Crop Monitoring as an
E-agricultural tool in
Developing Countries

WHEAT YIELD PREDICTION MODELS FOR THE HUAIBEI PLAIN

Reference: *E-AGRI_D44.1_Wheat_Yield_Prediction_on_Huaibei*
Author(s): Qinghan Dong, Herman Eerens and Beier Zhang
Version: 2.0
Date: 24/02/2014

DOCUMENT CONTROL

Signatures

Author(s) : Qinghan Dong

Reviewer(s) : Qinghan Dong

Approver(s) :

Issuing authority :

Change record

Release	Date	Pages	Description	Editor(s)/Reviewer(s)
1.0	01/01/2014			Qinghan Dong
2.0	25/02/2014			Qinghan Dong

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Crop yields can be assessed by using the low-resolution imagery registered by synoptic observation systems, such as NOAA-AVHRR (active since early 80's) or SPOT-VEGETATION (available since 1998). Such an assessment can usually be achieved by examining so-called vegetation indices retrieved from these systems, as a measure for plant growth and development. Since the appearance of the oldest and simplest vegetation variable NDVI (Normalized Difference Vegetation Index), other model based indices such as, Fraction of Absorbed Photosynthetically Active Radiation (fAPAR) or Dry Mass Productivity (DMP) are widely used. The statistical variables (maximum, minimum or mean etc.) derived from these vegetation times series enabled to detect anomalies in growth conditions. On the other hand, these biophysical variables are also widely used as predictors in crop yield forecasting models. This work-package has objectives to analyses all biophysical, technological and meteorological factors that affect wheat yield on the Huaibei plain and establish the yield prediction models for each of the prefectures using these driving factors as explanatory variables.

1. Introduction

1.1. Remote sensing and the crop yield prediction

Agricultural production can be affected by many factors, such as technological, biophysical or climatological factors. Many of these drivers can be assimilated /modelled or estimated using remote sensing approaches.

In this domain, agro-meteorological models had long been the main approach for yield assimilation before the advent of satellite imagery. Since the year 80', the introduction of coarse resolution satellite sensors, starting with NOAA-AVHRR, provided a potential tool for an objective and near real-time observation on vegetation growth. Since then, other satellites payloaded with sensors of similar observation capability have joined the class, including SPOT-VGT in 1998 and Terra-MODIS in 2000. This category of sensors with a spatial resolution between 250 m and 1 km assume particularly well the task of vegetation monitoring thanks to their large geographic coverage and high revisiting frequency.

The monitoring can be usually achieved by analysing so-called biophysical variables or vegetation indices retrieved from these sensors. Normalized Difference Vegetation Index is the oldest and simplest vegetation variable to be computed and applied in the crop monitoring. Another widely used biophysical variable is the fraction of absorbed photo-synthetically active radiation (fAPAR) based on inversion of canopy reflectance models. The third vegetation state attribute is the "Dry Matter Productivity" (DMP, in kg/ha/day) or the increase in dry matter biomass on a daily base is calculated following the Monteith approach. It takes into account the metrological conditions such as daily minimal, maximal temperature and solar radiation.

In addition, the long-term average or so-called historical year for these vegetation state variables at monthly and decadal steps is produced for the time series of NOAA-AVHRR and SPOT-VEGETATION. The difference values between the actual data and long-term average is computed to determine the status of actual growth in comparison with the historical average and their probability. That is why the time series produced by these sensors should be sufficiently long (8-10 years) so that an analysis with reference to historical data can be performed with a reasonable accuracy

Moreover, these earth observation systems have to cover the target areas with a relevant frequency so that a multi-day synthesis of registrations is possible. This synthesis is essential to remove the cloud perturbation and fill in any missing data.

As our institute is in charge of the pre-processing and distribution of SPOT-VGT data worldwide, the SPOT-VGT imagery becomes the reference raw for vegetation monitoring at European level and used in this research.

1.2. The study region of Huaibei Plain and its agriculture

The Huaibei Plain is located on the North of the Anhui province and on the south edge of the North China Plain, composed of 6 prefectures.

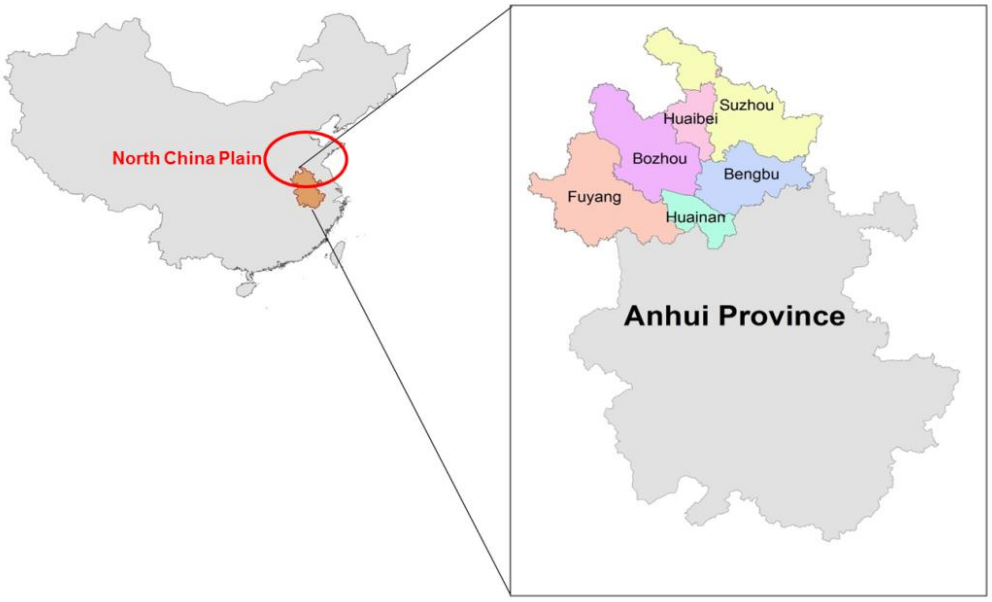

Figure 1: China is subdivided in 30 provinces and further in prefectures and counties. The figure shows the province of Anhui and its six prefectures. Main land-use feature is cropland, pastures is very exceptional.

Agriculture on the plain is well diversified because of the prevailed climate and occupies an important place in the country. The main agriculture products in the region are wheat, soybean, maize and cotton. Two growth seasons per year is the dominant cropping pattern with winter wheat followed by maize or soybean in the summer.

Figure 2: Mixed cropping pattern in the summer season in Mengcheng County

2. Data and methods

2.1. Data

2.1.1. Statistical data for wheat yield on the Huaibei Plain

The statistics of the wheat yield from 2000 to 2011 for six prefectures on the Huaibei Plain were collected from the National statistical Bureau.

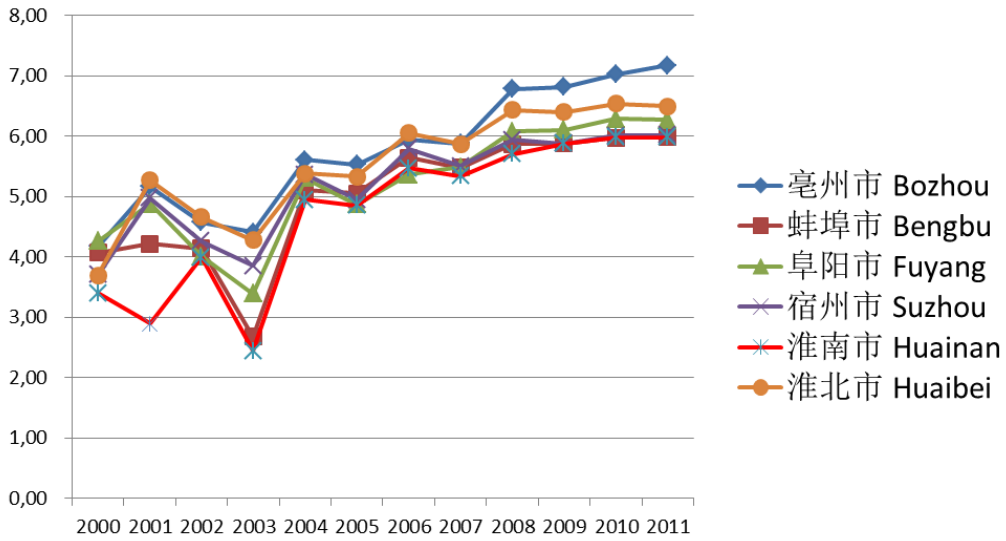

Figure 3: Wheat yield (ton/hectare) during the 2000 and 2011 for six prefectures of Huaibei Plain.

2.1.2. Remote sensing data

SPOT-VEGETATION imagery: the **Centre for Image Processing** (CVB), hosted at VITO, is responsible for the processing, archiving and distribution of all VGT-information. The two basic products are:

- VGT-P: Individual registrations ("segments"), calibrated and geo-corrected, but without atmospheric correction, hence they contain top-of-atmosphere (TOA) reflectance.
- VGT-S10: ten-daily global composites with TOC (top-of-canopy) reflectances and NDVI, are derived from all P-segments registered in a given dekad (10 days). For each pixel, the "best available" observation is selected (using *maximum value compositing*) and atmospherically corrected. The global S10 also comprises a Status Mask image (SM) which indicates the conditions of the selected observation: clear, cloud, snow/ice, error, etc...

The VGT data series started in April 1998 and has been continued until today without any interruption. All data are projected in the WGS84-Lon/Lat system with a resolution of 1°/112 (i.e. about 1 km around a great circle).

The VGT-data for the Huaibei area were extracted from the global S10 dataset. Although the region of interest only covers the northern part of Anhui, we extracted the data of whole province with following boundaries:

- Longitude: from 114.799107° to 119.799107°
- Latitude : from 29.2991071° to 34.8080357°
- Nr. of columns=560, nr. of records=617, nr. of pixels=345 520.

Images format: all images are converted to the ENVI-format represented by an .img file associated with a metadata (or header) file with extension .hdr.

NDVI layer: the layer in byte is rescaled (V=digital number) according to:

NDVI = -0.08 + 0.04*V, with V=0-250

The upper range (V=251-255) is used to label pixels with aberrant observations (251=error/missing, 252=cloud, 253=snow, 254=water, 255=other exception). This information is extracted from the Status Mask and from the land cover map GLC2000 (for the distinction between land and sea).

fAPAR and DMP layers: two new layers (images), relative to two biophysical variables, fAPAR (fraction of absorbed PAR, 400-700 nm) and DMP (Dry Matter Productivity), are computed. DMP is generated using a Monteith approach from fAPAR and external meteorological information (daily solar radiation and T_{min}/T_{max}). Daily meteorological data are needed for the DMP-computations. They are derived from ECMWF information and expressed in global grids with a resolution of 0.25°.

Table 1 summarizes the involved biophysical variables in this study. All variables expressed in image format follow the naming convention: ***spyyttv*.img/hdr**, where

- *s* = sensor: s=v for VGT
- *p* = periodicity: p=t for 10-daily periodicity
- *yy* = registration year**, yy=98,99,00,01,...,11**
- ***tt* =dekad** in year: tt=01, 02,..., 36. (every month is subdivided in three dekads, the first two always comprise 10 days meaning 1st-10th and 11st-20th, while the third one has variable length, 8/9 days for February, 10 or 11 days for other months).
- *v* = suffix for the type of "variable" (see table 1).

For instance: vt9810i contains the flagged (but non-smoothed) NDVI from the standard VGT-S10 for the first dekad of April 1998 (10^{th} dekad of the year).

Table 1: **Biophysical variables derived from SPOT-VGT (DT (Data-type): 1=unsigned byte with V=0 → 255, 2=signed short integer with V=-32768 → +32767)**

v	CONTENTS	DT	SCALING	FLAGS
i	NDVI non-smoothed	1	NDVI [-] = -0.08 + 0.004•V (V=0-250)	251=error/missing, 252=cloud, 253=snow/ice, 254=water, 255=other exception
k	NDVI smoothed			
a	fAPAR non-smoothed	1	fAPAR [%] = 0.5•V (V=0-200)	
b	fAPAR smoothed			
p	DMP noon-smoothed	2	DMP [kgDM/ha/day] = 0.01•V (V=0-32767)	-1 = error/missing, -2 = cloud, -3 = snow, -4 = water, -5 = other exception
y	DMP smoothed			

Auxilary geographical and environmental data:

- The digital terrain model **GTOPO30**.
- The **GLC2000** land use map (hard classification at 1km resolution).
- **GlCropV2**: JRC-FoodSec of the European Commission produced a new global 0/1-cropmask at 250m resolution in 2011. This crop mask was a compilation from different data sources: GlobCover V2.2, CORINE-2000, AfriCover, the SADC data set and the USGS Cropland Use Intensity data set. The mask covers exactly the same region as the global SPOT-VGT dataset, with a 4-times higher spatial resolution, 1°/112/4 or roughly 250m along a great circle. The mask image is framed such that each VGT-pixel exactly covers 4x4 mask pixels. From this 0/1-mask we derived a 1km resolution "Area Fraction Image" (AFI) which indicates for each 1km pixel the area fraction covered by cropland. This AFI is called GlCropV2 and is used in this research.

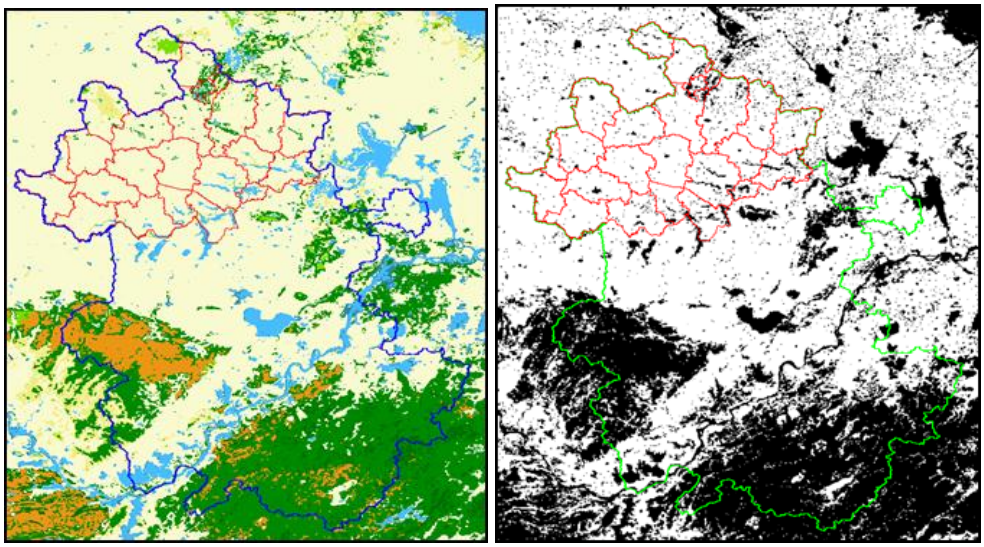

Figure 4: The left map shows an Extract of GLC2000 covering Anhui province (blue line). Cultivated areas are shown in white, forests and shrubs in green and brown. The right map represents an extract of the cropland AFI GlCropV2. The patterns are strongly similar for both maps .

2.1.3. Meteorological data

Besides the remote sensing variables that objectively reflected the vegetation growth status, other natural factors, especially the meteorological factors impact in a significant way the crop growth. In this study, the variables rainfall, mean temperature and the solar radiation were taken into consideration. The decadal values of these variables were generated by the Level 1 of the CGMS model investigated in the Work-Package 2. The spatial interpolation was performed on a grid of 25*25 Km.

2.1.4. Technological variables

As shown on Figure 3, the wheat yield has demonstrated a clear upward trend during the last decade. One of the main drivers for this so-called technological trend is the use of chemical fertilisers. In order to include this important driver for yield increase, the statistical data for use of chemical fertilisers were also retrieved from the National statistical Bureau (Fig. 5).

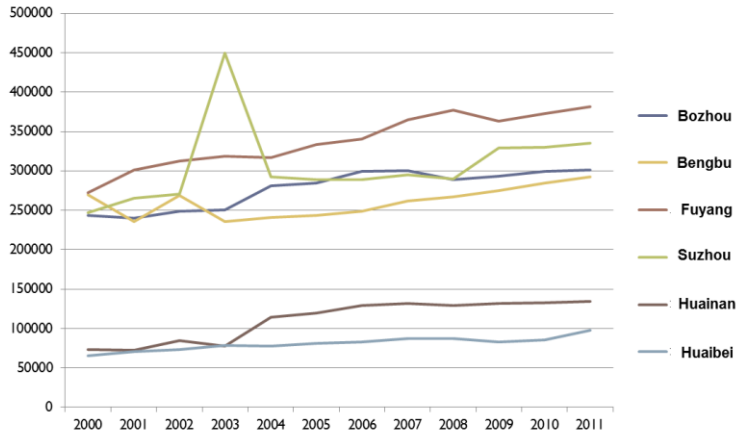

Figure 5: Annual chemical fertiliser input (ton) in six prefectures from 2000 till 2011.

2.2. Pre-processing

Smoothing for NDVI and fAPAR images: in this research, NDVI and fAPAR are pre-processed with smoothing algorithms, in order to further ease the aberrant pixel values caused by cloud or other missing values. A modified version of the Swets-algorithm was used, which inspects each pixel's time profile, detects all the cloudy observations and interpolates with values that are more appropriate. In these cases, DMP is derived from the smoothed fAPAR together with meteorological input.

Regional Unmixed Means (RUM): RUM is the mean value of a biophysical variable (for example NDVI) contributed by a class of vegetation (for example wheat), within a certain administrative unit (for example province), in course of certain period (for example first

dekad of May). These data are compiled in format of a database and can be charted in graphs (fig.3).

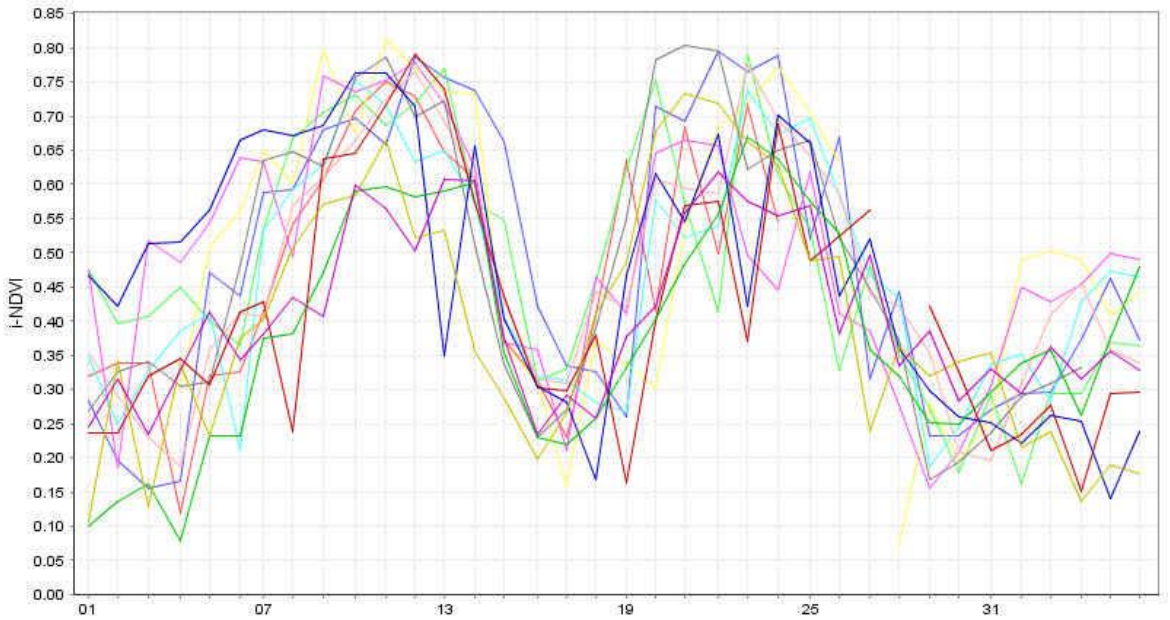

Figure 6: RUM-profiles of the non-smoothed i-NDVI for one of the counties in the HuaiBei zone. The different lines represent the subsequent years (1999-2011).

3. Results

3.1. Pre-processing and RUM database construction

The ancillary data within the Anhui province administrative boundary are subset from: GLC2000, GTOPO30 and the GlCropV2 global dataset. The SHAPE-file with the boundaries of the six districts was converted to raster format. In the resulting raster image, each pixel contains the REG_ID number (1-6) of the prefecture to which it belongs. See Fig. 4.

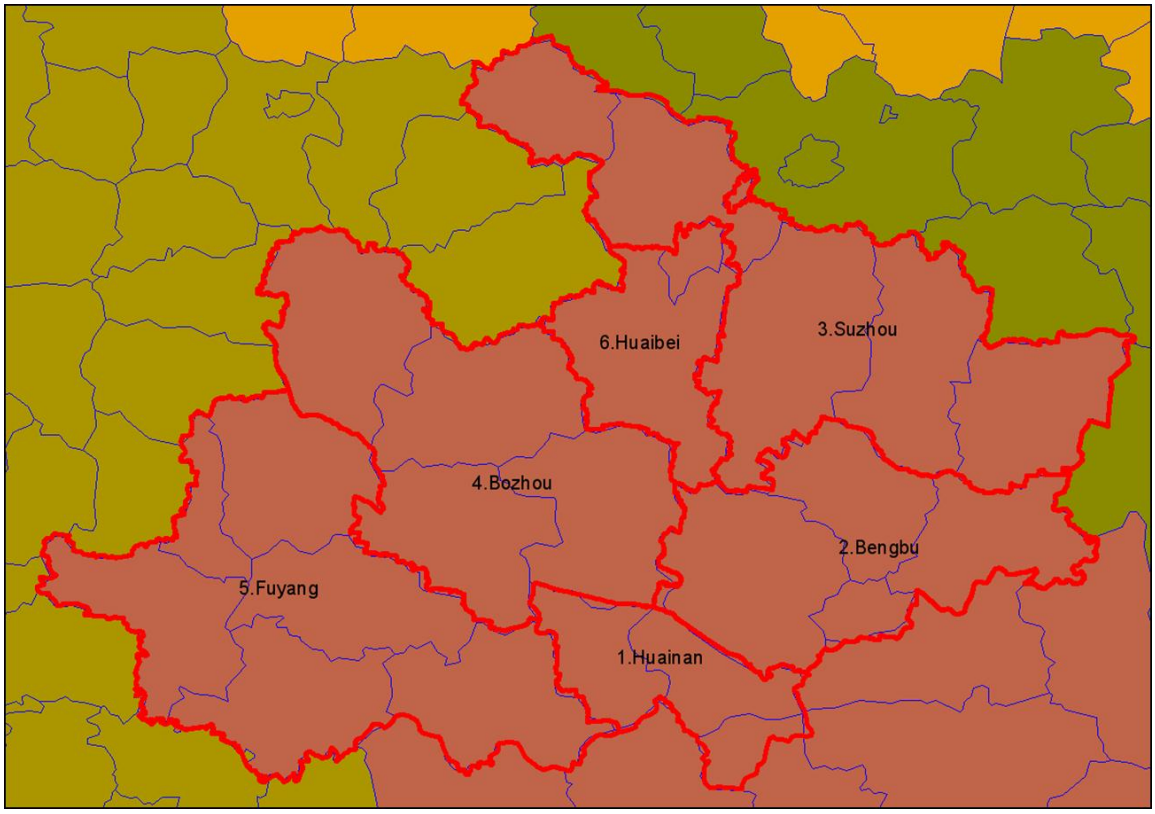

Figure 7: The Huaibei Plain, one of the study areas of E-AGRI project, is located the northern part of Anhui province. It encloses six prefectures (the administrative level between province and county).

Original non-smoothed NDVI and fAPAR times series(named as i-NDVI and a-fAPAR) were extracted and pre-proceeded (smoothed) using Swets-algorythm. The smoothed times series were named respectively k-NDVI and b-fAPAR (figure 5). The biophysical variable y-DMP images were then computed using b-fAPAR and the ECWMF daily global meteorological information (MARSOP-project). The whole imagery covers the full period 1999-2011 (extensible every year with 36 dekads per year) was generated and extended every year.

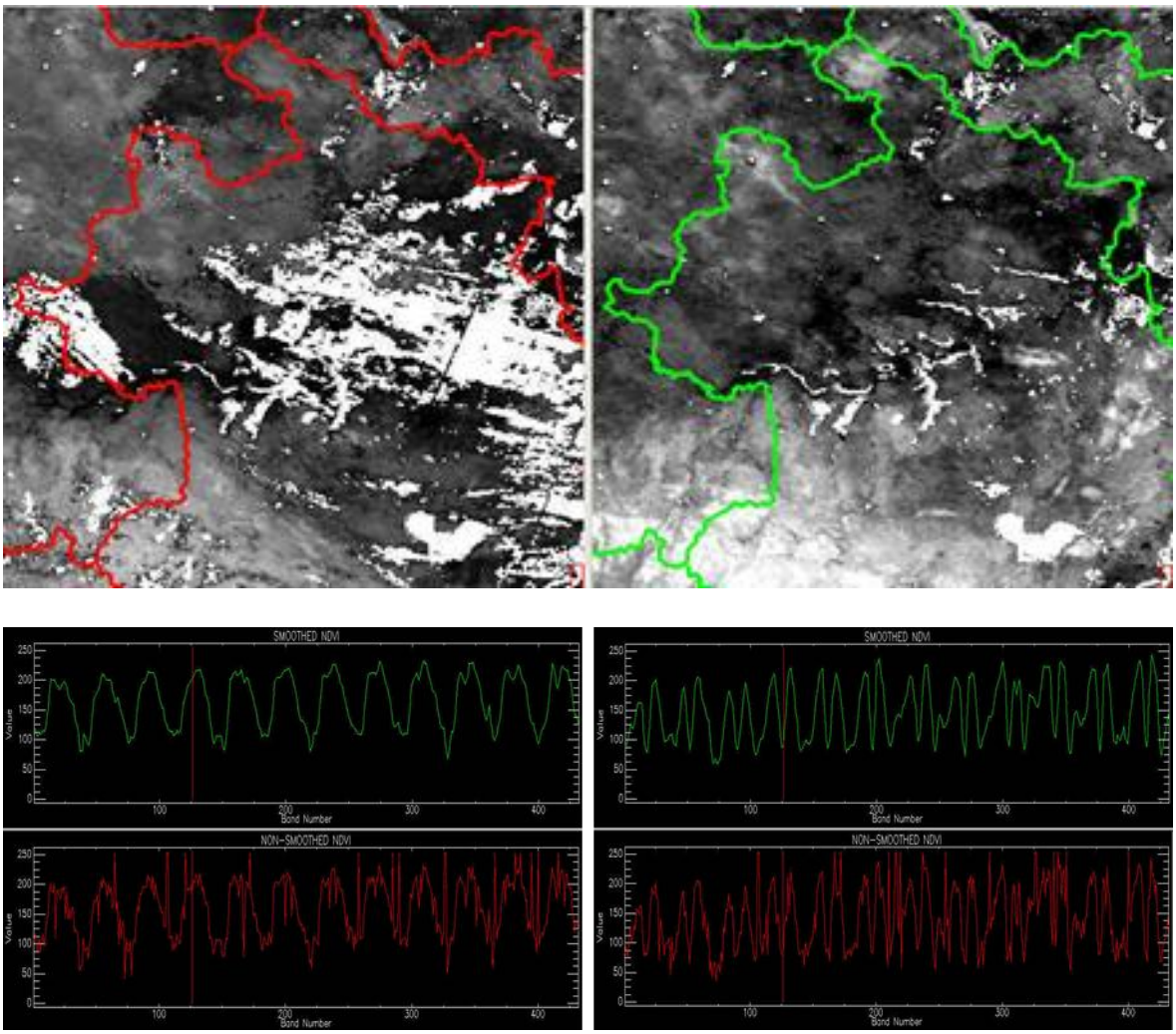

Figure 8: Effect of the Swets-smoothing for the variable NDVI in the region of Huaibei. The upper frames are the images of NDVI-S10 for the dekad 18 in 2002. The non-smoothed i-NDVI (left) contains lots of missing values and non- detected clouds. *The smoothed k-NDVI (right) filled much of the missing values. The effect of* smoothing can also be showed with a multi-annual profile for a forest pixel (lower-left) and a cropland pixel (lower-right). The red line represents non-smoothed NDVI, the green line records smoothed NDVI version.

Finally, databases with RUM-values (Regional Unmixed Means) were constructed for the biophysical variables i-NDVI, a-FAPAR, k-NDVI, b-FAPAR and y-DMP (see table 1 – hence not for p-DMP), according the rules:

- The mean values were taken for all “cropland pixels” in each of the six districts.
- All pixels having an area fraction of 100% in the cropland AFI were considered as “cropland” (Figure 2).

With a dedicated program, the initial RUM-databases were then converted to three simplified formats for easier handeling: X1.csv, X2.csv, X3.csv, with X=Huaibei, all in "comma-separated values" (CSV):

- Format 1 (Fig. 9):
  - Five tables, for each of five biophysical variables (i, a, k, b, y).
  - Per table: one line per region. It starts with Region_ID and then contains the values for all subsequent dekads, in all years.
- Format 2 (Fig.10):
  - Five tables, one for each variable (i, a, k, b, y).
  - Per table: one line per region & year.
  - Each line starts with the concerned Region_ID and year. Then follow the values for all the 36 dekads.
- Format 3 (Fig. 11):
  - One single table.
  - Separate lines for each combination of region & year & dekad
  - Each line starts with Region_ID, year and dekad-number. Then follow the corresponding values of the five target variables: i-NDVI, a-FAPAR, k-NDVI, b-FAPAR, y-DMP.

i-NDVI

,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,
Reg,	1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12,
1,	0.281,	0.325,	0.221,	0.344,	0.453,	0.342,	0.336,	0.353,	0.401,	0.543,	0.545,	0.498,
2,	0.244,	0.298,	0.220,	0.306,	0.441,	0.319,	0.282,	0.383,	0.386,	0.464,	0.500,	0.477,
3,	0.287,	0.345,	0.265,	0.359,	0.439,	0.347,	0.338,	0.378,	0.343,	0.495,	0.552,	0.542,
4,	0.235,	0.311,	0.226,	0.322,	0.402,	0.322,	0.323,	0.406,	0.376,	0.538,	0.517,	0.511,
5,	0.240,	0.310,	0.231,	0.328,	0.407,	0.342,	0.376,	0.425,	0.392,	0.583,	0.551,	0.525,
6,	0.304,	0.371,	0.258,	0.398,	0.473,	0.390,	0.386,	0.444,	0.432,	0.554,	0.615,	0.579,

a-FAPAR

,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,
Reg,	1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12,
1,	0.188,	0.238,	0.133,	0.225,	0.323,	0.262,	0.264,	0.251,	0.318,	0.475,	0.494,	0.430,
2,	0.137,	0.192,	0.116,	0.168,	0.290,	0.230,	0.175,	0.293,	0.303,	0.371,	0.443,	0.414,
3,	0.197,	0.264,	0.182,	0.227,	0.301,	0.282,	0.215,	0.296,	0.262,	0.390,	0.471,	0.486,
4,	0.127,	0.212,	0.126,	0.195,	0.260,	0.245,	0.240,	0.318,	0.304,	0.469,	0.469,	0.457,
5,	0.128,	0.211,	0.129,	0.210,	0.294,	0.264,	0.313,	0.339,	0.330,	0.519,	0.511,	0.462,
6,	0.219,	0.294,	0.176,	0.276,	0.342,	0.339,	0.289,	0.375,	0.379,	0.492,	0.552,	0.529,

k-NDVI

,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,
Reg,	1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12,
1,	0.302,	0.328,	0.357,	0.394,	0.453,	0.395,	0.383,	0.396,	0.449,	0.544,	0.562,	0.560,
2,	0.270,	0.299,	0.330,	0.367,	0.441,	0.386,	0.378,	0.392,	0.421,	0.473,	0.509,	0.527,
3,	0.320,	0.346,	0.370,	0.400,	0.441,	0.395,	0.381,	0.397,	0.431,	0.505,	0.557,	0.576,
4,	0.274,	0.311,	0.330,	0.359,	0.402,	0.369,	0.371,	0.414,	0.461,	0.538,	0.559,	0.574,
5,	0.275,	0.310,	0.334,	0.364,	0.407,	0.387,	0.400,	0.445,	0.500,	0.583,	0.599,	0.606,
6,	0.342,	0.373,	0.399,	0.433,	0.474,	0.439,	0.429,	0.454,	0.496,	0.567,	0.618,	0.616,

b-FAPAR

,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,
Reg,	1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12,
1,	0.218,	0.239,	0.252,	0.276,	0.323,	0.292,	0.293,	0.318,	0.378,	0.478,	0.505,	0.486,
2,	0.174,	0.193,	0.204,	0.226,	0.290,	0.269,	0.279,	0.305,	0.345,	0.400,	0.448,	0.453,
3,	0.248,	0.264,	0.258,	0.271,	0.306,	0.297,	0.293,	0.313,	0.344,	0.409,	0.477,	0.506,
4,	0.179,	0.212,	0.212,	0.228,	0.261,	0.263,	0.281,	0.328,	0.383,	0.469,	0.493,	0.502,
5,	0.177,	0.212,	0.224,	0.248,	0.295,	0.294,	0.324,	0.371,	0.429,	0.519,	0.537,	0.536,
6,	0.271,	0.294,	0.293,	0.309,	0.344,	0.349,	0.356,	0.387,	0.432,	0.503,	0.555,	0.552,

y-DMP

,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,	1999,
Reg,	1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12,
1,	11.3,	11.4,	17.5,	21.1,	28.7,	23.5,	31.4,	31.3,	39.4,	91.2,	91.4,	97.8,
2,	8.5,	8.9,	13.8,	16.7,	24.9,	20.4,	29.6,	30.9,	36.0,	75.9,	83.3,	94.4,
3,	11.1,	11.3,	16.9,	19.1,	25.3,	22.6,	31.5,	30.9,	37.8,	77.7,	95.1,	110.2,
4,	8.7,	9.6,	14.8,	17.2,	23.3,	21.9,	31.1,	30.6,	42.6,	90.4,	95.9,	104.8,
5,	9.4,	9.9,	16.5,	19.7,	27.2,	26.2,	35.3,	33.5,	47.4,	99.5,	100.4,	106.9,
6,	12.5,	12.9,	19.8,	22.5,	29.6,	27.7,	39.2,	36.9,	48.2,	96.7,	110.3,	118.8,

Figure 9: Format 1 of the RUM database for five biophysical variables generated in the E-AGRI project for the Huaibei Plain.

```
i-NDVI
Reg,Year,      1,      2,      3,      4,      5,      6,      7,      8,      9,     10,     11,     12,
   1,1999, 0.281, 0.325, 0.221, 0.344, 0.453, 0.342, 0.336, 0.353, 0.401, 0.543, 0.545, 0.498,
   1,2000, 0.041, 0.168, 0.089, 0.145, 0.131, 0.210, 0.241, 0.352, 0.480, 0.536, 0.617, 0.480,
...
   1,2010, 0.217, 0.196, 0.144, 0.094, 0.404, 0.387, 0.516, 0.531, 0.655, 0.629, 0.549, 0.760,
   2,1999, 0.244, 0.298, 0.220, 0.306, 0.441, 0.319, 0.282, 0.383, 0.386, 0.464, 0.500, 0.477,
...
   6,2010, 0.358, 0.349, 0.386, 0.220, 0.636, 0.509, 0.608, 0.636, 0.730, 0.724, 0.695, 0.831,

a-FAPAR
Reg,Year,      1,      2,      3,      4,      5,      6,      7,      8,      9,     10,     11,     12,
   1,1999, 0.188, 0.238, 0.133, 0.225, 0.323, 0.262, 0.264, 0.251, 0.318, 0.475, 0.494, 0.430,
   1,2000, 0.011, 0.070, 0.012, 0.042, 0.040, 0.089, 0.119, 0.233, 0.372, 0.414, 0.545, 0.437,
         ...
   1,2010, 0.128, 0.103, 0.044, 0.020, 0.274, 0.281, 0.436, 0.459, 0.587, 0.578, 0.479, 0.658,
   2,1999, 0.137, 0.192, 0.116, 0.168, 0.290, 0.230, 0.175, 0.293, 0.303, 0.371, 0.443, 0.414,
...
   6,2010, 0.306, 0.305, 0.286, 0.127, 0.509, 0.453, 0.559, 0.596, 0.674, 0.677, 0.653, 0.742,

k-NDVI
Reg,Year,      1,      2,      3,      4,      5,      6,      7,      8,      9,     10,     11,     12,
   1,1999, 0.302, 0.328, 0.357, 0.394, 0.453, 0.395, 0.383, 0.396, 0.449, 0.544, 0.562, 0.560,
   1,2000, 0.196, 0.177, 0.155, 0.151, 0.156, 0.210, 0.265, 0.358, 0.482, 0.562, 0.617, 0.592,
...
   1,2010, 0.251, 0.233, 0.230, 0.300, 0.404, 0.452, 0.522, 0.580, 0.656, 0.684, 0.728, 0.761,
   2,1999, 0.270, 0.299, 0.330, 0.367, 0.441, 0.386, 0.378, 0.392, 0.421, 0.473, 0.509, 0.527,
...
   6,2010, 0.406, 0.395, 0.432, 0.520, 0.636, 0.634, 0.650, 0.667, 0.731, 0.757, 0.801, 0.833,

b-FAPAR
Reg,Year,      1,      2,      3,      4,      5,      6,      7,      8,      9,     10,     11,     12,
   1,1999, 0.218, 0.239, 0.252, 0.276, 0.323, 0.292, 0.293, 0.318, 0.378, 0.478, 0.505, 0.486,
   1,2000, 0.085, 0.074, 0.054, 0.047, 0.050, 0.090, 0.145, 0.245, 0.375, 0.465, 0.545, 0.497,
...
   1,2010, 0.172, 0.141, 0.119, 0.176, 0.274, 0.333, 0.438, 0.505, 0.588, 0.612, 0.646, 0.666,
   2,1999, 0.174, 0.193, 0.204, 0.226, 0.290, 0.269, 0.279, 0.305, 0.345, 0.400, 0.448, 0.453,
...
   6,2010, 0.368, 0.349, 0.349, 0.410, 0.509, 0.537, 0.576, 0.616, 0.675, 0.698, 0.731, 0.749,

y-DMP
Reg,Year,      1,      2,      3,      4,      5,      6,      7,      8,      9,     10,     11,     12,
   1,1999,  11.3,  11.4,  17.5,  21.1,  28.7,  23.5,  31.4,  31.3,  39.4,  91.2,  91.4,  97.8,
   1,2000,   2.6,   2.1,   1.4,   2.3,   2.7,   8.1,  15.8,  33.1,  75.6,  98.2, 107.4, 117.2,
...
   1,2010,   6.8,   7.8,   8.0,   6.3,  16.0,  38.4,  20.4,  84.0,  65.6, 109.8,  81.3, 141.8,
   2,1999,   8.5,   8.9,  13.8,  16.7,  24.9,  20.4,  29.6,  30.9,  36.0,  75.9,  83.3,  94.4,
...
   6,2010,  12.9,  17.9,  21.7,  16.7,  29.3,  58.9,  27.9,  95.6,  83.6, 131.6,  93.4, 154.6,
```

Figure 10: Format 2 of the RUM database for five biophysical variables generated in the E-AGRI project for the Huaibei Plain.

```
Reg,Year,Dk,      i,      a,      k,      b,      y
   1,1999, 1, 0.281, 0.188, 0.302, 0.218,   11.3
...
   1,1999,36, 0.218, 0.100, 0.219, 0.103,    5.5
   1,2000, 1, 0.041, 0.011, 0.196, 0.085,    2.6
...
   1,2010,36, 0.297, 0.207, 0.313, 0.212,   11.2
   2,1999, 1, 0.244, 0.137, 0.270, 0.174,    8.5
...
   6,2010,36, 0.477, 0.451, 0.521, 0.469,   22.5
```

Figure 11: Format 3 of the RUM database for five biophysical variables generated in the E-AGRI project for the Huaibei Plain.

These RUM-tables with mean values of the biophysical variables can be used to carry out crop yield assessment based on remote sensing.

3.2. Calculating the cumulative values and extending the RUM databases

3.2.1. Extended RUM databases

The RUM databases, which contain the biophysical and meteorological variables, were extended with the cumulative values of the five variables listed in Table 1. The cumulative values were calculated for every possible consecutive 2 to 9 dekads growth period of wheat (between the 10th October and 20th June the following year), as a phenological stage for winter wheat could last till 9 dekads.

3.2.2. Establishing datasets for linear regression

Three databases have been established. The database 1 includes all possible cumulative variables issued from NDVI (ΣNDVI): cumulative variables from 1 to 9 consecutive dekads, as well as the technology variable CFI the chemical fertilizer input of sowing year in each prefecture. The variable will be symbolized by the month and the number of dekad (o: October, n: November, d: December, j: January, f: February, m: March, a: April, y: May, u: June). For example, o2d3 represents the cumulative variable of NDVI from the 2nd dekad of October to the 3rd dekad of December. The database 1 contains 198 explanatory variables.

The database 2 is composed by the cumulative variables derived from the biophysical variable DMP and CPI as in the case of the database 1. It contains 198 explanatory variable in total.

The database 3 is based on the database 1, and further includes several key meteorological variables. These meteorological variables represent the mean values of these indicators during a phenological stage. The following symbols are used: R: rainfall, S: solar radiation, T: temperature. The phenological stages are represented by s: sowing, e: emergence, t: tiller, w: winter period, g: turning green, j: jointing, h: heading, m: maturity, v: harvest. The database 3 has 225 variables in total.

3.2.3. Step wise linear regression

The stepwise regression was implemented. The probability significance thresholds for entry and removal of candidate predictors in the model were set to α_e= 5% and α_r= 10%. The prediction accuracy of this method was evaluated using Leave-one-out cross validation test. It will leave one year data out, then using the rest of data to predict the left year; and define that year's error. When this procedure is repeated for all the years i (i = 1–n), an independent error estimate can be obtained by the following formula, where AE represents absolute error, $\hat{Y}_i$ predicted wheat yield, and Y_i the official wheat yield

$$AE = \frac{\sum_{i=1}^{n} \left| \hat{Y}_i - Y_i \right|}{n}$$

3.3. Yield forecasting models for six prefectures on the Huaibei Plain

Based on three above-mentioned databases, multiple linear regression models were established using stepwise approach for six prefectures on Huaibei Plain. The results are displayed in Tables 1 to 3. All regression models showed high significant R^2 (from 0.653 to 0.990).

Table 2: **The best-fit regression models for wheat yield prediction at the prefecture level on Huaibei Plain using NDVI and CFI as predictors**

Prefecture	Models			R^2	Absolute Error (ton)
	Constant	CFI	ΣNDVI		
Bengbu	-3.925		+5.694*O2N2+1.934*F2F3	0.804	0.299
Bozhou	-5.040	+0.031*CFI	+7.376*N1	0.851	0.291
Fuyang	-8.265	+0.029*CFI	+4.255*O2N1	0.800	0.270
Huaibei	-2.619	+0.068*CFI	+0.702*J2M3	0.765	0.337
Huainan	-0.422	+0.047*CFI		0.918	0.261
Suzhou	-1.913		+11.396*M3	0.653	0.396

Table 3: **The best-fit regression models for wheat yield prediction at the prefecture level on Huaibei Plain using DMP and CFI as predictors**

Prefecture	Models			R^2	Absolute Error (ton)
	Constant	CFI	ΣDMP		
Bengbu	2.439		+0.280*D3	0.736	0.388
Bozhou	0.468		+0.006*A1Y3+0.114*D3	0.941	0.197
Fuyang	0.782		+0.034*A3	0.786	0.325
Huaibei	-1.365		+0.008*A1Y3+0.016*M2	0.854	0.281
Huainan	-0.422	+0.047*CFI		0.918	0.261
Suzhou	-0.249		+0.008*M2Y1	0.700	0.359

Table 4: **The best-fit regression models for wheat yield prediction at the prefecture level on Huaibei Plain using NDVI, CFI and meteorological variables as predictors**

Prefecture	Models				R^2	Absolute Error (ton)
	Constant	ΣNDVI	CFI	Meteorology		
Bengbu	-3.875	+6.183*O2N2		-0.019*RHV+ 0.471*TJ-0.093*RW - 0.326* SW	0.990	0.062
Bozhou	-5.040	+7.376*N1	+0.031*CFI		0.851	0.291
Fuyang	-12.189	+3.374*O2N1	+0.029*CFI	+0.282*SS	0.907	0.183
Huaibei	-2.588	+0.730*J3M3	+0.071*CFI	-0.40*RJ	0.963	0.283
Huainan	2.691		+0.050*CFI	-0.053*RJ-0.135*SH	0.964	0.167
Suzhou	-2.623	+12.762*M3		-0.065RJ+0.055*RTG	0.936	0.213

4. Discussions

4.1. Accuracy of prediction

The absolute errors resulted from these regressions reflect the accuracy of models. Table 5 compares the absolute errors of the predictions for the each prefecture and using different sets of predictors. The prediction models based on the database 3, thus the including biophysical, technological and meteorological variables could reduce the absolute error of prediction to 0.2 ton ha^{-1}. Yield estimates with such level of error (below 0.2 ton ha^{-1}) are considered to be very good by European standard.

Table 5: **comparison of the accuracy of prediction in terms of the absolute errors (tons) using different sets of predictors**

Prefecture	k-NDVI &CFI	y-DMP &CFI	k-NDVI & CFI& Meteo
Bengbu	0.299	0.388	0.062
Bozhou	0.291	0.197	0.291
Fuyang	0.270	0.325	0.183
Huaibei	0.337	0.281	0.283
Huainan	0.261	0.261	0.167
Suzhou	0.396	0.359	0.213
Average	0.309	0.302	0.200

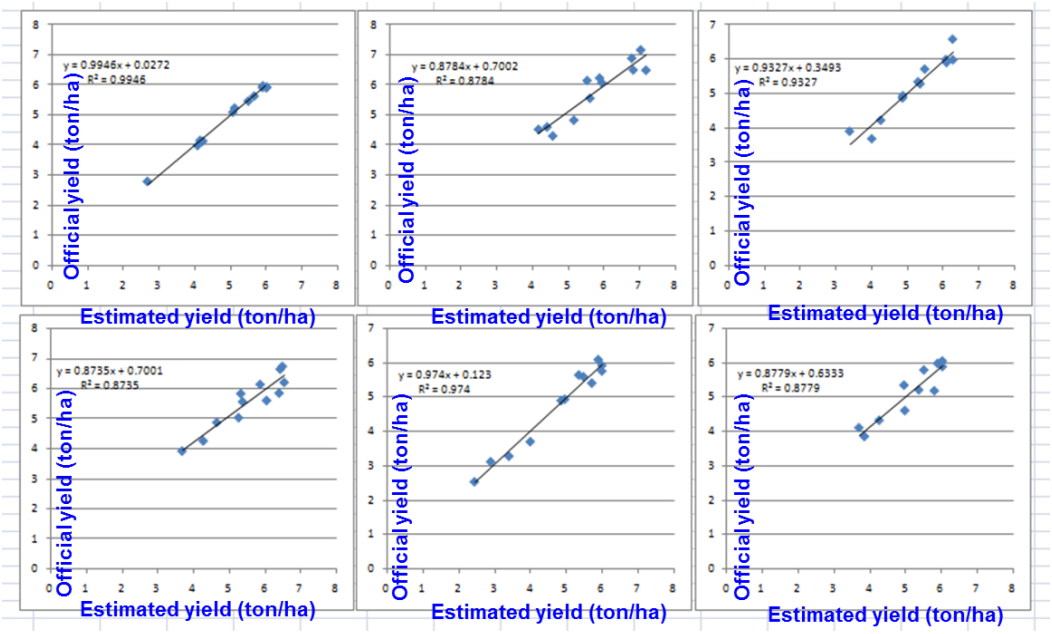

Figure 12: Correlation between the official and predicted yield using the set of predictors in the database 3.

Fig. 12 shows the true versus predicted yield scatter-grams for each prefecture by using the predictors in the database 3. The correlation between the official yield and predicted yield turns to be very high, with range of R^2 between 0.878 and 0.995.

4.2. The relevance of the predictors

From the wide range of candidate predictors in three databases, the stepwise regression retained the most significant explanatory variables. It turns out that the biophysical variables including ΣNDVI or ΣDMP are among the most relevant ones, especially in the phenological stages of emergence-tiller and jointing-heading. Among 18 cumulative biophysical variables appeared in the models listed above, six were emergence-tiller phase indicators, and nine of them were variables describing jointing-heading phase. The scientists have demonstrated the importance of emergence rate and tiller density on the ear number of crop, thus its final yield. The jointing-heading stage is another very important phase in crop growth, therefore is often be involved in yield prediction. It has been reported that the growth condition of winter wheat in jointing-heading stage contain more yield information than any other periods of wheat growth on the North China Plain. The nutrition accumulation during this growth period is considered as critical for the final crop yield.

The meteorological variables have their importance as well. The parameters of rainfall appear to be most relevant. However, in contrast with other areas on the north China Plain, especially the northern part of the plain, the variables of rainfall correlate negatively with the crop yield. This can be explained by the relative abundance of rainfall in the winter season and a higher irrigation rate (71% of sowing area) on the Plain. An excessive rainfall leads often to waterlogging and subsequent crop failure.

The technological drivers or so call technological trends, which cannot be explained by the natural or environmental factors, determine also the crop yield from one year to another. Figure 13 showed the extended winter wheat yield statistics since the year 1985.

As elsewhere in the country, the yield increased from 3.01 ton per ha in 1985 to 6.31 ton per ha in 2011 in Huaibei Plain. At the national level, for an earlier period, researchers found that the observed national wheat yield increased dramatically from 1845 kg ha^{-1} in 1978 to 3542 kg ha^{-1} in 1995 with a linear trend. The reason of this upwards trend is often be considered as consequences of technological and institutional progress including drastic increase of fertiliser input, better breeding technology and a surge of irrigation practice. As no irrigation practice or breeding improvement have been quantified and registered as statistical records, we assimilated the technological progress by the variable of chemical fertilisers input which is recorded by the local administrations.

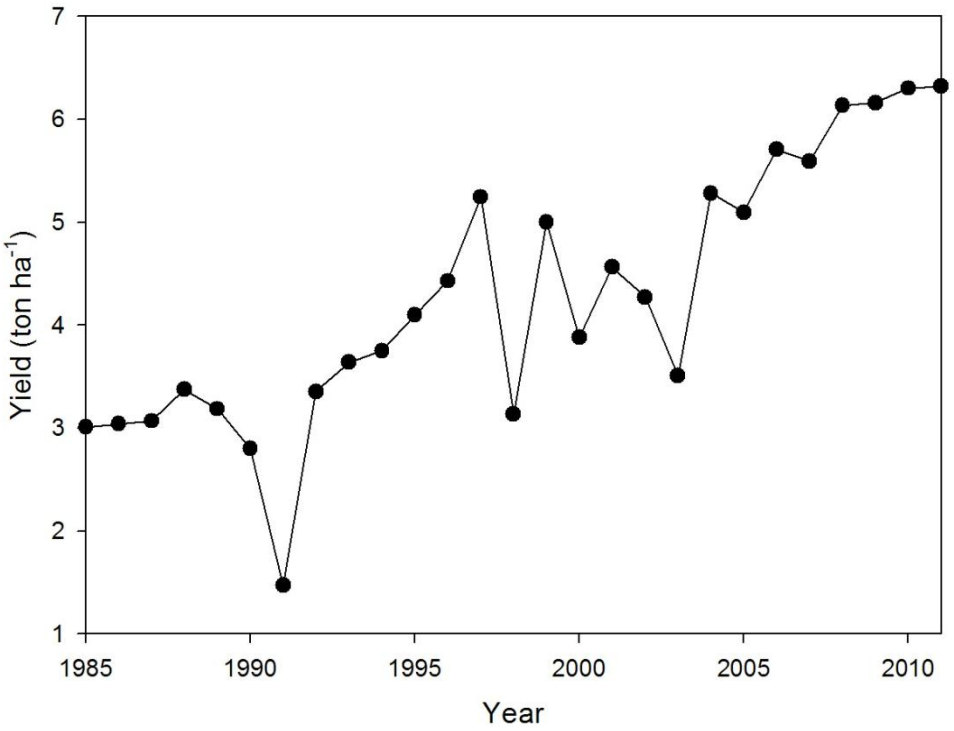

Figure 13: Average wheat yield on the Huaibei Plain (6 prefectures) from 1985 to 2010.

4.3. Early prediction on the Huaibei Plain

For the use operational use, we built and tested models, which are relevant for yield estimation at an early stage of crop growth season. For this test purposes, a database containing the mean values of all relevant predictors was built for whole Plain. Table 6 displays the potential regression models for early season yield forecasting with different prediction periods, from autumn to nearly harvest.

The cumulative variable NDVI in the emergence-tiller phenological stage and the variable CFI (Chemical Fertiliser Input) are revealed most crucial. It is understandable that the absolute errors of forecasting increase while the forecasts were conducted in earlier stages of growth.

In conclusion, for the whole Huaibei Plain the cumulative value of NDVI from second dekad of October to the second dekad of November is the most explanatory variable for wheat yield forecasting on the Huaibei Plain.

Table 6: **different scenarios for early yield forecasting**

Prediction periods	Models				R^2	Absolute Error
	Constant	ΣNDVI	CFI	Meteorology		
May 3rd	-6.879	+2.865*O2N2	0.314*CFI	+0.022*Sm	0.965	0.129
March 1st	-5.996	+2.526*O2N2	+0.355*CFI	-0.22*Rg	0.960	0.159
February 1st	-6.327	+3.683*O2N2	+0.299*CFI	-0.017*Tw	0.901	0.185
November 2nd	-5.959	+3.578*O2N2	+0.302*CFI		0.927	0.238