Space Research, School of Computational Sciences, George Mason University, Fairfax, VA 22030, USA

Received 21 October 2004; accepted 8 June 2005

Abstract

Numerous efforts have been made to develop various indices using remote sensing data such as normalized difference vegetation index (NDVI), vegetation condition index (VCI) and temperature condition index (TCI) for mapping and monitoring of drought and assessment of vegetation health and productivity. NDVI, soil moisture, surface temperature and rainfall are valuable sources of information for the estimation and prediction of crop conditions. In the present paper, we have considered NDVI, soil moisture, surface temperature and rainfall data of Iowa state, US, for 19 years for crop yield assessment and prediction using piecewise linear regression method with breakpoint. Crop production environment consists of inherent sources of heterogeneity and their non-linear behavior. A non-linear Quasi-Newton multi-variate optimization method is utilized, which reasonably minimizes inconsistency and errors in yield prediction.

Minimization of least square loss function has been carried out through iterative convergence using pre-defined empirical equation that provided acceptable lower residual values with predicted values very close to observed ones ($R^2 = 0.78$) for Corn and Soybean crop ($R^2 = 0.86$) for Iowa state. The crop yield prediction model discussed in the present paper will further improve in future with the use of long period dataset. Similar model can be developed for different crops of other locations.

Keywords: Crop yield prediction model; NDVI; Iowa; Soybean; Corn

1. Introduction

Monitoring of crop conditions is important for the economic development of any nation. The use of remote sensing has proved to be very important in monitoring the growth of agricultural crops and in irrigation scheduling. Efforts have been made to develop various indices for different crops of different regions throughout the globe. The production of crop and prediction of crop yield have direct impact on year-to-year national and international economies and play an important role in the food management (Hayes and Decker, 1996). Using remote sensing data, efforts have been made to develop various indices such as:
normalized difference vegetation index (NDVI), vegetation condition index (VCI) and temperature condition index (TCI). These indices are commonly used for drought detection, monitoring excessive soil wetness, assessment of weather impacts on vegetation and evaluation of vegetation health and productivity (Unganai and Kogan, 1998; Kogan, 2001, 2002; Kogan et al., 2003; Singh et al., 2003). The NDVI data have been used extensively in vegetation monitoring, crop yield assessment and forecasting (Hayes et al., 1982; Benedetti and Rossinini, 1993; Quarmby et al., 1993).

The US Corn Belt provides approximately 80% of the overall maize production for the entire US, and accounts for 36% of the global maize production (USDA, 1987). The United States Department of Agriculture (USDA) forecasts crop supply and demand estimates including expected crop yields every month during the crop season beginning in early January. The ability of the CERES-Maize model to estimate annual fluctuations in maize production for the US Corn belt was tested for the years 1982–1985 (Hodges et al., 1987). Spatial interactions in the CROPGRO-Soybean and CERES-Maize models and comparison of simulated and measured data are also studied (Batchelor et al., 2002). Optimizing crop-growth/yield models for Corn and Soybeans crops in USA was evaluated using the multi temporal high-resolution airborne digital imagery.

NDVI has been considered to be a useful way for crop yield assessment models using various approaches from simple integration to more complicated transformation. NDVI reflects vegetation greenness, thus it indicates levels of healthiness in the vegetation development. Although vegetation development of crop fields may differ from those of natural vegetation because of human influences involved such as irrigation, use of fertilizer and pesticides, NDVI is considered as a valuable source of information for the crop conditions. Different methods such as neural network (Stoikos, 1995), autoregressive (AR) state-space models (Wendroth et al., 2003), least-square regression (Jones, 1982), exponential-linear (EL) crop growth algorithm (Oroda, 2001) and numerical crop yield model (Hayes et al., 1982) have been used to predict crop yield with moderate success.

The ground and satellite (NOAA, Meteosat, etc.) measurements are commonly used to deduce various parameters such as evapotranspiration (Doorenbos and Kassam, 1979; Oroda, 2001), NDVI, soil type (Garcia-Paredes et al., 2000), light, carbon dioxide, temperature, water and the rate of growth and development (Monteith, 1981) and crop–weather relations (Watson, 1963; Baier, 1977; Frere and Popov, 1979; WMO, 1982; van Keulen, 1987; NCMRWF, 1990; Jain and Ranjana, 2000) are increasingly used to predict crop yield. Present crop yield estimation is based on various methods and data sources like field surveys, expert knowledge, trend analysis, regression analysis, statistical models and crop growth simulation models. In this paper, we have developed a crop yield prediction model based on Iowa Corn and Soybean yield estimates utilizing NDVI, surface temperature (ST), precipitation and soil moisture (SM).

2. Methods

2.1. Selection of a crop region

State of Iowa (Fig. 1a) belongs to the US Corn Belt, recorded the highest harvested cropland in 1997 (USDA, 1997). We have considered 10 top counties of Corn and Soybean production in 2000 of Iowa state, which is total combination of 14 counties to develop a crop yield assess model for Corn and Soybeans. Iowa state is lying in the humid temperate zone which can be approximated as a rectangular area with corner coordinates (97°W, 43.5°N), (97°W, 40.5°N), (90°W, 40.5°N) and (90°W, 43.5°N). The total area of Iowa is about 35,756,390 and 11,700,000 acres consists of planted Corn for grain in the year 2001, which is about 32.7% of the total area. Iowa state produced 17.9% of all Corn for grain in the year 2001, which is about 32.7% of the total area. Iowa state produced 17.9% of all Corn for grain in the year 2001, ranking the highest production in the US (USDA, 1997) and Soybean farming is the second largest crop after Corn, and about 10,920,000 acres were harvested in the year 2001.

2.2. Data

Crop yield data (Fig. 1b) of Corn and Soybean for Iowa state have been used, Corn crop is planted in early May, growth in biomass occurs from June to September and is ready for harvesting in September. Therefore, temporal annual average of NDVI, soil
moisture, surface temperature and rainfall (RF) data for period May to September have been used in the present analysis for 19 years from 1982 to 2001 (excluding 1994). Soybean crop season starts with active sowing period from mid of May to early June. Crop growth occurs from June to September and harvesting mostly in October. For Soybean, temporal annual average of NDVI, SM, ST and rainfall data for the months June to September for 19 years period (1982 to 2001, excluding 1994) have been used for analysis. Monthly composite NDVI data were spatially averaged over Iowa region (annual mean of growth season average; Corn (May to September) and Soybean (June to September)), from 1982 to 2001. Fig. 2a–d shows the annual growth season average of NDVI, SM, ST and RF from 1982 to 2001.

2.2.1. National Agricultural Statistics Service (NASS) Corn and Soybeans yield estimates

Corn and Soybean yield estimates have been obtained electronically from NASS/USDA database site (http://www.usda.gov/nass/). The Corn yield (Fig. 1b) estimates are available for the years from 1866 to 2004 whereas the Soybeans yield (Fig. 1b) estimates are available for the years from 1924 to 2004. Note that Soybeans belong to the category of ‘Oilseeds and Cotton’. The datasets provide planted and harvested acres, yield per acre and production in bushel.

2.2.2. NDVI

We have used 8 km × 8 km monthly composite continental NDVI datasets from the pathfinder advanced very high-resolution radiometer (AVHRR) collected by National Oceanic and Atmospheric Administration’s (NOAA) polar orbiters from 1982 to 2001. NDVI is level 3 data derived from Channel 1 (visible band) and Channel 2 (near-infrared band). The visible wavelength attenuation encountered in AVHRR observations do not show significant difference from the Landsat multi-spectral scanner (MSS) and Thematic Mapper (TM) sensors or the French Satellite Probatoire d’Observation de la Terre (SPOT) sensors. The attenuation is considerably more sensitive to water vapor in the near infrared than the other land observing sensors (Goward et al., 1991). The NDVI derived from NOAA 11 data shows values 0.05 higher than earlier NOAA missions for African desert. An error in the solar zenith angle (SZA) has also been discovered. However, NDVI in the pathfinder AVHRR land (PAL) dataset received by the users in HDF format are less affected by errors in the solar zenith angle (http://daac.gsfc.nasa.gov/CAMPAIGN_DOCS/LAND_BIO/zenith_angle_mem.html).

Quantitative interpretation of NDVI is complicated by numerous intervening factors such as instrument calibration, incident solar irradiance, nominal atmospheric attenuation, variable spatial
resolution, anisotropy with off-nadir views, and cloud occurrence. These factors combine to produce a global dataset that has, at best, a measurement precision of $\pm 0.1$ NDVI units (10% error) over 1 year at a temporal resolution somewhere between 10 days and 1 month (Goward et al., 1991).

The dry matter accumulation of the Corn leaves mature by mid July and almost constant from mid July to the beginning of September (Iowa State University of Science and Technology, 1996). The maximum vegetation greenness of Iowa region generally occurs in August. We have examined spatially averaged NDVI for the months of Corn growing season, which is used for Corn yield estimates. NDVI data are permanently missing for September to December 1994; therefore, 1994 data is excluded from model due to non-availability of September NDVI.

### 2.2.3. Precipitation

Rainfall data (http://www1.ncdc.noaa.gov/pub/data/cirs/) is available as monthly average within a climatic division calculated using equal weight to stations reporting both temperature and precipitation within a division. Monthly average total rainfall (unit: mm) is taken from above dataset from 1982 to 2001.

### 2.2.4. Temperature

Monthly surface temperature data is taken from NOAA NCEP-NCAR CDAS-1 monthly diagnostic surface temperature database (unit: Kelvin) (http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP-NCAR/).

### 2.2.5. Soil moisture

Soil moisture data have been taken from NOAA NCEP CPC global monthly soil moisture dataset (unit: mm) http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP/). Soil moisture is based on the water balance in the soil.

### 2.3. Methodology

Crop yield is considered as dependent variable that varies diversely with independent variables like NDVI, SM, ST and RF. Variations of NDVI, SM,
ST and RF data do not follow any distinct linear combination and with respect to crop yield. It is therefore difficult to model such a dynamic relationship using conventional linear methods like multi-variable multiple regression. Non-linear estimation approach is used to compute the relationship between a set of independent variables and a dependent variable. A two-piece empirical equations is devised and solved using non-linear Quasi-Newton method. Crop yield estimation equation with coefficients is derived by minimizing loss function for Corn and Soybean crop separately based on the 19 years dataset.

Non-linear piecewise linear regression with breakpoint (Quasi-Newton method) (Belegundu and Chandrupatla, 1999; Setiono et al., 2002) have been used to develop model for prediction of crop yield. Various steps of this model involve (1) identifying an initial model, (2) iterative convergence using the “stepping criteria,” and (3) terminating the search when either stepping criteria or number of iterations allowed reached its limit.

Empirical equation is based on piecewise linear regression method with breakpoint. Quasi-Newton methods have been used for multi-variate optimization (Belegundu and Chandrupatla, 1999). It is non-linear method that has been used to minimize least square loss function through iterative convergence of pre-defined empirical equation. In Quasi-Newton Method, the first-order derivative of the function at a point is computed to find the slope of a function at that point. Subsequent second-order derivative indicates how fast the slope is changing at the respective point and its direction. The Quasi-Newton method evaluates the function at different points at each step in order to estimate the first-order derivatives and second-order derivatives, which is used to find out the minimum of the loss function. Quasi-Newton is an iterative method that is primarily governed by minimization of chosen loss function (i.e., achieved global minima point where observed is closest possible to simulated value which in principle can be a 100% match). Chosen loss function can differ based on objective and here it is a commonly applied least square loss function, i.e., square of the difference between predicted and observed value (objective is to achieve lowest possible difference between observed and predicted value).

The iterative method works for multi-independent variables and dependent variable crop yield both above and below the breakpoint. A non-linear optimization approach achieves acceptable lower residual values with predicted values very close to observed values.

2.4. Model

Coefficients of empirical equation have been obtained using this method. Breakpoint (m) chosen is the mean of 19 year Corn or Soybean crop yield of Iowa. In rare case, if breakpoint (m) becomes equal to crop yield, condition of crop yield > breakpoint should be applied. The model empirical equations for Corn and Soybean crops, thus obtained with coefficients (Table 1) is given as:

crop yield = \( (c_1 + (a_1 \times NDVI) + (a_2 \times SM) + (a_3 \times ST) + (a_4 \times RF) ) \) for crop yield < breakpoint \( m \)

or \( (c_2 + (b_1 \times NDVI) + (b_2 \times SM) + (b_3 \times ST) + (b_4 \times RF) ) \) for crop yield > breakpoint \( m \)

where NDVI, normalized difference vegetation index; SM, soil moisture (mm); ST, surface temperature (Kelvin); RF, rainfall (mm); \( c_1, c_2, a_i, b_i \) for \( i = 1-4 \) are coefficients (Table 1); \( m \) = mean Corn or Soybean crop yield (1982–2001, excluding 1994) (breakpoint).

<table>
<thead>
<tr>
<th>Model variable</th>
<th>Coefficients</th>
<th>Corn</th>
<th>Soybean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>( c_1 )</td>
<td>-12733.9</td>
<td>445.2006</td>
</tr>
<tr>
<td>NDVI</td>
<td>( a_1 )</td>
<td>156.4193</td>
<td>-1.12634</td>
</tr>
<tr>
<td>SM</td>
<td>( a_2 )</td>
<td>-0.17942</td>
<td>-0.02641</td>
</tr>
<tr>
<td>ST</td>
<td>( a_3 )</td>
<td>51.13045</td>
<td>-1.25083</td>
</tr>
<tr>
<td>Rainfall</td>
<td>( a_4 )</td>
<td>-31.7037</td>
<td>-0.45429</td>
</tr>
<tr>
<td>Constant</td>
<td>( c_2 )</td>
<td>-4039.71</td>
<td>-21.01</td>
</tr>
<tr>
<td>NDVI</td>
<td>( b_1 )</td>
<td>80.32816</td>
<td>17.57264</td>
</tr>
<tr>
<td>SM</td>
<td>( b_2 )</td>
<td>0.017716</td>
<td>-0.02533</td>
</tr>
<tr>
<td>ST</td>
<td>( b_3 )</td>
<td>16.03083</td>
<td>0.967642</td>
</tr>
<tr>
<td>Rainfall</td>
<td>( b_4 )</td>
<td>-8.51827</td>
<td>-0.45243</td>
</tr>
<tr>
<td>Breakpoint</td>
<td></td>
<td>124.4737</td>
<td>40.36842</td>
</tr>
</tbody>
</table>

\( R \) = 0.88055, 0.93053

Variance accounted (%) = 77.53, 86.58

\( R^2 \) = 0.78, 0.86
Loss function used is least square, i.e. \( L_f = (\text{observed} - \text{predicted})^2 \).

The Quasi-Newton method utilizes this loss function to arrive at a solution closest possible to observed data. At each iteration loss function is computed to minimize square of difference between the observed and predicted crop yield using pre-defined empirical equation. The method is an optimization process, which runs as long as initial values, stepping values, number of iterations and convergence criteria are favorable. It terminates if any of these bounding conditions are fulfilled. Therefore, loss function can approach theoretically up to \( R^2 \) 100%. It depends on degree to which independent variables considered control dependable variable and absence of any other major governing factor affecting crop yield in a year. This approach can give results, which are closer to real value. Small dataset (7–10 years) gave better result than large dataset (say 19 years data) due to less variation in pattern of SM, ST, RF, NDVI and crop yield.

NDVI, SM, ST and RF data are major indicators or variables controlling the normal crop growth. This approach can be used to predict any shortfall in crop yield using empirical equations. Coefficients used in derived equation largely depend on pool of historical data. Studying NDVI, SM, ST and RF data region wise before harvesting period and employing above methodology (with modifications) can be used to predict crop yield for that season.

3. Results and discussion

The fitted model (Fig. 3b) agrees well with Soybean crop. Residual values in individual years are within acceptable limits with more or less even distribution of difference from observed crop yield. Similarly, Corn crop (Fig. 3a) also shows even distribution of residual values for all years except for 1993. Still the model agrees well for the year 1993 that has witnessed a steep fall (~46%) in observed crop yield compared with the preceding year (Table 2). It shows that crop yield is possibly also affected by some other factors that may dominate in some years causing steep fall. A moderate to high \( R^2 \)-values 0.78 for Corn and 0.86 for Soybean show that in most years, crop yield is largely governed by variables considered in model.

However, other factors like pests, diseases and human activities can cause local variations in predicted crop yield. This is a serious limitation to any forecasting method including this. However, inclusion of NDVI in model partly takes care of loss due to diseases or pests that directly affect vegetation. Crop yield prediction method is expected to yield good prediction results due to low residual values comparing with the historical data. The model can be optimized and evolve to be more rugged with growing historical data for better prediction. This exercise includes data for 19 years and yield better prediction results considering year-to-year variation of controlling factors.
This model can be further improved with usage of high-resolution data with availability of multi-year data. This method involves simple input dataset and can be suitably modified to add or drop variables depending on climate, region and crop type and can be easily extended for other crops. Modification in approach is required to use this methodology in prediction mode. For instance, a weekly or 10 days composite of input data (before harvesting month) can be compared with other measures of crop yield like dry matter productivity (DMP) and net primary productivity (NPP) for multiple years to arrive at predicted crop yield for current year.

4. Conclusion

Crop production environment consists of inherent sources of heterogeneity due to numerous parameters. The model discussed in the present paper reasonably minimizes inconsistency and errors in yield prediction giving high $R^2$-values with maximum accounting of variability in model. The model takes care of most of the parameters, which control the crop yield. This method can be used to predict crop yield for other crops as well as Corn and Soybeans. Based on data obtained before harvest, crop yield can be predicted with acceptable accuracy. Piecewise linear regression equation with breakpoint (Quasi-Newton method) can be extended to other countries as well, where crop production is primarily dependent on weather and climatic conditions. The model developed in the present paper shows a promising result, which can be useful for forecasting crop yields such as Corn and Soybeans and other crops in regional and global scales.

Acknowledgements

We are grateful to Dr. Katarzyna Dabrowska-Zielinske and one anonymous Reviewer for their comments and suggestions, which have helped us to improve the original version of the manuscript.

References


