Modelling soybean yield for the early prediction in the Russian Far East using remote sensing data

A S Stepanov1* and K N Dubrovin2

1 Far Eastern Agriculture Research Institute, Vostochnoe, 680521 Khabarovsk, Russia
2 Computing Center Far Eastern Branch of the Russian Academy of Sciences, 680000 Khabarovsk, Russia

*E-mail: stepanxx@mail.ru

Abstract. The paper presents an assessment of the model for predicting soybean yield at the level of municipalities in the Far East for the Oktyabrskiy and Leninskiy districts of the Jewish Autonomous Region, as well as the Khabarovsk and Vyazemskiy districts of Khabarovsk Territory. The share of soybean in the total arable land structure of these municipalities in 2018 ranged from 58% to 97%. According to 2010–2018 data, regression models were constructed for each region. The model used statistical data on district soybean yield, as well as data from remote sensing of the Earth. The values of the maximum NDVI (Normalized Difference Vegetation Index) of arable land and the growing duration at the week that reached maximum NDVI were used as independent variables in the regression model. We used weekly NDVI composites obtained for delineated arable lands through the Vega-Science system. According to long-term observations, it was found that in the study area the maximum NDVI was reached in weeks 30–33 (end of July to mid-August). The RMSE for different regions ranged from 0.06 to 0.15 t/ha, and the MAPE did not exceed 10%. The developed model can be used for predicting soybean yield and planning export operations by farms and territorial authorities.

1. Introduction
Recently, the use of Earth remote sensing data for modelling the vegetation cycle of various crops has become widespread [1,2]. Moreover, the values of vegetation indices are closely related to crop yields [3]. The most common characteristic of the vegetation process is the dynamics of NDVI (Normalized Difference Vegetation Index); the annual NDVI maximum is usually used to predict yield [4]. In the latest studies, regression models with vegetation indices and climatic characteristics of the area as independent predictors have been proposed [5,6]. For example, the MARS-Crop Yield Forecasting System (M-CYFS) has been used since 1993 to forecast the yields of all major crops in the European Union (EU). The system uses the minimum temperatures, maximum temperatures, precipitations and global radiation interpolated at a resolution of 25 km as predictors [7].

The main difficulty in using remote sensing data to predict crops at the level of territorial entities is the need to create a mask for each crop for the study area. This task is quite difficult. Earlier, we proposed the use of arable land masks to predict the leading regional crop at the municipality level [8]. This approach has been tested for the Khabarovsk District, where soybean is a dominant crop. At the same time, soybean is the most important crop of the Russian Far East, and one of the most important for Russia. Thus, one of the main objectives of this study is to test the proposed model, using remote
sensing data of the Earth and meteorological characteristics, for the neighbouring districts in the south of the Khabarovsk Territory and the JAR (Jewish Autonomous Region).

To solve this problem, it is necessary: to assess the share of soybean in arable land of the studied municipalities and outline the proposed boundaries of the studied land; to study the dynamics of NDVI and the climatic characteristics of the territories in the study period; assess the quality of models and the possibility of their application.

2. Materials and methods
The study area included two districts in the south of Khabarovsk Territory (Khabarovsk and Vyazemskiy District) and two districts in Jewish Autonomous Region (Oktyabrskiy and Leninskiy District). The main industry of these municipalities is agriculture, especially soybeans, grains and vegetables. Khabarovsk District and Vyazemskiy District are in the top three municipalities in the soybean-sown area. The Leninskiy District and Oktyabrskiy District are the leading soybean producing municipalities in the Jewish Autonomous Region.

The Khabarovsk Municipal District is located in the Middle Amur Region and consists of two parts: on the left bank of Amur river and the right bank around the city of Khabarovsk. Most of the sown area is located in the southern right-bank part of the municipality between 48° 20‘ – 48° 50’ N and 134° 10‘ – 135° 40’ E (figure 1). The Vyazemskiy District is located in the south of the Khabarovsk Territory. Most of the arable lands are located along the Ussuri River Valley to the west of the municipality between 47° 15‘ – 47° 45’ N and 134° 30‘ – 135° 15’ E (figure 1).

The average annual temperature in these two districts in July is near 21°C, and the average growing duration is 142–144 days. The amount of precipitation during the growing season is 614 mm in the Khabarovsk District and 557 mm in Vyazemskiy District (less than in Khabarovsk District; the highest rainfall intensity is observed at the end of summer). Plenty of moisture, sunlight, and alluvial and meadow soils allow for cultivating a wide range of crops, including soybean.

Figure 1. Study area: Khabarovsk District and Vyazemskiy District.
Figure 1 shows the contours of arable land in the Khabarovsk Territory, where we calculated vegetation indices. For example, the lands of the Big Ussuri Island were not included in the model. This is due to the lack of significant areas of soybean in this territory. At the same time, in the Vyazemsky District, we allocated two areas of arable land. This was done to avoid the distortion of NDVI values due to the influence of meadows with forage grasses.

Leninskiy District is located in the Middle Amur Lowland between 47° 40′ – 48° 25′ N and 131° 40′–133° 00′ E (figure 2). Meadow and alluvial soils, warm summers (average temperature in July is above 21°C), and a less monsoon climate (compared to the southern districts of Khabarovsk District) allow growing soybean throughout the district. Therefore, the Leninskiy District is the main soybean-producing region in the eastern part of the Russian Far East (Khabarovsk, Primorskiy territories and the Jewish Autonomous Region).

The Oktyarskiy District is located in the south-west of the Jewish Autonomous Region and has borders with Leninskiy District in the East. The sown area is located in the south of District between 47° 40′ – 48° 05′ N and 130° 50′ – 131° 45′ E (figure 2). Climatic conditions (fairly warm summers and an optimal amount of precipitation) in the agriculture area are suitable for soybean producing. Figure 2 also shows the contours of the arable land for studied municipalities of JAR.

Field average NDVI calculations were performed using weekly field composite values (in a CSV file) from the VEGA-Science web service (http://sci-vega.ru/). Climatic characteristics of the studied areas (total active temperatures (sum of temperatures), precipitation and photosynthetically active radiation (PAR)) during the growing season were obtained using reanalysis data. Table 1 presents the values of climatic characteristics, NDVI maxima, and growing duration (days with average daily temperature above 10 °C) for the four studied regions.

As can be seen from table 1, the average total active temperature is in the range 2467–2558 °C, the minimum value of 2217 °C was noted in the Oktyabrskiy district, and the maximum (2759 °C) in Leninskiy District. The average amount of precipitation ranged from 474 mm in the Oktyabrskiy District to 614 mm in the Khabarovsk District. The PAR for all municipalities was approximately at the same level (from 1.34 to 1.39 GJ · m²). The number of days favourable for crop growth (with T >
10 °C) ranged from 140 to 144. In general, the climatic conditions of the four regions were quite similar. At the same time, the NDVI values of district arable lands amounted to 0.82–0.83 for JAR municipalities and the Vyazemskiy District, and only 0.78 for the Khabarovsk District. As was established in previous studies, the structure of crops in the arable fields of the region affects the value of the maximum NDVI and the week of reaching the maximum [8].

Table 1. NDVI and meteorological characteristics (2010-2018).

| District       | NDVI max | Total active temperature, °C | Amount of precipitation, mm | Growing, days | PAR, 10\(^9\) GJ-m\(^2\) |
|----------------|----------|-----------------------------|-----------------------------|---------------|----------------------|
| Khabarovsk     | \(\bar{x}\) | 0.78 | 2532 | 614 | 142.0 | 1.34 |
|                | \(\Delta \bar{x}\) | 0.01 | 99 | 69 | 6.1 | 0.07 |
|                | min | 0.76 | 2353 | 424 | 133.0 | 1.22 |
|                | max | 0.81 | 2751 | 720 | 157.0 | 1.47 |
| Vyazemskiy     | \(\bar{x}\) | 0.83 | 2558 | 557 | 143.9 | 1.39 |
|                | \(\Delta \bar{x}\) | 0.01 | 85 | 48 | 5.4 | 0.07 |
|                | min | 0.81 | 2343 | 475 | 134.0 | 1.27 |
|                | max | 0.85 | 2713 | 657 | 155.0 | 1.54 |
| Oktyabrskiy    | \(\bar{x}\) | 0.82 | 2467 | 474 | 139.6 | 1.34 |
|                | \(\Delta \bar{x}\) | 0.01 | 95 | 82 | 5.1 | 0.09 |
|                | min | 0.80 | 2217 | 318 | 128.0 | 1.21 |
|                | max | 0.84 | 2605 | 652 | 146.0 | 1.45 |
| Leninskiy      | \(\bar{x}\) | 0.83 | 2580 | 519 | 144.4 | 1.39 |
|                | \(\Delta \bar{x}\) | 0.02 | 96 | 73 | 5.2 | 0.06 |
|                | min | 0.80 | 2374 | 378 | 136.0 | 1.27 |
|                | max | 0.87 | 2759 | 659 | 153.0 | 1.50 |

According to table 2, the total area of arable land in the Leninskiy and Oktyabrskiy districts of the JAR is 80,892 ha and 48,893 ha, and the share of soybeans exceeded 95% in 2018. In the Vyazemskiy District, this indicator amounted to 63%, and in the Khabarovsk District it was 58%. Accordingly, a shift in the maximum soybean in the municipalities of the Khabarovsk Territory to an earlier date was expected (the influence of early crops).

Table 2. The share of soybean in district arable land (2018).

| District       | Total sown area, ha | Soybean sown area, ha | Share of soybean, % |
|----------------|---------------------|-----------------------|---------------------|
| Khabarovsk     | 28,222              | 16,460                | 58.3                |
| Vyazemskiy     | 17,268              | 10,859                | 62.9                |
| Oktyabrskiy    | 48,893              | 46,456                | 95.0                |
| Leninskiy      | 80,982              | 78,971                | 97.5                |

Figure 3 shows the dynamics of the weekly NDVI arable land composites of the four studied areas in 2018. The NDVI seasonal progress graph has a characteristic shape corresponding to soybean [9]. The maximum NDVI in the Khabarovsk District in 2010-2018 was usually reached at the 30th calendar week and in the Vyazemskiy district at the 31st week. In the Oktyabrskiy and Leninsky districts of the EAO, according to the 2012–2018 data, the maximum corresponded to the calendar weeks 32–33.

A linear regression model was developed to assess the soybean yield. The soybean yield was chosen as a dependent variable, while the remote sensing data and meteorological characteristics were chosen as independent variables. We considered one dependent and six independent variables:

- \(y\) – average annual soybean yield estimation by municipality, t/ha;
- \(x_1\) – the maximum NDVI value by the mask of the municipality’s arable land;
- \(x_2\) – growing duration.

The multivariate regression model was constructed as follows:
\[ y = b + a_1 x_1 + a_2 x_2. \]

To evaluate the accuracy of the predictions, we used the coefficient of determination \( R^2 \), RMSE and MAPE between modelled and observed data. The MAPE was calculated to estimate model accuracy according to data for the observation period, expressed as a percentage, as follows:

\[
APE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_{i}^{\text{pred}} - y_{i}^{\text{obs}}}{y_{i}^{\text{obs}}} \right| \times 100,
\]

where:
- \( n \) – observation period duration (years);
- \( y_{i}^{\text{pred}} \) – predicted yield in the \( i \)th year;
- \( y_{i}^{\text{obs}} \) – observed yield in the \( i \)th year.

The RMSE (root-mean-square error) indicator was used to model accuracy estimation as follows:

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{i}^{\text{obs}} - y_{i}^{\text{pred}})^2}.
\]

**Figure 3.** Weekly Normalized Difference Vegetation Index (NDVI) composites (calendar weeks 15–42) in 2018 for: (a) Khabarovsk District, (b) Vyazemskiy District, (c) Oktyabrskiy District and (d) Leninskiy District.
3. Results
Table 3 presents the coefficients of the regression equations for the proposed model. The $R^2$ values for the Khabarovsk, Vyazemskiy and Leninskiy districts were 0.45–0.49, and for the Oktyabrskiy District $R^2 = 0.62$.

Table 3. Regression coefficients for studied districts (2010–2018)\(^a\).

| District   | $x_1$ | $x_2$ | b   | $R^2$ |
|-----------|-------|-------|-----|-------|
| Khabarovsk| 8.22  | 0.026 | -7.31 | 0.48  |
| Vyazemskiy| 2.33  | 0.030 | -3.30 | 0.45  |
| Oktyabrskiy| 1.94  | 0.017 | -2.23 | 0.62  |
| Leninskiy | 3.5314| 0.015 | -3.41 | 0.49  |

\(^a\) 2012-2018 for Leninskiy, Oktyabrskiy

Figure 4 presents the actual and model yield in 2010–2018 for the districts of the Khabarovsk Territory and 2012–2018 for the districts of the JAR.

Figure 4. Actual yield and model estimation for: a) Khabarovsk District, b) Vyazemskiy District, c) Oktyabrskiy District and d) Leninskiy District.
In general, the soybean yield in the Khabarovsk Territory is higher than in the JAR. At the same time, the quality of the model in some areas was influenced to some extent by the data on soybean yield during the catastrophic flood on the Amur in 2013, as well as its consequences in 2014. The average district yield varied greatly depending on the method of accounting for flood-affected fields. In addition, the deviation of the climatic characteristics of the August-September of individual years from the average annual norm can influence the quality of the model. At the same time, the methodological approach used involves the early prediction of yield, before the end of the soybean vegetation cycle. Therefore, the use of climatic characteristics of the full vegetation cycle in the model would increase the accuracy; however, this model would be of limited practical use.

In general, regression models showed high accuracy. The RMSE ranged from 0.06 t/ha for the Oktyabrskiy District to 0.15 t/ha for the Khabarovsk District (table 4). The MAPE for the Khabarovsk District was 9.7%, and for the Vyazemskiy District, it was 8.8%. The highest accuracy was observed for the Oktyabrskiy district of the JAR at 5.2%.

Table 4. Proposed model accuracy.

| District       | RMSE, t/ha | MAPE, % |
|----------------|------------|---------|
| Khabarovsk     | 0.15       | 9.7     |
| Vyazemskiy     | 0.14       | 8.8     |
| Oktyabrskiy    | 0.06       | 5.2     |
| Leninskiy      | 0.10       | 7.6     |

4. Conclusions
A regression model for estimating soybean productivity was built for four soybean-producing municipalities of the Far East, where the values of the maximum NDVI of arable land and the growing duration up to a week corresponding to the NDVI maximum were included in the model as independent variables. It was found that the maximum NDVI values were reached in weeks 30–31 in the Khabarovsk Territory, and weeks 32–33 in the JAR. An assessment of the accuracy of the model according to the 2010–2018 showed good results; the MAPE model was in the range 5.2–9.7%, and the RMSE ranged from 0.06 to 0.15 t/ha. The developed model can be used in practice for predicting soybean yield.

References
[1] Sun L et al. 2017 Remote Sens. RS9(2017)317
[2] Wehlage D C, Gamon J A, Thayer D and Hildebrand D V 2016 Remote Sens. RS8(2016)872
[3] Panek E and Gozdowski D 2020 Remote Sensing Applications: Society and Environment 17 100286
[4] Hassan M A, Yang M, Rasheed A, Yang G, Reynolds M, Xia X, Xiao Y and He Z 2019 Plant Science 282 95–103
[5] Nagy A, Feher J and Tamas J 2018 Computers and Electronics in Agriculture 151 41–49
[6] Balaghi R, Tychon B, Eerens H and Jlibene M 2008 International Journal of Applied Earth Observation and Geoinformation 10(4) 438–452
[7] Cao J, Zhang Z, Tao F, Zhang L, Luo Y, Han J and Li Z 2020 Remote Sens. RS12(2020)750
[8] Lecerf R, Ceglar A, Lopez–Lozano R, van der Velde M and Baruth B 2019 Agricultural Systems 168 191–202
[9] Stepanov A, Dubrovin K, Sorokin A and Aseeva T 2020 Remote Sens. RS12(2020)1936
[10] Novorotskii P 2007 Russ. Meteorol. Hydrol. 32 102–109