Application of the decision tree method for predicting the yield of spring wheat

V K Kalichkin¹, O K Alsova² and K Yu Maksimovich¹,³

¹ Siberian Federal Scientific Centre of Agro-BioTechnologies of the Russian Academy of Sciences, P.O. box 267, Krasnoobsk, 630501, Russia
² Novosibirsk State Technical University, Novosibirsk, 20 Prospekt K. Marksa, 630073, Russia
³ E-mail: kiri-maksimovi@mail.ru

Abstract. The results of the development of predictive models of the yield of spring wheat based on the use of the decision tree method are presented. When constructing the models, qualitative factors were taken into account (the level of intensification, the system of soil cultivation, the placement of the crop after steam) and agrometeorological resources (the sum of active air temperatures, precipitation). The minimum number of input parameters (public data) was used for the generality of the system and its versatility for different natural and agricultural conditions. The efficiency of using decision trees for forecasting wheat yield is shown. The accuracy of the constructed models was evaluated on the training and test samples and the following indicators were achieved (CART method): average absolute error - 3.455 (training sample) and 4.446 (test sample); determination coefficient - 0.895 (training sample) and 0.811 (test sample). A set of rules has been obtained that determine the level of crop yield depending on the complex of control actions and the prevailing conditions.

1. Introduction

To increase the stability of agricultural production and increase the efficiency of the application of agrotechnical measures, it is necessary to create and implement approaches to their planning in the long term, as well as to predict the yield of cultivated crops in the coming growing season. Predicting the behaviour of complex systems of biological nature, such as agrosystems, in conditions of limited data and information, as well as attempts to create logical models of such systems is one of the most important tasks of the world scientific community. The task of forecasting is complicated by the presence of complex nonlinear relationships between the input factors and the resulting indicators of the functioning of the agro system, a large number of factors with a relatively small number of observations, and the practical complexity of carrying out multifactorial experiments. To this it should be added that agriculture is conducted under conditions of risk and uncertainty, the latter of which is a fundamental property of nature [1-3]. It so happened that economists were mainly engaged in modelling and assessing the effectiveness of agricultural systems when developing programs for the development of agricultural production and management. The main tool for the implementation of this type of tasks were economic and mathematical optimization models. A feature of such modelling is that it allows you to choose one of the best from the area of feasible solutions according to a predetermined efficiency criterion. However, most of these models have limited application, since they describe the state of the agro system at a particular moment in time and in a certain space [4]. Various methods are currently used in
forecasting crop yields: analysis of the trend and cyclicality in the dynamics of yields; identification of the analogue year; construction of regression relationships between various statistical data, including those obtained on the basis of remote sensing and meteorological observations; modelling of agrophysical processes, etc. [1,5,6]. The first and second approaches are distinguished by long-term forecasting, but insufficient accuracy. The most commonly used are regression and simulation. In most cases, the data of agrometeorological resources and their forecast are used as input information for constructing a regression relationship or for modeling plant growth, and the approaches themselves are based on the use of indirect factors [7-10].

Modern research in the direction of predicting crop yields has formed two main principles: empirical - which is characterized by the wide use of heuristic descriptions of determining factors and a mechanistic (ecophysiological) principle, which involves considering the essence of processes and causal relationships in the agroecosystem with a description of their dynamics [11,12]. Over the past decades, a colossal experience of the agro-industrial complex has been accumulated, which contains useful knowledge, but their practical implementation with the formalization of approaches and knowledge engineering is a methodological problem. In domestic agricultural science and practice, the formation of qualitatively new approaches and methods of land use management based on information technologies is taking place [10]. The development of information technology has opened up a space of new opportunities on the way to understanding and predicting the essence of the ongoing processes in agricultural systems [10-14]. This is due to the emergence of the concept of "smart" agriculture, the use of IOT technologies and data mining [15]. The development of models close to the logical reasoning of a person and the use of this method in the construction of decision support systems is one of the main directions of digitalization of agriculture. One of such approaches can be the method of decision trees - a graphical representation of the logical rules for the development of a process depending on the prevailing conditions (operating factors of influence), reflecting alternative options for the development of events.

When constructing decision trees, a machine learning method is used - the selection of algorithm parameters based on a training sample (a set of obtained rules and conclusions) [16,17]. This method is already used in the field of ecology, medicine, economics, sociology. It is successfully used to predict investment risks in agriculture; sustainable development of the agrarian sector, etc. The most interesting from the point of view of studying biological objects was the use of this method for predicting phytosanitary situations for the development of wheat septoria, intellectual analysis of phytosanitary information [18-20]. The introduction of decision trees for particular problems in the field of crop production, one of which is forecasting crop yields, is becoming relevant. The approach we propose will make it possible, with a certain accuracy, to forecast the yield of a cultivated crop, using only the available data. Such an approach can potentially help in making informed decisions on the choice of agricultural technologies and in managing risks in agricultural production.

The purpose of the study is to evaluate the effectiveness of using decision tree methods to predict crop yields (for example, spring wheat).

2. Materials and methods

The work is based on the analysis of long-term data on the yield of spring wheat in the forest-steppe zone of Western Siberia. The materials of long-term field experiments of the Siberian Research Institute of Agriculture and Chemicalization of Agriculture (54° 54'46 " N 82° 56'50 " E) carried out in the Novosibirsk region, the working village of Krasnoobsk and information about the weather Novosibirsk post of meteorological observations near Novosibirsk (the data source was the web resource "weather and climate" (http://www.pogodaiklimat.ru/)). The study used data on the yield of spring wheat and factors determining the yield of the crop for 2001-2019. In the course of the analysis, the following qualitative factors were taken into account: the level of intensification (extensive, intensive); remoteness of culture from steam (first, second, or third); tillage system (dump, dump-free, combined minimum, no tillage), and quantitative climatic factors: the sum of effective air temperatures and precipitation for the period from the third decade of May to the end of July, the sum of effective air temperatures for August.
The choice of climatic factors was carried out on the basis of an analysis of their relationship with the yield of spring wheat; in the course of the study, factors were used that have the maximum correlation with the level of crop yield. In the course of the study, methods of constructing decision trees were used: the CART method - Classification and Regression Tree [21], the CI method - Conditional Inference [22]). All calculations were performed using the R language and the R-Studio statistical data analysis environment. In total, the original database included 448 observations, which were divided into training and test samples in a ratio of 80% and 20%.

3. Results and discussion
The yield of an agricultural crop is one of the indicators of the productivity of the environment, functioning under the interaction of groups of abiotic and biotic factors. The beginning of the creation of a digital model of the process (in our example, this is the yield of an agricultural crop per unit area, expressed in centners / ha) is its informatization, i.e., understanding what data, information and knowledge are needed for further digital control and forecasting of this process. Using the methods of constructing decision trees CART and CI, the models for forecasting the yield of spring wheat were built in two versions - only by qualitative factors and by a combination of qualitative and quantitative factors. The accuracy of the models was assessed on the training and test samples for three main indicators: MAE - mean absolute error; RMSE is the root of the standard deviation of the error and Rsq is the coefficient of determination, expressing the percentage of the explained variance of the resulting feature in fractions of one. Taking into account only qualitative factors when constructing a predictive model (table 1) made it possible to achieve an average absolute forecasting error of 6.446 and 6.211, respectively, on the training and test samples, the coefficient of determination is 0.594 and 0.557. At the same time, the accuracy indicators of the models built by the CART and CI methods do not differ.

Table 1. Statistical indicators of model accuracy.

| №  | Method | Characteristics        | Sample | Measure | MAE  | RMSE  | Rsq  |
|----|--------|------------------------|--------|---------|------|-------|------|
| 1  | CI     | qualitative            | train  |         | 6.446| 8.373 | 0.594|
|    |        |                        | test   |         | 6.211| 8.800 | 0.557|
| 2  | CART   | qualitative            | train  |         | 6.446| 8.373 | 0.594|
|    |        |                        | test   |         | 6.211| 8.800 | 0.557|
| 3  | CI     | qualitative and        | train  |         | 4.748| 6.125 | 0.783|
|    |        | quantitative           |        |         | 5.297| 6.687 | 0.744|
| 4  | CART   | qualitative and        | train  |         | 3.455| 4.258 | 0.895|
|    |        | quantitative           |        |         | 4.446| 5.755 | 0.811|

Figure 1 shows a decision tree constructed by the CI method. The level of intensification has the greatest influence on the yield of the crop, in second place is the placement of the crop after the fallow; the factor “soil cultivation system” did not have a statistically significant effect and was not singled out when constructing the decision tree. The maximum yield of spring wheat (37.561 ± 0.827 kg / ha) corresponds to a combination of an intensive level of intensification and the first or second crop after fallow. The inclusion of climatic features (the sum of effective temperatures and precipitation) in the model allows to increase the accuracy of the models (see Table 1): the mean absolute forecasting error is 4.748 and 5.297, respectively, on the training and test samples (CI method); 3.455 and 4.446 (CART method); determination coefficient - 0.783 and 0.744 (CI method); 0.895 and 0.811 (CART method).
Figure 1. Decision tree (method CI).

Figure 2 shows a decision tree constructed by the CART method. The CART method showed the best results in terms of model accuracy in comparison with the CI method (table 1).

As a result of the analysis, it is possible to identify the ranges of total temperatures and precipitation that correspond to an increase in crop yield: total precipitation for the period from the third decade of May to the end of July - more than 100 mm; the total air temperature from the third decade of May to the end of July is in the range of 1200-1260° C; total air temperature in August - 500-550° C. Exit from the optimal range of trait values, both to the lower and to the higher side, leads to a decrease in the yield of spring wheat.

Figure 2. Decision tree (method CART).

Notes: Intensification Level (IL); Intensive (I), Extensive (E); Culture Placement After Steam (CPAS); Cumulative Temperature May July (CTMJ); Cumulative Precipitation May Jule (CPMJ); Cumulative Temperature August (CTA).
The constructed decision trees describe a set of logical rules that determine the combination of the values of the factors leading to a certain level of crop yield. For example (figure 2), if the level of intensification is intense, the total precipitation for the period from the third decade of May to the end of July is more than 51 mm, the sum of active air temperatures from the third decade of May to the end of July is in the range of 1251-1259° C, then the average yield is 55 ± 1.27 c / ha (maximum yield, CART method). If the level of intensification is extensive and the third crop after fallow, then the average yield is 11 ± 0.48 c / ha (the lowest yield, CART method). In total, the decision tree built by the CART method defines 13 rules.

4. Conclusion
As a result of the study, decision trees were constructed using different methods to predict the yield of spring wheat depending on qualitative factors and weather conditions (the sum of active air temperatures, precipitation). An assessment of the accuracy of the models on the training and test samples is carried out, and a meaningful interpretation of the results obtained is carried out. The efficiency of using decision trees for predicting crop yield is shown, the coefficient of determination is 0.783 (training sample) and 0.744 (test sample) when using the CI method and 0.895 (training sample), and 0.811 (test sample) when using the CART method. A set of logical rules has been obtained that determine the level of crop yield depending on the complex of emerging conditions (influencing factors). This approach is of scientific novelty and practical value in the planning of agricultural technologies, as well as in the creation of knowledge bases that will be used in the work of expert decision support systems in crop production.

References
[1] Lutsenko E V, Loiko V I and Velikanova L O 2008 Forecasting and Decision Making in Crop Production Using Artificial Intelligence Technologies (Krasnodar: KubSAU)
[2] Jones J W, Antle J M, Basso B et al. 2017 Brief history of agricultural systems modelling Agricultural Systems 155 240-54
[3] Kuzmin E A 2012 Uncertainty and Certainty in the Management of Organizational and Economic Systems (Yekaterinburg: Institute of Economics of the Ural Branch of the Russian Academy of Sciences)
[4] Lenkov I I 2011 Modelling and Forecasting the Economy of the Agro-Industrial Complex (Minsk: BSATU)
[5] Savin I Yu, Bartaleev S A, Lupyans E A, Tolpin V P and Khvostikov S A 2010 Forecasting the yield of agricultural crops based on satellite data: opportunities and prospects Current Problems in Remote Sensing of the Earth from Space 7 (3) 275-85
[6] Badenko V L, Topazh A G, Yakushev V V, Mirshel V and Mendel K 2017 Simulation model of agroecosystem as a tool for theoretical research Agricultural Biology 52 (3) 437-45
[7] Asalkhanov P G, Ivanov Ya M and Polkovskaya M N 2017 Models for predicting the yield of agricultural crops in the problems of parametric programming Proceedings of Irkutsk State Technical University 21 2 (121) 57-66 doi: 10.21285 / 1814-3520-2017-2-57-66
[8] Neverov A A 2018 Alternative models of long-term forecasting of grain yield for the steppe zone of the Orenburg region Bulletin of the Orenburg Scientific Center of the Ural Branch of the Russian Academy of Sciences 1 1-9 doi: 10.24411 / 2304-9081-2018-11002
[9] Dunnett A, Shirsath P B, Aggarwal P K et al. 2018 Multi-objective land use allocation modelling for prioritizing climate-smart agricultural interventions Ecological Modelling 381 23-35 doi: 10.1016/j.ecolmodel.2018.04.008
[10] Keating B A and Thorburn P J 2018 Modelling crops and cropping systems – Evolving purpose, practice and prospects European Journal of Agronomy 100 163-76 doi: 10.1016/j.eja.2018.04.007
[11] Bukhovets A G, Semin E A and Biryuchinskaya T Ya 2016 Modern Approaches and Methods in Predicting the Yield of Certain Types of Grain Crops (Voronezh: Voronezh SAU)
[12] Badenko V L, Garmanov V V and Ivanov D A 2015 Prospects for the use of dynamic models of agroecosystems in the tasks of medium and long-term planning of agricultural production and land management *Russian Agricultural Sciences* **1-2** 72-76

[13] Walter A et al. 2017 Opinion: Smart farming is key to developing sustainable agriculture *Proceedings of the National Academy of Sciences* vol. **114** 24 6148-150

[14] Khvorova L A and Topazh A G 2011 Building models of agroecosystems and their adaptation to specific conditions *St. Petersburg Polytechnical University Journal. Computer Science. Telecommunication and Control Systems* **1** (115) 99-105

[15] Sushanth G and Sujatha S 2018 IOT based smart agriculture system *Proc. of the Int. Conf. on Wireless Communications, Signal Processing and Networking (WiSPNET)* (IEEE) pp 1-4

[16] Berestneva O G and Muratova E A 2004 Construction of logical models using decision trees *Bulletin of the Tomsk Polytechnic University. Geo Assets Engineering* **307** (2)

[17] Zharkova O S et al. 2016 Construction of decision support systems in medicine based on decision trees *Modern High Technologies* **6-1** 33-37

[18] Zubarev A 2007 Methods of analysis and assessment of the level of investment potential and risk of the region *Russian Journal of Entrepreneurship* **4-1**

[19] Popova E V, Kumratova A M and Chikatueva L A 2013 Development stability of the agricultural sector: a complex of mathematical methods and models *Polythematic Online Scientific Journal of Kuban State Agrarian University* **90**

[20] Ibragimov T Z and Sanin S S 2019 Digital plant protection and intellectual analysis of phytosanitary information *Plant Protection and Quarantine* **4** 15-18

[21] Breiman L, Friedman J H, Olshen R A and Stone C J 1984 *Classification and Regression Trees.* *Wadsworth* (Routledge) https://doi.org/10.1201/9781315139470

[22] Torsten Hothorn, Kurt Hornik and Achim Zeileis 2006 Unbiased Recursive Partitioning: A Conditional Inference Framework *Journal of Computational and Graphical Statistics* **15**(3) 651–74 Retrieved from: https://www.zieileis.org/papers/Hothorn+Hornik+Zeileis-2006.pdf