DOI: 10.37943/19PPFN3256

Sapar Toxanov
PhD in Information Systems, Vice-Rector for Educational Work
sapar.toxanov@astanait.edu.kz , orcid.org/0000-0002-2915-9619
Astana IT University, Kazakhstan

Dilara Abzhanova
Director of the Center of Competence and Excellence
dilara.abzhanova@astanait.edu.kz , orcid.org/0000-0002-7988-3971
Astana IT University, Kazakhstan

Alexandr Neftissov
PhD, Associate Professor, Director of the Science and Innovation Center "Industry 4.0"
alexandr.neftissov@astanait.edu.kz , orcid.org/0000-0003-4079-2025
Astana IT University, Kazakhstan

Andrii Biloshchytskyi
Doctor of Technical Sciences, Professor, Vice-rector for Science and Innovation
a.b@astanait.edu.kz , orcid.org/0000-0001-9548-1959
Astana IT University, Kazakhstan.
Professor of Information Technologies Department, Kyiv National University of Construction and Architecture, Ukraine

METHODS OF FORECASTING GRAIN CROP YIELD INDICATORS TAKING INTO ACCOUNT THE INFLUENCE OF METEOROLOGICAL CONDITIONS IN THE INFORMATION-ANALYTICAL SUBSYSTEM

Abstract: Forecasting crop yields is one of the key challenges for the agricultural sector, especially in the context of a changing climate and unstable weather conditions. Kazakhstan, possessing significant territories suitable for growing grain crops, faces many challenges related to the effective management of agricultural activities. In this regard, yield forecasting becomes an integral part of planning and decision-making processes in agriculture. Information and analytical subsystems that integrate yield forecasting methods allow agribusinesses to estimate future production more accurately, minimise risks associated with climate change and optimise resource use. An important component of such systems is the consideration of weather conditions, as weather factors have a direct impact on crop growth and development. The purpose of this article is to develop and evaluate modern methods of forecasting grain yields taking into account the influence of weather conditions, as well as their integration into information-analytical subsystems to improve the accuracy of agricultural forecasting. To achieve this goal, the article addresses the following tasks: to analyse existing methods of yield forecasting and identify their advantages and disadvantages, to develop forecasting models, including machine learning methods such as gradient bousting and recurrent neural networks, to validate the developed models on the basis of historical data using cross-validation methods, to evaluate the effectiveness of the proposed methods and compare them with basic models such as linear regression and simple average, to evaluate the effectiveness of the proposed methods and to compare them with the basic models such as linear regression and simple average. This article reviews modern methods of forecasting grain crop yields in

Received: 05.09.2024 Accepted: 24.09.2024 Published: 30.09.2024

Kazakhstan, as well as technologies used in information-analytical subsystems. Particular attention is paid to the analysis of the influence of meteorological conditions on yields and the development of models that take this factor into account. The presented review and research results are aimed at improving the existing approaches to the management of agricultural processes under conditions of growing uncertainty caused by climate change. The article explores an important scientific task related to the development of methods for step-by-step forecasting of agrometeorological factors and grain yields, relying on the principle of analogy.

Keywords: forecasting, grain crops, meteorological conditions, Kazakhstan, agricultural technologies, climate, forecasting algorithms, agrarian activity management.

Introduction

In recent years, the state policy of Kazakhstan is aimed at developing and ensuring the stable functioning of agriculture and agro-industrial production. In his message of September 1, 2023, entitled "The Economic Course of Fair Kazakhstan", President Kasym-Jomart Tokayev emphasized the need for a significant breakthrough in the country's agro-industrial complex [1]. Despite the huge potential of the domestic agricultural sector, its opportunities have not yet been fully realized. Given the fact that Kazakhstan is surrounded by large markets with a need for quality food products, the strategic goal of the country is to become one of the leading agricultural centers of the Eurasian continent.

Also, to improve the competitiveness of Kazakhstan's agriculture, to ensure food security of the country and sustainable development of rural areas, the "Concept of development of agro-industrial complex of the Republic of Kazakhstan for 2021-2030" is realized [2]. Modernization and innovation act as a key direction aimed at the introduction of modern technologies, innovative solutions and digitalization in agriculture. This includes the application of precision farming, effective resource management and the use of advanced agro-technologies, which will significantly increase productivity and quality of agricultural products. The development of export potential is a priority within the strategy, which envisages an increase in the volume of exports of Kazakhstan's agricultural products. This is aimed at strengthening the country's position in international markets and increasing the competitiveness of domestic goods by improving their quality and compliance with international standards. An important component is the improvement of agro-technologies, which covers optimization of crop cultivation methods, improved use of fertilizers, plant protection, improved quality of seed material and development of breeding programs. These measures are aimed at increasing yields and sustainability of agricultural production. The Concept pays considerable attention to infrastructure development, including improving the transportation and logistics network, increasing the availability of modern agricultural machinery and equipment, and developing storage and processing systems. This will ensure more efficient management of production processes and minimize losses at all stages of the supply chain. Environmental sustainability is an important aspect of the concept, providing for the conservation of natural resources, sustainable management of land and water resources, reduction of negative impact on the environment and support for environmentally friendly production. These measures are aimed at ensuring the long-term sustainability of agriculture and preserving the country's natural wealth. State support for the agricultural sector provides for increased subsidies, preferential lending, development of insurance programs and other measures aimed at creating favorable conditions for agrarian business. This will increase the investment attractiveness of the sector and stimulate its development. Special attention in the concept is paid to the development of human capital, which includes the improvement of personnel qualifications, development of rural education, as well as the introduction of training and retraining programs for specialists for the agrarian sector. This is important for the creation of a highly professional labor force capable of effectively using modern technologies and managing production processes. Finally, social development of rural areas is aimed at improving the quality of life in rural areas, developing social infrastructure and supporting small and medium-sized businesses. These measures will create conditions for sustainable development of rural areas, contributing to poverty reduction and improving the living standards of the rural population. Thus, the concept is focused on achieving sustainable and long-term growth of the agro-industrial complex of Kazakhstan, ensuring food independence and improving living standards in rural areas, which corresponds to the strategic objectives of the country's development.

In this context, the development and implementation of information and analytical subsystems using methods and algorithms for forecasting grain crop yields taking into account the influence of weather conditions is an important element of the strategy of agricultural development in Kazakhstan. These systems will help to unlock the potential of the agricultural sector and ensure sustainable growth that meets the objectives set by the state.

Literature review

Forecasting plays a central role in agronomy, helping to optimise agricultural processes and increase yields. A variety of methods are used to analyse and generate forecasts, which include both traditional statistical approaches and modern machine learning and artificial intelligence techniques.

Traditional forecasting methods, widely used before the advent of neural networks, remain important tools in various research areas. One such method is the optimisation and simulation method proposed in [3]. This method is based on the application of Monte Carlo method using optimal molecular descriptors to predict the retention time of pesticides in gas chromatography. The study demonstrates the high accuracy and reliability of this model, confirming its ability to adapt to different types of pesticides.

Time series analyses and extrapolation methods also occupy an important place in forecasting. For example, in a research paper [4] proposed a trend decomposition method for forecasting solar energy generation. This method increases the accuracy of forecasts and improves the ability of machine learning algorithms to generalise data.

The process of time series forecasting using the proposed architecture starts by processing the raw data, which is then decomposed into trend, seasonal and residual components. Once the trend component is extracted, the data becomes more stable, allowing existing machine learning models to be used for forecasting. If extrapolation is necessary, linear models are applied to predict the trend. The final step is to combine the stable data and the predicted trend values to obtain the final prediction. The method is characterised by low computational complexity, which makes it effective for practical applications in solar energy forecasting.

Exponential smoothing methods find application in traffic forecasting in cellular networks, as shown in the study of Tran Q. T. and co-authors [5]. Their work demonstrates the effectiveness of this method, its low computational complexity and its versatility for different types of traffic. Exponential smoothing provides high prediction accuracy for both voice and data traffic in cellular networks.

The Holt-Winters method also deserves attention in the context of analysing and forecasting time series taking into account trends and seasonal changes. A study [3] emphasises its importance by allowing short-term and long-term data patterns to be taken into account, making it useful for forecasting various aspects including crop yields and economic performance.

Recurrent neural networks with long short-term memory (LSTM) are another important tool in time series forecasting. A study by Okur and Mori showed that LSTM efficiently processes and analyses sequential data by capturing complex non-linear patterns, which makes it useful for forecasting in areas such as financial data and climate change [6].

Data preprocessing plays an important role in improving the prediction accuracy of neural networks. A study [7] showed that proper preprocessing including normalisation and outlier removal significantly improves the quality of forecasts.

In conclusion, prediction using neural networks is actively developing as a research area. There are many methods and algorithms that can be used depending on the specific problem. With the increase in the amount of data and the development of deep learning algorithms, we can expect further improvement in the quality of predictions using neural networks.

A study [8] compared different types of neural networks for wind speed prediction and showed that a combination of convolutional and recurrent neural networks gave the best results in terms of accuracy, although the ARIMA method showed better time performance.

Forecasting using neural networks is also reflected in the works of Kazakhstani scientists, for example, in the work [9], the author uses recurrent neural networks to analyse the stock market, concluding that the quality of forecasts depends on the processing and structuring of input data. The study [10] makes a comparative analysis of oil price forecasting methods and emphasises the accuracy of forecasts provided by neural networks. And papers [11], [12] investigate the application of neural networks for predicting soil moisture in Northern Kazakhstan, noting the high accuracy of forecasts.

Paper [13] reviews various machine learning methods for wheat yield forecasting, emphasising their effectiveness in a changing climate.

Currently, highly efficient modelling systems for predicting the production process have been developed, such as AGROTOOL, EPIC (Soil & Water Research Laboratory, USDA-ARS), AGROSIM (Centre for Agricultural Landscape Research, Müncheberg, Germany) and others. These complexes allow predicting the consequences of agro-technological measures even before their practical implementation, being integrated directly into decision-making processes. However, their application is limited by the lack of necessary agrometeorological data for future periods.

Research methodology

In this study, modern methods of grain crop yield forecasting that take into account the influence of meteorological conditions were used. The main attention is paid to the analysis and application of information-analytical subsystems that integrate different approaches to forecasting.

1. Data collection and processing. Historical data on grain yields and weather conditions for the last 5 years collected from various sources, including state meteorological services and agricultural organizations, were used to develop forecasting models. The data were preprocessed using cleaning, normalization, and de-emphasis techniques to improve their quality and ensure the accuracy of the models.

2. Prediction methods used both traditional statistical methods such as linear regression and time series, as well as advanced machine learning techniques including:

- gradient boosting was used to create models that account for complex interactions between inputs, including weather conditions and agronomic factors.
- recurrent neural networks were used to predict crop yields based on time series of weather data, allowing for long-term dependencies and seasonal fluctuations.
- The Holt-Winters method was used to analyze and forecast time series, taking into account trend and seasonality, in order to more accurately account for climatic factors.

3. Model validation. The models were tested and validated against historical data using cross-validation techniques. Metrics such as mean square error (MSE) and coefficient of determination (R^2) were used to assess the accuracy of the predictions.

4. Implementation into the information and analytical subsystem. The developed models were integrated into the information-analytical subsystem, which is used by agricultural enterprises for planning agricultural activities. The subsystem was configured to automatically update data and adapt the models to changing conditions.

5. Performance evaluation. The effectiveness of the proposed methods and models was evaluated by analyzing their applicability in real conditions and comparing them with existing approaches. The results show that the integration of forecast models taking into account meteorological conditions allows to significantly improve the accuracy of forecasts and optimize agrarian processes.

Results

The process of developing a method of step-by-step yield forecasting includes several key steps, starting from theoretical research and ending with the construction of grain crop productivity models for yield forecasting. The method developed by the authors is based on the use of modern computer technologies and includes the following key steps:

1. Creation of information support. At this stage, agrometeorological indicators are collected and calculated, preliminary statistical analysis is performed, and the reliability of experimental data, which are necessary for the identification of the algorithm, is assessed.

2. Technology of selecting years-analogs. This stage includes clustering of data, selection of optimal partitioning into clusters and formation of a class of year-analogues for the year under study.

3. Modeling of weather scenarios. Modeling is carried out on the basis of the principle of similarity and stochastic methods, which allows taking into account different weather scenarios.

4. Software development. At this stage, a set of instrumental software tools is created to ensure the implementation of the developed algorithm.

5. Yield forecasting. On the basis of the developed models of productivity of grain crops the forecast of yield is performed.

6. Yield estimation. Forecasted data are checked with the help of simulation and modeling complex, which allows to assess the accuracy and reliability of the obtained forecasts.

Consideration is given to the implementation of individual stages in more detail:

1. Creation of information support. At this stage, the methodology of preliminary statistical analysis and reliability assessment of experimental agrometeorological data is developed. An important role is played by processing and analysing the accumulated data, which is especially important, given their heterogeneity, significant variations and connectivity. Particular attention is paid to the study of multivariate series of grain yields to identify cyclical properties and non-stationarity of time series.

2. The technology of selecting years-analogues. This stage is based on the principle of similarity and classification of agrometeorological factors. The initial step consists in the classification of objects on the basis of the analysis of signs or indicators characterising these objects and their attribution to a certain class.

Let the set Ω of all studied objects:

$$\Omega = \left\{(X^r, Y): X^r = \left\{x_{ij}^r\right\}, Y = \{y_r\}, r = \overline{1,n}; i = \overline{1,t}; j = \overline{1,m}; x_{ij} \in R, y_r \in R\right\}$$

it is necessary to form non-intersecting subsets $Ak \in \Omega$ – classes of similar objects - year-analogues of the species according to the decisive rule S:

$$A_k = \{(X^N, Y_N): X^N \in X^r, Y_N \in Y, N = \overline{1, a_k}, a_k \leq n\}. \quad (1)$$

Here $X^r = \{x_{ij}^r\}$ is a matrix of values of agrometeorological factors of size $t \times m$, determined for each set of observed features affecting the vector y_r; y_r vector of values of actual yield; n – number of years under study; m – number of available agrometeorological characteristics; t – discrete moment of time; k – number of formed classes possessing a set of factors close to each other in terms of influence on the resultant feature of the object y_N; a_k – number of years-analogues in the corresponding class.

Under the decisive rule S we will understand providing the extremum of the functional $F(A_k) \rightarrow min(max)$ - a measure of homogeneity of objects, where $F{:}X{*}X \rightarrow$. In the case of dependent features, the Mahalanobis distance is taken as a measure of homogeneity of objects:

$$D_m = \sqrt{\left(x_i' - \bar{x}\right)^T \Sigma^{-1} \left(x_i' - \bar{x}\right)}, \qquad i = \overline{1, t}, \tag{2}$$

where x'_i – multivariate feature vector; Σ – correlation matrix; $\bar{x}$ – class (cluster) centre.

The objects of classification are years, and agrometeorological factors act as features or indicators characterising these objects.

In order to compare several typifications and select the optimal one, a criterion – a numerical measure of classification quality - is needed. The quality of classification can be assessed as an indicator:

$$\mathrm{K} = \mathrm{K}_{\mathrm{w}} / \mathrm{K}_{\mathrm{b}}, \tag{3}$$

where $K_w = \frac{2}{k*(k-1)} \Sigma_{i=1}^{k} \overline{d_{\iota\iota}},\ K_b = \frac{1}{k} \Sigma_{i=1}^{k} \Sigma_{j=i+1}^{k} \overline{d_{\iota J}}$.

Here K_w – intra-cluster and K_b – inter-cluster distances; k – number of clusters; $\overline{d_{\iota\iota}}$ – average distance between points within the *i-th* cluster; $\overline{d_{\iota J}}$ – average distance between pairs of points of the *i-th* and *j-th* clusters.

As a result of this procedure, first, a training sample $X_{l1}, X_{l2}, \ldots, X_{ln}, l = 1, k,$ *is* formed, where X_i is a vector of multivariate observations; k is the total number of classes identified in the process of preliminary typologisation; and it is known about the observations X_{li} that they all characterise objects belonging to the *l-th* class. Second, a classifier (discriminant function) of each classifiable object, given by the values of its descriptive features, is constructed.

3. Modelling of weather scenarios. At the third stage of step-by-step forecasting of grain crop yields, the system of modelling weather scenarios based on the use of two approaches: the principle of similarity and stochastic methods is considered.

Modelling weather scenarios according to the principle of similarity is the task of the second stage of the technology of determining year-analogues, which consists in selecting from all subsets of $Ak \in \Omega$ the class of objects A_{k_0}, that best matches (according to certain criteria) the new element:

$$X^{n+1} = \{x_{i_0 j}^{n+1}\}, i_0 = \overline{1, l_0}; j = \overline{1, m}, l_0 < l; \tag{4}$$

where l is usually taken as 365 days; l_0 is the number of the day from which the weather scenario is modelled. In order to assess the impact of weather conditions on crop formation, it is necessary to classify the situation in a certain period of time on the basis of the study of a set of agrometeorological parameters, taking into account its impact on the state of plants, more precisely, on yield. As a consequence, the formed class of objects A_{k_0} forms an ensemble of possible realisations of weather conditions, which can be described by vector $G = \{G^0, G^1, \ldots, G^{k_0}\}$ where G_0 – the year under study; where $G^1, \ldots, G^{k_0}$ – years-analogues.

Then the forecast $\hat{G}^0(t + l'), l' = \overline{l_0, l}$ scenario for the year under study can be constructed using the optimisation procedure:

$$O = \Sigma_{t=1}^{l_0-1}\left[\hat{G}^0(t) - G^0(t)\right]^2 \rightarrow min, \tag{5}$$

where $\hat{G}^0(t) = \Sigma_{i=1}^{k_0}\alpha_i G^i, \Sigma_{i=1}^{k_0}\alpha_i = 1,\ \alpha_i \geq 0$ similarity parameters.

The use of this method assumes that when the model is running, actual weather data are input to the model up to the point at which forecasting starts. In order to match (smooth) the actual data and the data of the year-analogues, deviations of the actual data at the forecast date and deviations of the data of the year-analogues are recorded. These deviations are smoothed using a first-order dynamic link that filters out fluctuations.

The second method is stochastic (probabilistic) in nature. The source of new weather realisations is the so-called weather generator, in which daily meteorological data such as maximum and minimum air temperatures, minimum air humidity, wind speed, precipitation and solar radiation attenuation coefficient are modelled as a multivariate random non-stationary process. This approach is based on the autoregressive model introduced by Richardson [14].

The stochastic weather generator simulates synthetic daily series of meteorological elements with statistical characteristics close to those in historical data of actual weather realisations for 20-30 past years.

The result of weather scenario modelling is a 'fan' of possible trajectories of crop formation and corresponding sets of possible values of potential probabilistic forecast of productivity resources. The notion of weather scenario generation does not mean that the result will be a realisation of weather conditions ever encountered in a given area.

The main purpose of weather scenario modelling is that this procedure, being used as input data for mathematical models of the productivity process, will produce a result that solves the problem of forecasting this or that parameter of the productivity process. Thus, the meteorological situations obtained as a result of modelling are joined to the available actual meteorological conditions, forming a complete set of daily input data for mathematical models of grain crop productivity.

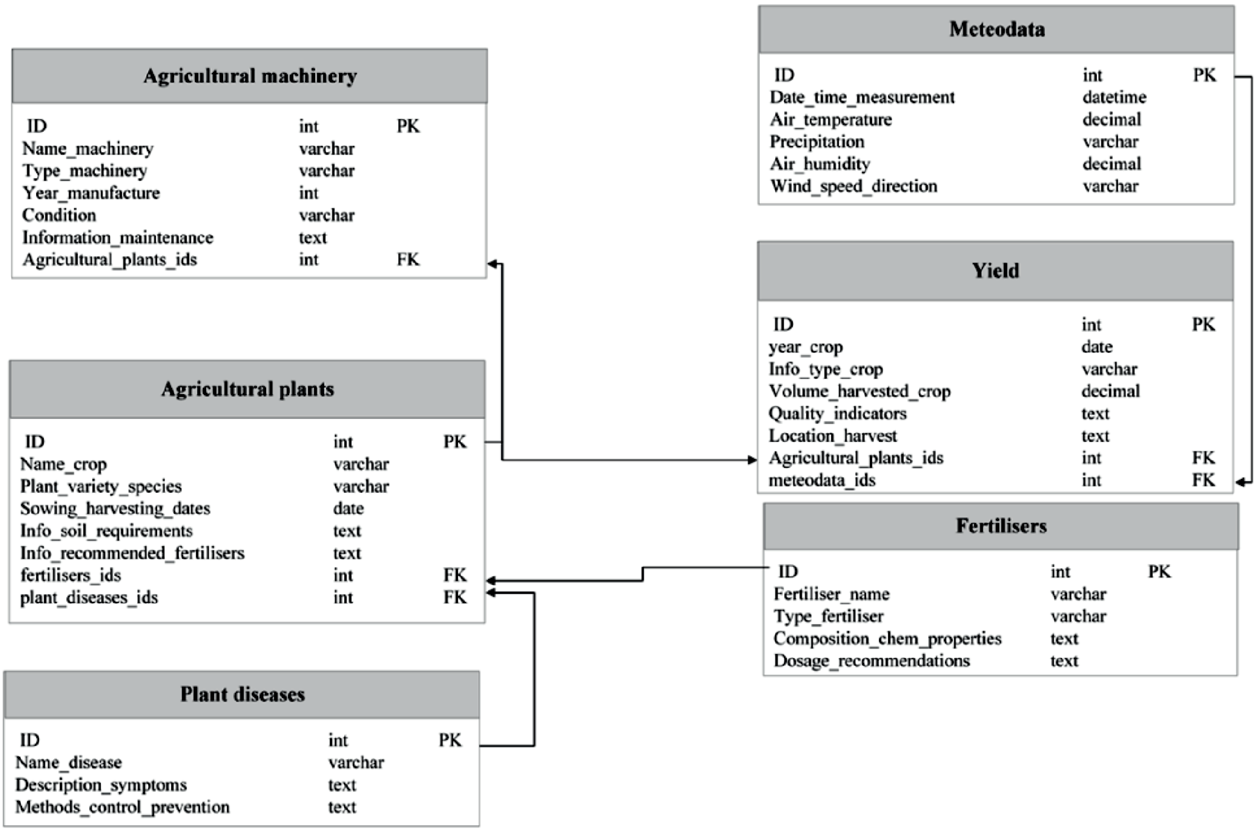
4. Development of a set of instrumental software tools. A special role in the system of operational agrometeorological support of agricultural production is assigned to information and prognostic systems of processing and analysis of agrometeorological information, which allows, as a result of the generalisation of this information, to carry out the forecast of agrometeorological factors and grain yields. In this regard, there was a need to develop a software package for processing experimental agrometeorological data (and information support models of productivity of grain crops. This software complex includes:

1) a database of experimental data;

2) block of formation and primary processing of agrometeorological factors;

3) a block of implementation of technology for determining the letanalogues (application of the principle of similarity for the formation of weather scenarios using cluster and discriminant analyses).

For computer implementation of the complex and relational model of the database the Java computing platform was chosen. The system developed an interface that allows exporting and importing to external sources, editing and forming data in the database. The experimental data base is presented as a hierarchically organised set of control and subordinate data tables (Figure 1).



Agricultural machinery		
ID	int	PK
Name_machinery	varchar	
Type_machinery	varchar	
Year_manufacture	int	
Condition	varchar	
Information_maintenance	text	
Agricultural_plants_ids	int	FK

Meteodata		
ID	int	PK
Date_time_measurement	datetime	
Air_temperature	decimal	
Precipitation	varchar	
Air_humidity	decimal	
Wind_speed_direction	varchar	

Yield		
ID	int	PK
year_crop	date	
Info_type_crop	varchar	
Volume_harvested_crop	decimal	
Quality_indicators	text	
Location_harvest	text	
Agricultural_plants_ids	int	FK
meteodata_ids	int	FK

Agricultural plants		
ID	int	PK
Name_crop	varchar	
Plant_variety_species	varchar	
Sowing_harvesting_dates	date	
Info_soil_requirements	text	
Info_recommended_fertilisers	text	
fertilisers_ids	int	FK
plant_diseases_ids	int	FK

Fertilisers		
ID	int	PK
Fertiliser_name	varchar	
Type_fertiliser	varchar	
Composition_chem_properties	text	
Dosage_recommendations	text	

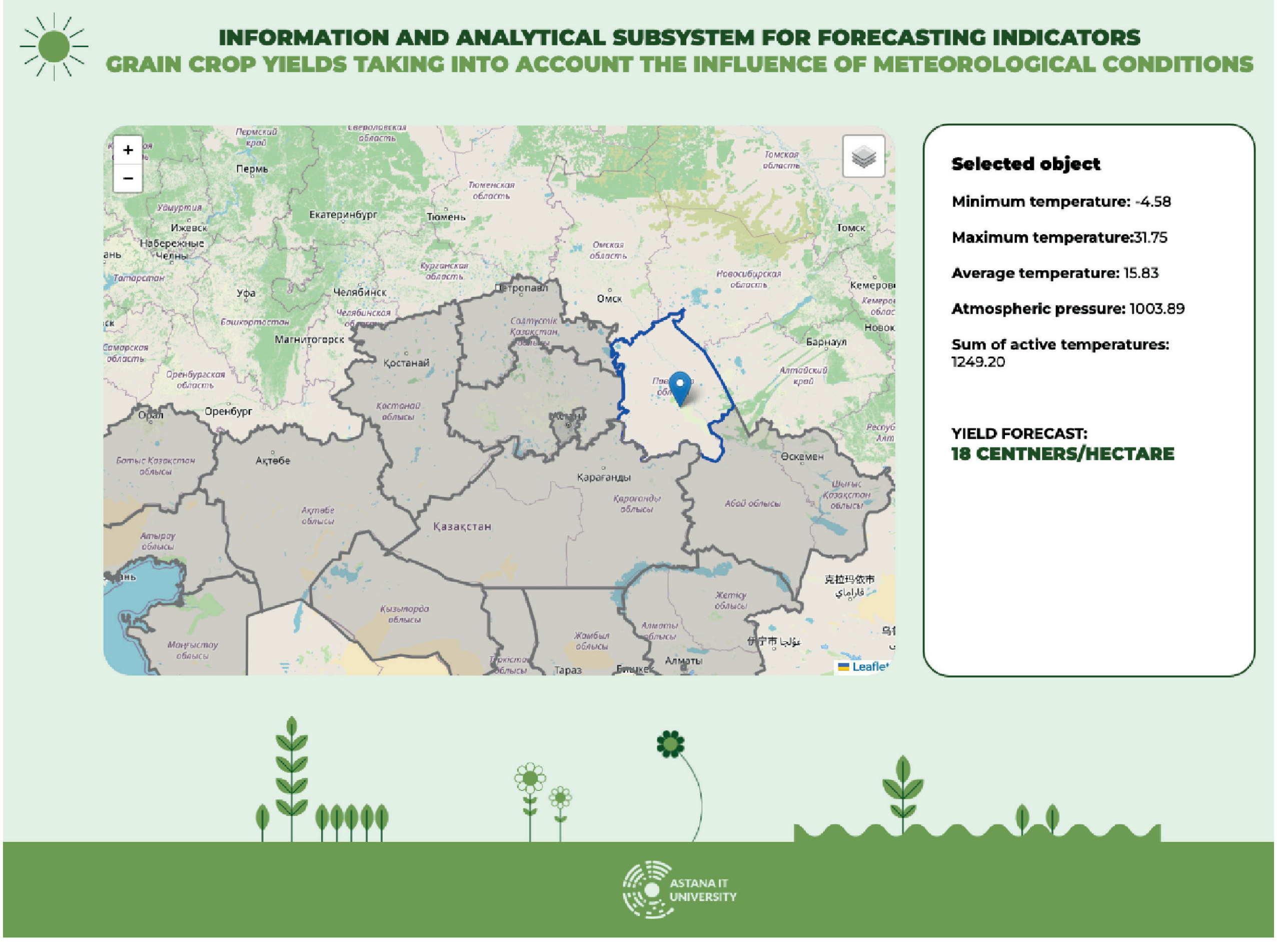
Plant diseases		
ID	int	PK
Name_disease	varchar	
Description_symptoms	text	
Methods_control_prevention	text	

Figure 1. Schematic of the conceptual database model

The information model is designed with the help of a specialised SQLite library, which allows full use of modern tools of database tools according to SQL-standard.

The architecture of this information system (Figure 2) is divided into three levels: the level of the user interface of the precision farming information system for managing agricultural activities, the level of data processing and management, and the level of the database.

The level of data processing and management consists of the module of forecasting the influence of meteorological data on the efficiency of growing agricultural plants – weather station, data collection sensors, forecasting model. The user interface level is an information panel displaying all indicators. It is integrated with databases for data visualisation.



INFORMATION AND ANALYTICAL SUBSYSTEM FOR FORECASTING INDICATORS
GRAIN CROP YIELDS TAKING INTO ACCOUNT THE INFLUENCE OF METEOROLOGICAL CONDITIONS

Selected object

Minimum temperature: -4.58

Maximum temperature:31.75

Average temperature: 15.83

Atmospheric pressure: 1003.89

Sum of active temperatures: 1249.20

YIELD FORECAST:
18 CENTNERS/HECTARE

Figure 2. System interface

In this study, modern tools and libraries were used to efficiently perform tasks related to data analysis and model building for yield prediction.

Pandas and NumPy are key tools for working with data in Python. Pandas provides powerful data structures, such as DataFrames, that allow you to easily manipulate, filter, group, and aggregate data. This library is particularly useful for time series processing, including performing time interpolation operations, resampling data, and calculating moving averages. NumPy, on the other hand, provides basic operations on multivariate data sets and provides a wide range of mathematical functions that can be used for statistical analysis of time series.

In this study, Pandas was used for agrometeorological data preprocessing, including data cleaning, format conversion and temporal referencing, which is a critical step to ensure correct analysis and model building. NumPy has been extensively used to perform mathematical operations such as normalising data, calculating statistics and processing multidimensional arrays.

Two popular frameworks, TensorFlow and PyTorch, have been used to build machine learning models, especially recurrent neural networks (RNNs). These frameworks provide flexible and scalable tools for building, training and evaluating deep learning models.

TensorFlow is developed by Google and is a powerful tool for building machine learning models, including deep neural networks. Its broad capabilities and support for large datasets make TensorFlow a suitable choice for complex forecasting tasks such as time series modelling of agro-meteorological indicators.

PyTorch, developed by Facebook, offers a more intuitive and user-friendly interface for creating and training models. It is widely used in research projects due to its flexibility and support for dynamic computational graphs, which facilitates the implementation of complex architectures such as recurrent neural networks.

In this research, the TensorFlow and PyTorch frameworks were used to develop and train recurrent neural network models that provide weather-aware yield time series forecasting. These models demonstrate high accuracy and generalisability based on time series analysis.

Two recurrent neural network architectures, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), were applied to solve the yield forecasting problems. These architectures are widely used for time series analysis and forecasting due to their ability to take into account the temporal dependence of the data.

LSTM is one of the most popular RNN architectures specifically designed to address the vanishing gradient problem, making it particularly useful for problems where long-term memory is important. LSTM uses memory "cells" that can remember information for long periods of time, allowing models to capture and analyze long-term dependencies between data.

GRU is a simplified version of LSTM that is also capable of remembering long-term dependencies, but with fewer parameters, making it more efficient in terms of computational resources. GRU is used in tasks where speed of learning and minimization of computational cost are important.

In this study, LSTM and GRU models were applied to build yield forecasts based on time series of agrometeorological data. Both types of models were trained on the data using TensorFlow and PyTorch, which achieved high accuracy of the forecasts.

Python programming language was used to develop all models and perform data analysis. Python is a staple tool in scientific research due to its simplicity, code readability, and extensive ecosystem of libraries and frameworks for data processing and machine learning. Its popularity in the machine learning research and development community stems from its ability to quickly implement complex algorithms and its ease of working with large amounts of data.

An important part of the process of data analysis and presentation of results is data visualization. In this study, Matplotlib and Seaborn libraries were used to visualize data and prediction results.

Matplotlib is a standard tool for creating graphs and charts in Python, providing rich possibilities for customizing visualizations. In the study, Matplotlib was used to generate time series, histograms, and other graphs showing the dynamics of agrometeorological data and yield prediction results.

Seaborn extends the capabilities of Matplotlib by providing higher-level tools for building complex statistical visualizations. In this study, Seaborn was used to create heat maps, correlation matrices, and visualize data distributions, contributing to a better understanding of data structure and relationships between agrometeorological indicators.

To evaluate the performance of the developed forecasting methods, validation was performed using historical data. Cross-validation methods were used in the validation process to provide a more reliable assessment of the accuracy of the models, minimising the probability of overfitting.

Cross-validation involves several steps:

1. Data partitioning. The original data set is divided into k subsets (folds), where each subset is used in turn to test the model, while the other subsets are used to train it. For example, if k=5, the data will be split into 5 folds.

2. Training and testing. At each of the k iterations, the model is trained on k-1 folds and tested on the one remaining fold. This process is repeated for all folds, and each piece of data serves for testing exactly once.

3. Collection of results. After all iterations are completed, the test results are collected, including the values of metrics such as mean square error (MSE) and coefficient of determination (R^2) that were described earlier.

4. Overall performance evaluation. To obtain an overall performance evaluation of the model, the average of the metrics across all k test sets is calculated. This provides a more accurate representation of the model's ability to generalise to new data.

Cross-validation methods provide a reliable assessment of forecast quality and help to determine how well the model will perform on new, unseen data. The validation results confirm that the proposed methods are effective for forecasting grain yields taking into account the influence of meteorological conditions.

Two main metrics were used in this study to evaluate the accuracy of the predictions: mean square error (MSE) and coefficient of determination (R^2).

Mean Square Error (MSE). The mean square error is an important metric that estimates the average error between predicted values and actual values. MSE is calculated using the following formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2, \tag{6}$$

where:

n – number of observations,

y_i – actual yield value for the i-th observation,

$\widehat{y_i}$ – predicted yield value for the i-th observation.

MSE shows how much the predicted values deviate from the actual values. The smaller the MSE value, the more accurate the predictions are.

2. Coefficient of determination (R^2). The coefficient of determination R^2 reflects the proportion of variation in the dependent variable that is explained by the independent variables in the model. It is calculated by the formula:

$$R^2 = \frac{SS_{res}}{SS_{tot}}, \tag{7}$$

where:

– $SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ – sum of squares of residuals,

– $SS_{tot} = \sum_{i=1}^{n}(y_i - \hat{y})^2$ – total sum of squares, where $\hat{y}$ – average value of actual data. The experimental results are presented in Table 1. The table shows the MSE and R^2 values for all methods.

Table 1. Experimental results

Method	MSE	R²
Gradient bousting	120.5	0.85
Recurrent neural networks	115.3	0.87
Holt-Winters method	130.2	0.82
Linear regression	150.4	0.75
Simple medium	180.6	0,60

The R^2 coefficient takes values between 0 and 1, where a value close to 1 indicates that the model explains the variation in the data well. An R^2 value of 0 indicates that the model does not explain the variation, while a value greater than 0.5 is generally considered satisfactory for predictive models.

The experimental results show that recurrent neural networks (LSTM) and gradient bousting showed the lowest MSE values and the highest R^2 values, indicating that the predictions are highly accurate. Holt-Winters methods and linear regression showed less satisfactory results, and simple mean was the least accurate of all the models tested. The experiment confirmed that the developed yield prediction methods are more effective than the baseline models. These results emphasise the importance of applying modern machine learning techniques in the agricultural sector for more accurate forecasting of grain yields. The use of these tools and technologies not only enabled the construction of accurate yield forecasting models, but also provided visualisation of the results, which is an important step in agronomic decision-making.

Conclusion

This paper presents a method developed by the authors, which includes the technology of forecasting agrometeorological factors affecting the productivity of grain crops. An important part of the study was the mathematical formalisation of the principle of similarity, which allows the effective identification of years similar to the study period based on agrometeorological indicators. This ensured the possibility of creating realistic weather scenarios based on the data on analogous years and the application of weather data generators.

The application of the proposed technology for determining the summer-analogues in combination with methods of modelling the productivity of grain crops allows for more accurate forecasting of yields. This is especially important for agricultural enterprises and farmers, as it allows to estimate potential yields in advance and to plan agrotechnical measures taking into account possible changes in weather conditions.

The experimental results show that recurrent neural networks (LSTM) and gradient bousting demonstrated the lowest MSE values and the highest R^2 values, indicating that the predictions are highly accurate. Holt-Winters methods and linear regression showed less satisfactory results, and simple mean was the least accurate of all the models tested. Thus, the proposed algorithm can become an effective tool for decision support in agriculture, providing reliable forecasting of grain crop yields based on a comprehensive analysis of agrometeorological factors. The introduction of this technology into agricultural practice can contribute to increasing the resilience of agriculture to climatic changes and optimising the management of production processes.

Acknowledgment

The paper was written within the state order for the implementation of the scientific program under the budget program of the Republic of Kazakhstan 217 “Development of Science”, subprogram 101 “Program-targeted funding of the scientific and/or technical activity at the expense of the national budget” on the theme: AP19678730 Development of precision farming information technology for agricultural management using “The Internet of Things”.

References

[1] Akorda. (2023). President Kassym-Jomart Tokayev’s State of the Nation Address “Economic course of a Just Kazakhstan”. Retrieved from https://www.akorda.kz/ru/poslanie-glavy-gosudarstva-kasym-zhomarta-tokaeva-narodu-kazahstana-ekonomicheskiy-kurs-spravedlivogo-kazahstana-18588

[2] Concept for the development of the agro-industrial complex of the Republic of Kazakhstan for 2021-2030. (2021). Retrieved from https://adilet.zan.kz/rus/docs/P2100000960

[3] Zdravković, M., et al. (2018). QSPR in forensic analysis – The prediction of retention time of pesticide residues based on the Monte Carlo method. *Talanta*, 178, 656-662. https://doi.org/10.1016/j.talanta.2017.09.064

[4] Kavakci, G., Cicekdag, B., & Ertekin, S. (2024). Time Series Prediction of Solar Power Generation Using Trend Decomposition. *Energy Technology*, 12(2), 2300914. https://doi.org/10.1002/ente.202300914

[5] Tran, Q.T., Hao, L., & Trinh, Q.K. (2019). Cellular network traffic prediction using exponential smoothing methods. *Journal of Information and Communication Technology*, 18(1), 1-18. https://doi.org/10.32890/jict2019.18.1.1

[6] Ookura, S., & Mori, H. (2020). An efficient method for wind power generation forecasting by LSTM in consideration of overfitting prevention. *IFAC-PapersOnLine*, 53(2), 12169-12174. https://doi.org/10.1016/j.ifacol.2020.12.1008

[7] Ma, X., et al. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, 54, 187-197. https://doi.org/10.1016/j.trc.2015.03.014

[8] Peña-Gallardo, R., & Medina-Rios, A. (2020). A comparison of deep learning methods for wind speed forecasting. *2020 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, 1-6 https://doi.org/10.1109/ropec50909.2020.9258673

[9] Sarbasov, Y.A. (2015). Application of recurrent neural networks in data analysis and financial market forecasting. *Banks of Kazakhstan*, 4, 14-16.

[10] Albrecht, V. S. (2016). Neural networks as a method for predicting oil prices. *Banks of Kazakhstan*, 3, 2-7. – ISSN 2307-0323

[11] Mimenbayeva, A., et al. (2023). Neural network model of soil moisture forecast for North Kazakhstan region. *Scientific Journal of Astana IT University*, 149-159. https://doi.org/10.37943/15TYEQ8191

[12] Aubakirova, G., et al. (2022). Application of artificial neural network for wheat yield forecasting. *Eastern-European Journal of Enterprise Technologies*, 117(4). https://doi.org/10.15587/1729-4061.2022.259653

[13] Kumar, D., et al. (2022). Wheat Crop Yield Prediction Using Machine Learning. *2022 International Conference on Data Analytics for Business and Industry (ICDABI)*, 433-437. https://doi.org/10.3390/drones8070284

[14] Richardson, C.W., & Wright, D.A. (1984). WGEN: A model for generating daily weather variables. *US Department of Agriculture, Agricultural Research Service, ARS-8*, Washington DC, 83 p.