A set of data on retrospective grain yield for neural network modeling

A F Rogachev¹,²,*, E V Melikhova¹ and I S Belousov¹

¹Volgograd State Agricultural University, 26, Universitetskiy Avenue, Volgograd, 400002, Russian Federation
²Volgograd State Technical University, 28, Lenina Avenue, Volgograd, 400005, Russian Federation

*E-mail: rafr@mail.ru

Abstract. Theoretical, methodological and applied aspects of creating a database (DB) of long-term retrospective information about the yield of grain crops are considered. The database is designed for statistical processing and modeling of multi-year data. Statistical information on annual levels of grain yield is the basis for planning, forecasting, management and optimization of agricultural production. To implement these tasks and their information support, we use design methods and models that allow us to build relational databases that provide statistical and neural network modeling of interannual variability of yield levels. The data storage format (*.csv) is justified, which provides in-depth processing and statistical analysis using built-in Python libraries. The review statistical analysis of productivity on the example of grain crops grouped by annual intervals is presented, and their features are revealed.

1. Introduction

Specialized databases of retrospective crop yields containing reliable information about long-term values of their yields are important for predicting gross collections used for agricultural production management. Management of agricultural production requires reliable knowledge of the forecast level of crop yields to assess their gross collections. Multi-year time series (TS) of retrospective data on annual yield levels are the basis for planning, forecasting and optimizing agricultural production [1,3,17,12]. Information support for solving forecasting problems is a set of databases (DB) of retrospective information that allow automated modeling, analysis and forecasting of yield levels for the future. The results of research on the problems of forecasting the yield of various groups of crops are published in the works of Russian and foreign scientists A. Gerasimov, I. Vintzenko, E. Popova, V. Perepelitsa, B. Shulika, A. Porvan, O. Vysotska, A. Nekos, A. Zhemerov [14, 11, 15]. The research of A. Gagarin, A. Pavlova, V. Kalichkin and others is devoted to the methodology, theory and practice of creating specialized databases, mainly of relational type [1, 2, 7, 8].

Various aspects of the creation and functioning of information and analytical tools aimed at supporting and justifying management decisions in the agricultural economy, based on specialized information systems (is) and corresponding databases, are considered in the works of A. Aksenov, A. Altybayev, A. Gerasimov, T. Bashkatova [10,13,15,16]. In particular, the work performed at the IBM research center proposed and tested an approach to the joint use of yield statistics and space
photography data for training neural networks, but the results are limited by the natural and climatic conditions of the regions of North America [18]. Taking into account internal patterns of interannual changes in yield levels for Russian conditions that are often characterized by a lack of humidity, it will allow creating a system for forecasting yield taking into account natural and climatic, technical and technological, and organizational and economic conditions. Such an information and analytical system and a specialized database can be based on modern methods of machine learning and artificial neural networks (ANN).

2. Materials and methods
The creation of specialized databases based on the system approach is based on specific methods-functional-morphological and autocorrelation analysis, conceptual design and takes into account the use of neural network technologies [5,6,10]. The database ontology and relational structure were formed taking into account its functional purpose – forecasting parameters and optimizing agricultural production [7,9]. To save data in a subject-oriented database, as well as subsequent statistical processing and mathematical modeling, the CSV data storage format was justified. For the purpose of subsequent analysis of retrospective yield data, the preliminary preparation and transformation of the source information was provided, providing computer processing of the target information using specialized ML-libraries of machine learning in Python V. 3.7 [10].

3. Results and discussions
To create the database, the data storage format «*.csv» (Comma-Separated Values) was selected, which is text files for representing table data. The CSV file format allows you to use universal MS Excel files as source files and then edit them using various tools, such as text editors. This solution makes it easier to load the generated database into the Python environment to use its statistical processing libraries, such as Pandas. A typical picture of the values of yield levels, for example, of grain crops collected according to official data from ROSSTAT for natural and climatic conditions in the Volgograd region, is shown in the diagram figure on 1.

![Figure 1. Average grain yield values, Volgograd region.](image)
The presented data are characterized by a pronounced positive trend and a significant spread of values relative to the linear approximation with a coefficient of variation exceeding 30%. The summary results of the review statistical analysis of yield levels for 69 years, grouped by annual intervals of 1950...2018, are shown in Fig. 2 and 3. Overview statistical analysis was performed using the Seaborn library of the Python programming environment. The yield distribution is characterized by a significant variation relative to the average yield level (12.8 t / ha) with a coefficient of variation $v = 0.45$. Note that this value is below the 50% quantile of yield levels, which is 13.1 t / ha (figure 2).

Analysis of yield levels grouped by annual intervals using the Seaborn library visualization tools makes it possible to more clearly assess their variability over a nearly 70-year period (Fig. 2).

The analysis of the configuration of interval values of the spread of yield levels shows some stabilization of the spread of yield in the period of 1984 ... 2001. At the same time, in recent years (2001 ... 2018), it is possible to note individual emissions falling outside the typical normal distribution range of ± 3σ, characterized by increased relative to the average values of yields reaching 26.9 t/ha.

Some aspects of the internal patterns of yield variation are presented in a two-dimensional diagram of their paired interactions, where the same annual categories are also highlighted in color (figure 3). It is possible to note a significant intersection of the levels of analyzed yields by annual groups, which may require the use of more advanced tools to study and account for their internal patterns.

![Figure 2. Overview statistical analysis of yield levels grouped by annual intervals.](image-url)

Along with the BP of yields of various winter and spring crops (wheat, rye, corn, barley, triticale, etc.), the database formed for the possibility of multi-factor research should include natural and climatic, as well as organizational and technological factors. Implementation of such a complex database will require collecting the listed factors by aggregated groups, as well as forming it as a relational structure that includes several linked tables that are combined by key fields [13].

Taking into account the identified internal patterns of interannual changes in yield levels stored in the database under development will allow us to develop a more reasonable system of mathematical modeling and forecasting.
Figure 3. Diagram of paired interactions of the analyzed factors.

As the latter, it is assumed to use the «Years» fields in Pandas dataframes that characterize the research years. The number of tables that make up a complex database will be determined by the number of one-dimensional factors–indicators (agriculture, soil type, temperature, precipitation, etc.). These factors determine the structure of the relational database being developed.

4. Conclusion

1. Based on the conducted research, a relational database structure is recommended, including a set of one-dimensional factors – indicators (agriculture, soil type, temperature, precipitation, etc.), combined by key fields-years of observations.

2. In order to generate information about the long-term retrospective yield of grain crops and use it for statistical analysis and mathematical modeling, the database should include the fields listed above.

3. For subsequent computer analysis and mathematical modeling, data storage is justified in the *.csv (Comma-Separated Values) format, which is a text format for representing tabular data. This format provides in-depth processing, analysis, and visualization using specialized Python eco-environment libraries.

5. Acknowledgments
The publication has been prepared with the financial support of the RFBR and the Administration of the Volgograd region under project No. 19-416-340014.

References
[1] Aksenov A 2019 Analysis of intelligent decision support systems in agriculture *Electrotechnology and electrical equipment in agriculture* 3(36) 46-51
[2] Borodychev V V 1999 Method of cultivation of soy for grain on irrigated lands R. U. Patent No. 2132600

[3] Bendik N and Ivanyo Ya 2018 Conceptual model of data storage for effective agriculture in the region Climate, ecology, agriculture of Eurasia Materials of the VII international scientific and practical conference 159-166

[4] Veklenko V, Petrenko N, Pashkov V and Eremenko O 2013 Designing a database of models for optimal planning of agricultural production Bulletin of the Kursk state agricultural Academy 1 33-36

[5] Gagarin G and Rogachev A 2018 Application of artificial neural networks for crop yield forecasting based on cross-regional data analysis Proceedings of the Lower Volga agrodiversity complex: Science and higher professional education 2 339-346

[6] Rogachev A and Shubnov M 2012 Estimation of the forecast level of crop yield on the basis of neural network models of dynamics Proceedings of the Lower Volga agrodiversity complex: Science and higher professional education 4 226-231

[7] Gerasimov A and Bashkatova T 2016 Information and analytical tools for substantiating managerial decisions in the economy of a traditionally agrarian region Bulletin of the Stavropol agro-industrial complex 20(1) 8-12

[8] Glazunov G, Afonchenko N and Sanzharov A 2017 Structure of the database of natural resource potential of agricultural landscapes Bulletin of the Kursk State Agricultural Academy 8 6-11

[9] Goryukhina E and Litvinova L 2013 Application of information technologies using databases in the agricultural sector of the agro-industrial complex Economics and management in the agricultural sector of the agro-industrial complex: problems and solutions. Collection of scientific 56-59

[10] Egorova I, Gagarin A, Kuznetsov S, Simonov A and Velikanov V 2019 Successful commercialization of innovations as a basis of development of modern human society. Advances in Intelligent Systems and Computing 726 1156-1162

[11] Igoshin A, Cheremukhin A and Savrukov N 2019 Building a mathematical model of long-term dynamics of grain yield Bulletin of NGIEI 3 (94) 74-83

[12] Komratova A M 2018 A predictive model of the effects of external imperatives of the natural environment on grain yield in the South of Russia Bulletin of the Adygeya state University. Series 5: Economics 3 (225) 129-136

[13] Melikhova E 2017 Modeling and justification of resource-saving parameters of drip irrigation in the cultivation of root crops: monograph (Volgograd: Volgograd state agricultural university) 112 p

[14] Lysenko S 2019 Forecasting crop yields based on satellite monitoring of carbon dynamics in terrestrial ecosystems Exploring the Earth from space 4 48-59

[15] Rogachev A 2011 Parametrization of econometric dependencies by the method of least modules Management of economic systems: electronic scientific journal 3 0421100034

[16] Rogachev A 2018 System analysis and forecasting of yield time series based on autocorrelation functions and neural network technologies Proceedings of the Lower Volga agrodiversity complex: Science and higher professional education 3 (51) 309-16

[17] Shulika B, Porvan A, Vysotska O, Nekos A and Zhemerov A 2017 Control over grape yield in the north-eastern region of Ukraine using mathematical modeling Eastern-European Journal of Enterprise Technologies 2/3(86) 51-59

[18] Russello H 2018 Convolutional neural networks for crop yield prediction using satellite images. IBM Center for Advanced Studies. https://esc.fnwi.uva.nl/thesis/cenatraal/files/f1570224447.pdf

[19] Tokarev K et al 2019 Economic and mathematical modelling of regional market of grain crops. IOP Conf. Ser.: Earth Environ. Sci. 341 012212