Turkish Journal of Agriculture and Forestry

Volume 47 | Number 5 Article 7

10-16-2023

Crop weed infestation forecasting using data mining methods

KIRILL MAKSIMOVICH

OLGA ALSOVA

VLADIMIR KALICHKIN

DMITRY FEDOROV

Turkish Journal of Agriculture and Forestry

http://journals.tubitak.gov.tr/agriculture/

Research Article

Turk J Agric For
(2023) 47: 662-668

doi:10.55730/1300-011X.3118

Crop weed infestation forecasting using data mining methods

Kirill MAKSIMOVICH[1,*], **Olga ALSOVA**[2], **Vladimir KALICHKIN**[1], **Dmitry FEDOROV**[1,2]

[1]Siberian Federal Scientific Centre of Agro-BioTechnologies of the Russian Academy of Sciences, Krasnoobsk, Russia

[2]Novosibirsk State Technical University, 20 Prospekt K. Marksa, Novosibirsk, Russia

Received: 16.09.2023 • **Accepted/Published Online:** 02.06.2023 • **Final Version:** 16.10.2023

Abstract: The studies were carried out to develop logical rules for crop weed infestation forecasting using data mining methods within limited sampling conditions. The classical probabilistic-statistical methods and decision tree method were used. Selected methods were chosen considering the distribution of the input data, the diversity of its factors and attributes, the relationship structure peculiarities. The analysis was performed on the extensive field experiments data of the Kemerovo Scientific Research Institute of Agriculture – a branch of SFSCA RAS on weed infestation of agricultural soils by cereal weeds for 2013–2019. The qualitative factors (tillage system and first crop) and meteorological features (average ten-day air temperatures and amount of precipitation) determining the indicators of crop weed infestation were outlined. The statistical significance and contribution rate of each factor were evaluated. The decision tree based on the set of logical rules was built, which makes it possible to forecast the weediness index. The coefficient of determination of the model was 0.68, which is a sufficiently successful result for the forecast of the biological nature object. The results obtained (information about relations between indicators and factors, a set of logical rules) can be used in the design of knowledge-based decision-making support systems in crop production.

Key words: Data mining, agriculture analysis, weediness forecasting, cereal weeds, decision trees, R

1. Introduction

Current agricultural science and practice is increasingly being based on information technology. There are many problems associated with the analysis and processing of massive amounts of accumulated data using modern approaches, methods, and software. One of the relevant trends is the use of data mining and machine learning methods. Data mining is the process of extracting information from empirical data and synthesis of previously unknown, nontrivial, and practically useful data to achieve certain goals with the help of mathematical models, methods, software, and information technology (Rizaev et al., 2011; Lemeshko et al., 2011; Kravchenko et al., 2012). Mathematical statistics, pattern recognition, artificial intelligence, decision trees, neural networks, theory of databases, and many others are the main subjects of the data mining approach.

Until recently, the only methods used in agriculture for data analysis and processing were probabilistic-statistical methods. These methods tended to be parametric, requiring a number of a priori assumptions about the nature of the raw data. Probabilistic-statistical methods allow building models with a number of advantages.

The models are "transparent" and allow a meaningful interpretation with the possibility to assess the statistical significance of both the model parameters and the obtained results. Furthermore, many algorithms have been developed and a lot of experience has been accumulated in their application in solving applied problems. However, the use of probabilistic-statistical methods is not always mathematically strictly justified due to violation of a priori assumptions and data requirements, methods may not take into account the specific structure of data (for example, nonlinear type of relationship between the target feature and factors) and, consequently, not provide high accuracy when solving the problem. Hence, it becomes relevant to expand the set of used data mining methods for agricultural data analysis and processing (Van Wart et al., 2013; Chipanshi et al., 2015; Majumdar et al., 2017; Morota et al., 2018; Muangprathub et al., 2019; Kremer, 2019; Asad et al., 2019; Kravchenko et al., 2021; Orlov, 2021; Shahhosseini et al., 2021). The use of decision tree methods (classification and regression) is considered to be perspective (Yakovlev et al., 2018; Kalichkin et al., 2021). A decision tree is a logical model of data described in the form of a hierarchical structure of classification IF-THEN rules.

* Correspondence: kiri-maksimovi@mail.ru

The central concern of agrophytocenosis management is the ability to predict crop weed infestation and the effective use of crop protection measures. Weed vegetation is the most dynamic part of agrophytocenosis and at a certain level of its development becomes a limiting factor in producing rich crop yields. As a result of unfavourable phytosanitary conditions, the annual crop yield losses amount up to 25%–30% (Spiridonov et al., 2016). In Western Siberia, cereals occupy one of the leading positions among weeds in terms of amount and biomass. The prevailing position is due to the ability of intensive growth, unpretentiousness to biotic factors, and sufficiently high seed production or fecundity (Spiridonov et al., 2016; Luneva et al., 2017). This class of weeds is also characterized by the phenomenon of heterocarpy, which contributes to the prolonged emergence of weed seedlings in time and the preservation of seeds in the soil for up to three or more years. The ability to forecast the situation and predict the growth and development of crop weed infestations becomes an important aspect of crop production management. Classical methods of data processing and modeling cannot always provide the required accuracy, which justifies the relevance of searching for new ways to perform analysis and build complex forecasting models (Ibragimov et al., 2019).

The aim of the study is to develop logical rules for crop weed infestation forecasting using data mining methods within limited sampling conditions.

2. Materials and methods

The study was carried out on the extensive field experiments data of the Kemerovo Scientific Research Institute of Agriculture – a branch of SFSCA RAS, conducted in 2013–2019 in the Kuznetsk forest-steppe of Kemerovo region (Novostroeniya village), Russia. The complex of weed plants included mainly cereal species: proso millet (*Panikum miliaceum L.*), wild oats (*Avena fatua L.*), field sowthistle (*Sonchus arvensis L.*), and green foxtail (*Setaria viridis L.*). As factors having an influence on the level of weed infestation the systems of tillage: no-tillage (NO), deep tillage (DT), minimum tillage (MT), deep no-tillage (DNO), minimum no-tillage (MNO), combined minimum (CM), combined deep (CD), the first crop (rapeseed (*Brássica nápus*), melilot (*Melilótus officinális*), summer fallow) were examined. The study also used agrometeorological resource data: average ten-day air temperatures (°C) and precipitation (mm), sums of temperature (°C) and precipitation (mm) for the 3rd ten-day periods of April and 1-3 ten-day periods of May (web resource Weather and Climate (http://www.pogodaiklimat.ru/) provided the data). In the study, the probability-statistical methods and criteria were used to identify the distribution of data (Shapiro-Wilk and Anderson-Darling tests [2]), to identify the factors influencing the crop weed infestation and evaluate their contribution (Kruskal-Wallis (Kruskal and Wallis, 1952), median and Spearman rank correlation coefficient tests). The CART method (Breiman et al., 1984) was also used to build a decision tree for crop weed infestation level forecasting. Data visualization methods for graphical representation of the data structure and analysis results (box plots, histograms, and decision trees) were applied.

In order to avoid the effects of model overfitting or underfitting, a decision tree optimization procedure was performed using the tree complexity parameter and 10-fold cross-validation with 10 repetitions (Breiman et al., 1984). Cross-validation involved splitting the initial data set into ten subsamples (subsamples were formed using a random number generator), where nine parts were used as a training sample to build the model, and one part was used as a test sample. The test sample was used to calculate the accuracy of the model. Then another test part was selected, and the process was repeated 10 times, so each subsample was used 9 times as a training sample and once as a test sample. The whole procedure was repeated 10 times (a total of 100 tests). Based on the results of all tests, the accuracy values of the model were averaged.

For the qualitative characteristic of the crop weed infestation, the indicator of the number of weeds (obtained in the experiment) was introduced, taking the value of one of the 4 categories of exceeding the economic threshold of harmfulness (ETH = 15 pcs./m^2): N – no exceeding (up to 15 pcs/m^2); LE – low exceeding (up to 20 pcs./m^2); ME – moderate exceeding (2–3 times greater than the ETH); SE – significant exceeding (more than 6–7 times greater than the ETH).

All the calculations were performed using R language in R-Studio statistical data analysis environment.

3. Results and discussion

The first stage of the study identified the factors that have a statistically significant effect on the weeding of spring wheat crops and which should be taken into account when building a forecasting model. As it is known, a choice of statistical methods for correlation research depends on the type of features, the type of supposed correlation, and on the type of data distribution. In case of normal distribution of initial data and sufficiently large sample size (the required sample size depends on the research method), parametric methods are used, otherwise, it is necessary to use their nonparametric counterparts. A histogram of distribution overlaid with an approximating curve of the probability density function is shown in Figure 1. Visually, the histogram does not fit the normal data distribution model and has a pronounced left asymmetry. In addition, the mean, median, and mode estimates are not equal to

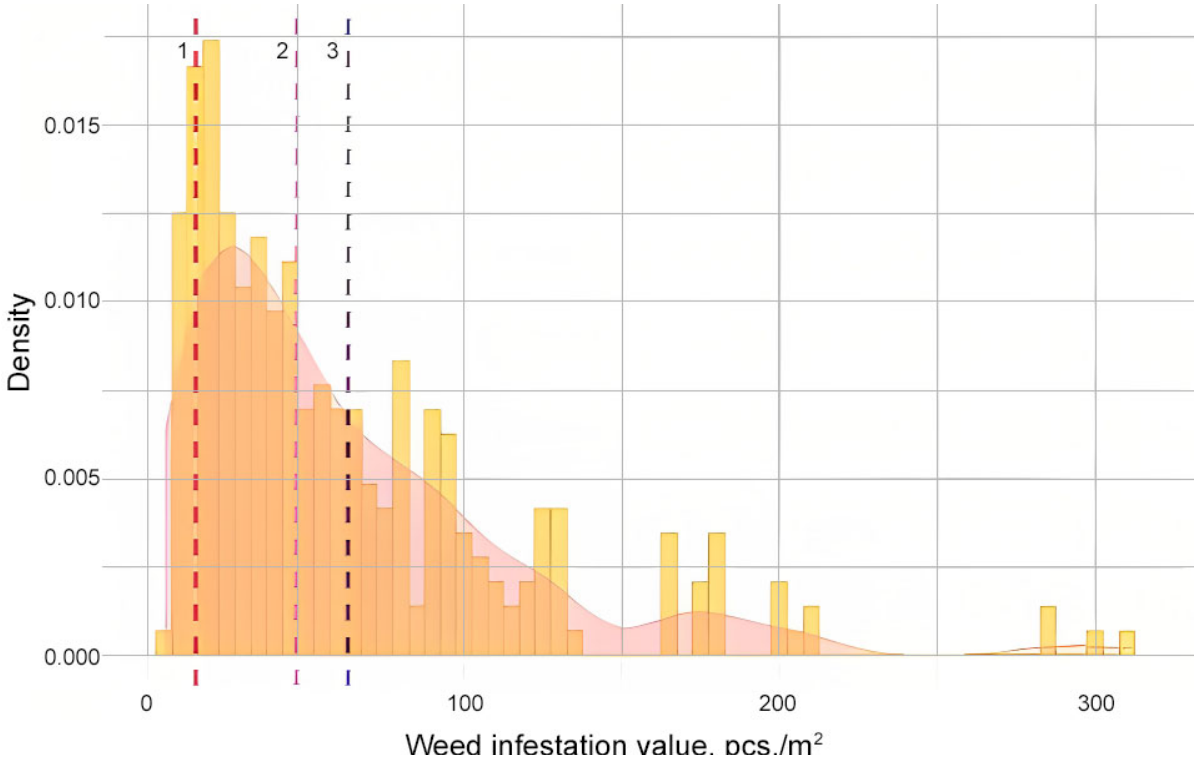

Figure 1. Histogram of weed infestation values with approximating probability density function (the graph shows statistical characteristics (from left to right): 1 – mode, 2 – median, 3 – mean).

each other, indicating that the distribution is asymmetric. The Shapiro-Wilk and Anderson-Darling tests were also used to examine the hypothesis that the distribution of crop weed infestation was consistent with the normal distribution model. The hypothesis was rejected as a result of both tests (p-value < 2.2e-16) at a significance level of 0.05.

Similar results were obtained when examining the law of distribution of the weediness level of crops in different subsamples formed depending on the values of qualitative factors (first crop and tillage system). Consequently, it is necessary to use nonparametric methods to analyse the relationships between the weediness level of crops and dependent factors. Therefore, the Kruskal-Wallis test was chosen to assess the relationship between the weediness level of crops and qualitative factors, and the Spearman rank correlation coefficient was chosen to evaluate the relationship between the weediness level of crops and meteorological data.

The changes in the statistical characteristics of the weediness of crops in dependence on qualitative factors were also analysed (Figure 2). On the box plot, the position of the central line determines the median value, the borders of the rectangle correspond to the lower and upper quartiles, the height of the rectangle is the interquartile range (IQR), the whiskers of the plot are located from the upper (lower) border of the rectangle to the highest (lowest) value, which is within 1.5 x IQR. Values outside the whiskers may be anomalous outliers.

The Kruskal-Wallis test and the median test have revealed a statistically significant relationship between the crop weediness values and the tillage system ($p < 0.0001$), as well as the differences in the year of observation ($p < 0.0001$). The first crop has no statistically significant effect on the weediness of spring wheat at the significance level of 0.05 ($p = 0.1726$ for the Kruskal-Wallis test and $p = 0.1784$ for the median test). The highest level of crop weed infestation (median = 148 pcs./m^2) is observed when using minimum no-tillage, the lowest–when using combined minimum (median = 35.5 pcs./m^2) and minimum tillage (median = 33 pcs./m^2). The relationship between the crop weediness value and the year of the observation confirms the hypothesis that the level of weediness of crops depends on weather changes.

The values of the Spearman linear rank correlation coefficient and evaluation of its statistical significance are shown in Table 1, which confirms the presence of a moderate correlation (more than 0.6) between data on crop weed infestation and meteorological data. Statistically significant negative correlations with the temperature

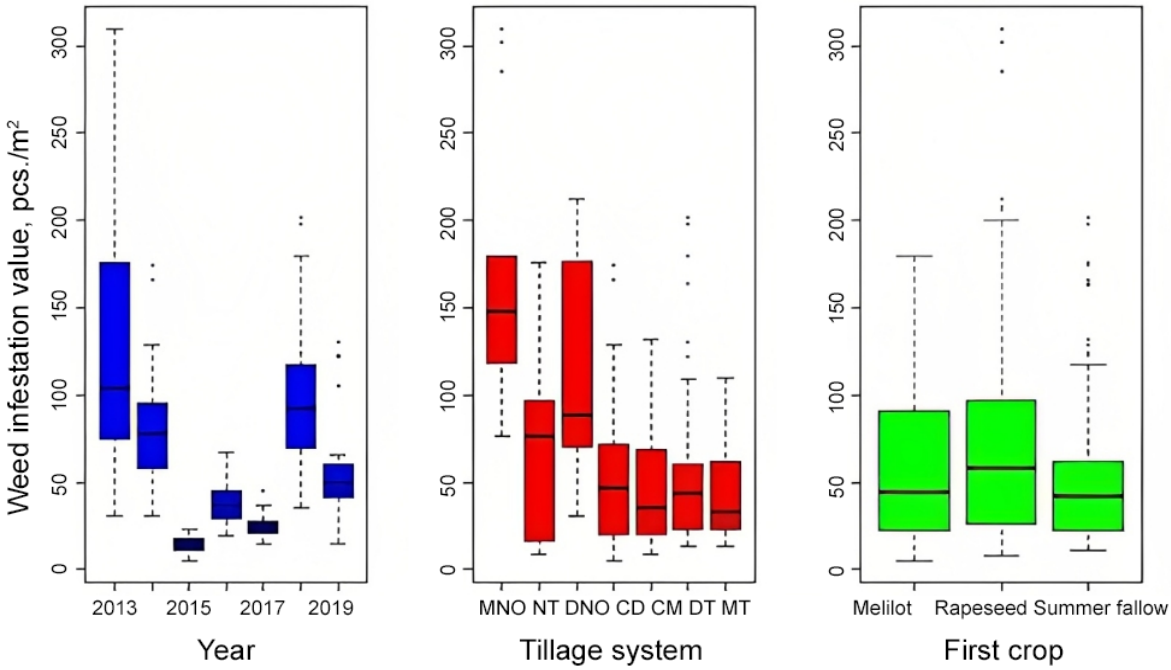

Figure 2. The box plots of the weed infestation levels depending on the qualitative factors' values.

of the 3rd ten-day period of April, the 2nd-3rd ten-day periods of May, the sum of temperatures for the period from the 3rd ten-day period of April to the end of May were revealed. At the same time, statistically significant positive correlations with the temperature of the 2nd ten-day period of May, the sum of precipitation for the period were revealed.

The results obtained are appropriate to the biology of cereal weeds. The sustained transition of air temperature through 10 °C determines the period of intensive biochemical processes in soil and coincides with the sprouting of weeds (Kalichkin et al., 2020). Mass sprouts of cereal weeds are usually associated with a sum of average daily air temperatures of 260–340 °C that accumulates in the forest-steppe zone of Western Siberia by the 2nd-3rd decade of May. Spring precipitation is usually one of the key factors increasing the intensity of growth and development of weeds. Low air and soil temperatures in the spring (1–2 decades of May) do not significantly impact the germination process but determine the intensity of weed vegetation growth.

Given the small sample size, the availability of both qualitative and quantitative predictors, and significant deviations in the distribution of the level of weediness of crops from the law of normal distribution, a decision tree method was chosen to build a predictive model.

The decision tree graphically represents a logical model of the relationship between crop weed infestation and contributing factors (Figure 3). Each node of the tree shows the predicted value of the infestation and the number of observations (n) in the node and the percentage of total observations.

Based on the constructed decision tree, the following conclusions emerge. High average daily air temperatures in the second decade of May reduce the intensity of weed growth in favorable conditions for the active growth of cultivated plants, which can create serious competition to weeds in the struggle for moisture and nutrients. Application of combined and mouldboard minimum tillage in conditions of average daily air temperatures in the second decade of May in the range of 7.4–9.6 °C also causes moderate excess of ETH, but the forecast indicator has a borderline value that is classified as "insignificant excess". This is due to the depth of tillage, particularly, at deep mouldboard and combined tillage, and low rates of soil warming, the method of tillage causes the germination of weed seeds. At minimum tillage, the opposite process occurs and the "seed bank" of weeds is also in the depth under a "preserved" state.

Application of DT, MT, and DNO in conditions of average daily air temperatures of the second ten-day period of May below 7.4 °C and precipitation of the first ten-day period of May below 23 mm on summer fallow also causes moderate exceeding of the ETH. The depth of tillage and the amount of precipitation are the limiting factors of growth intensity in this case. For summer fallow and melilot first crops in similar conditions of air temperature and precipitation, the forecasted category

Table 1. Spearman correlation coefficient values.

Parameter	Coefficient value	p-value
Temperatures for:		
3rd ten-day period of April	–0.67	<0.05
1st ten-day period of May	0.37	<0.05
2nd ten-day period of May	–0.84	<0.05
3rd ten-day period of May	–0.56	<0.05
3rd ten-day period from April to the end of May (the sum)	–0.81	<0.05
Precipitation for:		
3rd ten-day period of April	0.02	>0.05
1st ten-day period of May	0.32	<0.05
2nd ten-day period of May	0.52	<0.05
3rd ten-day period of May	0.08	>0.05
3rd ten-day period from April to the end of May (the sum)	0.60	<0.05

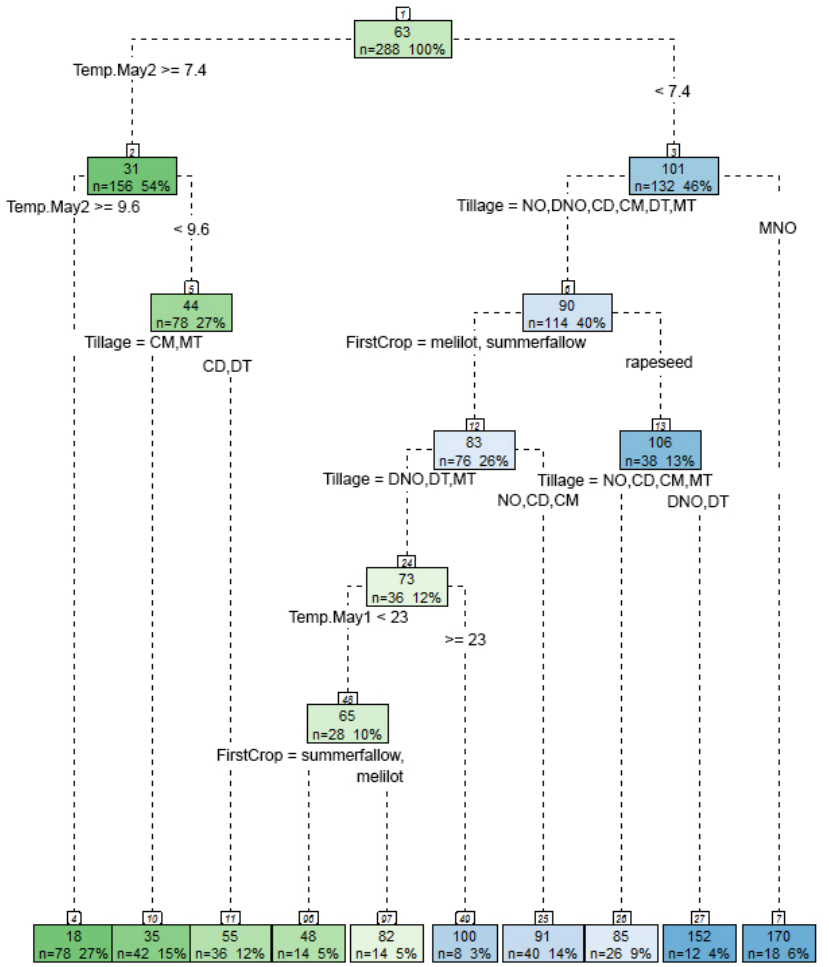

Figure 3. Decision tree for crop weed infestation forecast.

of weed infestation is also ME. An important factor of a significant excess of ETH becomes the presence of precipitation above 23 mm, which is evident in Figure 3 (tree leaf No. 49). In case of abundant precipitation for a decade, weediness takes the category of SE.

One of the highest indicators of weeds per square meter (152 pcs./m^2) is observed at average air temperature for the second ten-day period of May less than 7.4 °C, at no-tillage and deep tillage. This is explained by the influence of the depth of tillage and spring conditions, when weed vegetation, due to its unpretentiousness to the environmental conditions, is the most active and competitive in comparison with the cultivated plant. The method of tillage causes the intensity of germination and growth of weeds (increased rate of soil warming and improved oxygen regime). A higher level of weed infestation was also recorded when spring wheat was sown on rapeseed than on summer fallow or melilot.

It is possible to define the most significant logical rules describing the combination of relevant factors and forecasting the level of spring wheat weed infestation using the decision tree that has been built.

IF (temperature = "average daily air temperature of the 2nd ten-day period of May < 7.4 °C") AND (precipitation = "precipitation of the 1st ten-day period of May < 23 mm") AND (tillage = "deep tillage" OR "minimum tillage" OR "deep no-tillage") AND (first crop = summer fallow) THEN (crop weed infestation value = ME (category) 48 pcs./m^2).

IF (temperature = "average daily air temperature of the 2nd ten-day period of May < 7.4 °C") AND (precipitation = "precipitation of the 1st ten-day period of May < 23 mm") AND (tillage = "deep tillage" OR "minimum tillage" OR "deep no-tillage") AND (first crop = "melilot") THEN (crop weed infestation value = ME (category) 82 pcs/m^2).

IF (temperature = "average daily air temperature of the 2nd ten-day period of May < 7.4 °C") AND (tillage = "combined minimum" OR "minimum tillage" OR "combined deep" OR "no-tillage") AND (first crop = "rapeseed") THEN (crop weed infestation value = ME (category) 85 pcs./m^2).

IF (temperature = "average daily air temperature of the 2nd ten-day period of May = < 7.4 °C") AND (tillage = "deep no-tillage" OR "deep tillage") AND (first crop = "rapeseed") THEN (crop weed infestation value = SE (category) 152 pcs./m^2).

The accuracy of the decision tree model was determined: MAE – mean absolute error – 18.61; RMSE – root mean square error – 30.62 and R^2 – determination coefficient that expresses the percentage of explained variance of the resultant attribute in fractions of one – 0.68.

4. Conclusions

The analysis and forecasting of the crop weed infestation level based on the application of different classes of data mining methods were carried out within the framework of the study. The methods of the study were chosen by taking into account the features of data on the distribution of crop infestation as a target variable, different types of predictor variables (qualitative and quantitative factors and attributes), and the structure and type of relationships in the initial data. Qualitative factors (tillage system, first crop) and meteorological data (average ten-day air temperatures and amount of precipitation) determining the crop weed infestation value on the field were identified and the extent of their influence was assessed. A decision tree was constructed, which makes it possible to forecast the crop weed infestation values using a set of logical rules; the coefficient of determination of the model is 0.68, which is a sufficiently successful result for forecasting objects of biological nature in conditions of uncertainty.

The use of different data mining methods during the study made it possible to conduct a complete analysis of the distribution and structure of relationships in the data, to build an appropriate forecast model of spring wheat crop weed infestation, and to make reasonable conclusions and meaningful interpretations of the results obtained. The approach based on the integrated application of data mining methods is effective and has scientific and practical significance. The results obtained (information about relations between indicators and factors, a set of logical rules) can be used in the design of knowledge-based decision-making support systems in crop production.

References

AAsad MH, Bais A (2019). Weed detection in canola fields using maximum likelihood classification and deep convolutional neural network. Information Processing in Agriculture 7 (4): 535-545. doi:10.1016/j.inpa.2019.12.002

Asad MH, Bais A (2019). Weed detection in canola fields using maximum likelihood classification and deep convolutional neural network. Information Processing in Agriculture 7 (4): 535-545. doi:10.1016/j.inpa.2019.12.002

Bedi P, Gole P (2021). Plant disease detection using hybrid model based on convolutional autoencoder and convolutional neural network. Artificial Intelligence in Agriculture 5: 90-101. doi:10.1016/j.aiia.2021.05.002

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/ Cole Advanced Books & Software 40 (3): 874. https://doi.org/10.2307/2530946

Chipanshi A, Zhang Y, Kouadio L, Newlands N (2015). Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. Agricultural and Forest Meteorology 206: 137-150. https://doi.org/10.1016/j.agrformet.2015.03.007.

Ibragimov TZ, Sanin SS (2019). Digital plant protection and intelligent analysis of phytosanitary information. Plant Protection and Quarantine 4: 15-18.

Kalichkin VK, Alsova OK, Maksimovich KYu (2021). Application of the decision tree method for predicting the yield of spring wheat. IOP Conference Series: Earth and Environmental Science. Krasnoyarsk, Russia, 16-19 June 2021, IOP Publishing. https://doi.org/10.1088/1755-1315/839/3/032042

Kalichkin VK, Maksimovich KYu, Galimov RR (2020). Bayesian network for predicting the level of crops weediness with wild oat. Journal of Physics: Conference Series. IOP Publishing 1679 (2). https://doi.org/10.1088/1742-6596/1679/2/022097

Kravchenko YuA, Lezhebokov AA, Zaporozhets DYu (2012). Methods of intelligent data analysis in complex systems. News of the Kabardin-Balkar scientific center of RAS 3: 52-57. https://doi.org/10.21686/1818-4243-2014-3/104-49-54

Kremer NSh (2019). Probability theory and mathematical statistics as the foundation of a new integrated applied discipline "data analysis". Modern mathematics and concepts of innovative mathematical education 6 (1): 333-337.

Kruskal WH, Wallis WA (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association 47 (260): 583-621.

Lemeshko BYu, Lemeshko SB, Postovalov SN, Chimitova EV (2011). Statistical Data Analysis, Modeling and Investigation of Probability Patterns. Computer Approach. Monograph. Novosibirsk: NSTU Press.

Luneva NN, Mysnik EN, Bochkarev DV, Nikolsky AN, Kuzovatkin EM (2017). Ecological and geographical substantiation of the formation of species composition of weeds in the territory of the Republic of Mordovia. Agrarian Scientific Journal 6: 25-30. https://doi.org/10.7868/S0367059717010036

Majumdar J, Naraseeyappa S, Ankalaki S (2017). Analysis of agriculture data using data mining techniques: application of big data. Journal of Big Data 4 (1): 1-15. https://doi.org/10.1186/s40537-017-0077-4

Morota G, Ventura RV, Silva FF, Koyama M, Fernando SC (2018). Big data analytics and precision animal agriculture symposium: Machine learning and data mining advance predictive big data analysis in precision animal agriculture. Journal of Animal Science 96 (4): 1540-1550. https://doi.org/10.1093/jas/sky014

Muangprathub J, Boonnam N, Kajornkasirat S, Lekbangpong N, Wanichsombat A et al. (2019). IoT and agriculture data analysis for smart farm. Computers and electronics in agriculture 156: 467-474.

Orlov AI (2012). A new paradigm of applied statistics. Zavodskaya Laboratoriya, Diagnostika Materialov 78 (1-1): 87-93.

Pérez-Miñana E, Krause PJ, Thornton J (2012). Bayesian Networks for the management of greenhouse gas emissions in the British agricultural sector. Environmental Modelling & Software 35: 132-148. https://doi.org/10.1016/j.envsoft.2012.02.016

Rizaev IS, Rahal Y (2011). Intelligent data analysis for decision support. Monograph. Kazan: Editorial and Publishing Center "Shkola".

Shahhosseini M, Hu G, Huber I, Archontoulis Sv (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Scientific reports 11 (1): 1-15. https://doi.org/10.1038/s41598-020-80820-1

Spiridonov YuY, Budynkov NI, Sayfullin RG, Strizhkov NI, Ataev, SS Kh et al. (2016). Control of pests on crops of field crops. Agrarian Scientific Journal 9: 43-48. https://doi.org/10.28983/asj.v0i7.525

Trofimova IE, Balybina AC (2015). Regionalization of the West Siberian Plain from thermal regime of soils. Geography and natural resources 3: 27-38. https://doi.org/10.1134/S1875372815030038

Van Wart J, Grassini P, Cassman KG (2013). Impact of derived global weather data on simulated crop yields. Global change biology 19 (12): 3822-3834. https://doi.org/10.1111/gcb.12302

Yakovlev SS, Seredin OS (2018). Using decision trees to visualize multidimensional data. News of Tula State University, Technical Sciences 10: 137-145.