Application of statistical methods for the agriculture complex using the example of sugar beet production

Lyudmila Korobova* [1], *Irina* Tolstova[1], *Maxim* Ivliev[1], *Natalia* Datsenko[1], *Sergey* Chikunov[1]

[1]Voronezh State University of Engineering Technologies, Voronezh, Russia

Abstract. The paper considers the issue of applying statistical methods that allow processing large technological data. At the initial stage, the technological process of beet sugar production was analyzed. The indicators (more than 40 parameters) were taken, and those that have the greatest influence on the technological process were selected. The values of the main parameters determine the varietal yield (sugar grade). The statistical analysis method was chosen as the main research method. Characteristic samples of 40 parameters were considered, the sample size was 179 values. Two quality parameters that determine the sugar grade were selected. Correlations of the characteristic initial parameters with the parameters of varietal sugar were checked. A mathematical processor was chosen as a tool for conducting the analysis. The Excel program contains almost all the functions and techniques that are missing in application development tools and much more. Statistical characteristics of some of the experimental samples under study are presented. And based on the obtained models, a conclusion was made that the third-order power model adequately describes the experimental data.

1. Introduction

Currently, personal computers (PCs) are increasingly being introduced into everyday human activities. The devices performance is increasing, while the cost is decreasing. All this contributes to the personal computer applications expansion. Modern software is being developed and implemented for various human activity fields. At the same time, there is no need to maintain a large staff, since software can sometimes replace the specialists entire departments work. Sometimes it is not necessary to have the programmer knowledge to work with modern PCs. However, modern software doesn't yet have the human intelligence, so software can only do what a human has programmed [1].

The promising area in the information technology field is the creation of analytical systems that free people from routine, sometimes repetitive operations for processing large information amounts. Such systems can make predictions and make recommendations on the current situation [2].

Over the last fifty years, many studies have been conducted to improve the quality of production processes. A process can be considered controlled if it is possible to obtain information about changes in its main parameters. The information obtained is then used to control the process in order to prevent failures and achieve the best results.

Today, leading companies are looking for new sources of competitive advantage and understand that the vast amount of data collected during their manufacturing operations provides excellent information to improve the development of quality products and process performance. The quality of manufactured products must comply with the accepted Russian standards (RS) or technical conditions (TU) of enterprises or industries.

In addition, companies are under increasing competitive pressure. In order to improve the sustainability and competitiveness of their products and processes, a deep understanding of the physics of the processes is required. This can be achieved by using a more powerful analytical tool [2, 3].

The desire to achieve high profits forces industrial and processing companies to constantly look for ways to reduce costs. They are starting processing processes that are close to critical limits according to technological regulations, using cheaper components where possible, reducing energy consumption and minimizing waste and rework costs.

There are approaches to control and diagnostics of technological processes (using sugar beet production as an example) [4, 5]:

- technological process control and diagnostics;
- sugar beet production automation;
- sugar beet production information system modeling.

Decomposition and structural-parametric analysis of technological processes (using sugar beet production as an example) and tools for their control and diagnostics include the following:

- decomposition and structural-parametric analysis of technological processes of sugar beet production;
- structural modeling of the technological process of sugar beet production;

* Corresponding author: lyudmila_korobova@mail.ru

- structural synthesis of the model of the information system for diagnostics and control of the technological process;
- modeling of the functioning of the information system for monitoring and diagnostics of technological processes (using sugar beet production as an example);
- analysis of the stationarity of technological processes of sugar beet production using the nuclei of conflict, cooperation and indifference.

2. Materials and methods

Systematic collection of data and the ability to translate it into a format suitable for analysis was another common obstacle to the use of multivariate methods in the past. Today, this is less of a problem because most processes are instrumental, and complex control systems are widely used. In fact, the problem now is that so much data is collected that it is becoming increasingly difficult to cut through the vast amount of information to find underlying patterns, which encourages the use of multivariate methods [4, 5].

Most technological processes are very multidimensional in nature due to complex reactions. There are a large number of variables that are usually very interactive [6, 7]. To fully understand the process dynamic, complex systems require experiments series. The experiments produce descriptive statistics. The experimental data are used to quantify the process characteristics for population assessment. A convenient measure for understanding process capability is 6σ-measure. If the experimental data distribution is "normal" (i.e. bell-shaped) and its mean is denoted as "μ", then the process output is typically given by $\mu \pm 3\sigma$. This values range known as the "process spread" will (theoretically) cover the process output of 99.73%, when the process is stable.

Nevertheless, statistical process control (SPC) tools used in production engineering still often rely on one-dimensional methods. Despite collecting big data using instruments and control systems this tools don't show complex processes complete picture – it uses traditional statistical approaches (mean, standard deviation, Student's criterion, etc.), which only consider individual variables individually.

Although one-dimensional statistics can be useful for the simple systems study, it tends to fail when analyzing more complex systems. This is because it can't detect the relationship that may exist between variables, since they look at variables that are independent of each other. This relationship is known as covariance or correlation and is a central theme in the MVA. Covariance describes the effect that one variable has on others. Process disturbances are usually caused by multiple variables acting together.

The main multidimensional methods are Exploratory Data Analysis (EDA), regression/forecasting methods, and classification methods. EDA attempts to find a hidden structure or underlying patterns in large, complex datasets. This provides a better process understanding and can lead to insights that would not otherwise be observed. EDA methods include cluster analysis and principal component analysis (PCA). An EDA application example is checking for pollutants in a feedstock process or identifying by-products caused by incorrect process settings [7, 8].

Traditional one-dimensional control charts show many different variables at the same time, making it extremely difficult to get a clear, complete picture. Multidimensional control diagrams unite all this data into one or two graphs, taking into account the complex interactions between variables. If the process starts to drift, you can "delve deeper" into specific samples or emissions to quickly identify the problem root cause using a multidimensional and one-dimensional diagnostics combination [8, 9].

Multivariate data analysis can be used together with the software engineering applications through the entire chain of the product development (process model), scaling up or down (considering individual process operations at their own level and the entire process as a whole), process design and its optimization. One of the tasks at the software engineering stage is to understand the correlation of parameters with each other. It is necessary to select a pair or three input parameters on which the quality of one or more output parameters depends. The numerical values of the output parameters directly determine the manufactured products quality and grade (in this case, the produced sugar grade). The correlated parameters are determined by specialists (technologists) based on personal practical experience (human factor). But at the verification and the modeling stage, it is possible to refute many years' developments. Therefore, it is advisable to take the expert technologists opinion as a starting point and when conducting statistical analysis, rely on the results and conclusions obtained. Here, the emergence of a possible conflict is clearly visible in terms of taking into account the opinions of expert technologists and the conclusions of scientific research.

Currently, various statistical modules are used in all economics and management branches which help to solve pressing problems in the production sector. A lot of programs allow complex calculations using mathematical statistics methods (MapleSoft Maple, MathWorks Matlab, MathSoft Mathcad, Wolfram Researh Mathematica) and specialized statistical processing packages (StatGrapics, SPSS, SAS, BMDP and Statistica) developed by Russian programmers [2, 4]. Companies and manufacturing enterprises have the right to choose the necessary software themselves depending on the research objectives.

In the processing of experimental data, statistical criteria are key to identifying and eliminating abnormal measurement results, otherwise known as outliers. These outliers, which are values that significantly deviate from the general trend, are detected at the stage of data preprocessing. The quality of preprocessing directly affects the reliability of subsequent analysis and the construction of a model of the process under study. Ignoring outliers leads to distortion of the results, especially when using classical, non-robust methods of statistical analysis that are sensitive to the presence of

such deviations [10, 11]. This can lead to incorrect conclusions and erroneous interpretations of the obtained data, which is critical for many scientific and engineering problems. Gross errors in measurements are usually obvious and can be detected visually or by simple checks. However, in cases of uncertainty, when information on the quality of measurements is incomplete or questionable, the use of statistical methods becomes necessary. Among such methods, the Grubbs criterion is widely used, designed to check for anomalies of individual values in a sample [6, 10]. Statistical tests are based on the fact that the expected errors in the source data are normally distributed and allow one (minimum or maximum value) or two outliers (two minimum or two maximum values) to be identified in the sample. It is important to note that the application of Grubbs' criteria, like many other statistical tests, is based on standards. For example, the international ISO 5725 standard defines the procedure for performing such tests and interpreting the results. Deviation from these standards can lead to incorrect conclusions.

In this paper, we investigate the effect of data distribution deviation from normality on the Grubbs' test statistics distribution [7, 12]. This is an extremely important aspect, since the data normality assumption is fundamental to many statistical methods, including Grubbs' tests. This assumption violation can significantly distort the test results and lead to false positive or false negative conclusions. For example, if the data have an asymmetric distribution (e.g., exponential or lognormal), Grubbs' tests may be overly sensitive to outliers at one end of the distribution and less sensitive at the other. This may result in significant outliers at one end of the distribution being missed and minor deviations at the other end being erroneously classified as outliers.

Computer modeling methods can be used to study this influence [2, 10, 11]. Generating samples from various distributions (normal, exponential, log-normal, Cauchy, and others) with the addition of the outlier of different sizes and at different positions in the distribution, it is possible to demonstrate how the Grubbs criterion statistics values change and how this affects the detecting outliers at different probability of significance levels. Such analysis results can be presented as graphs showing the criterion power dependence (detecting a true outlier probability) on the data distribution parameters and the outlier size. It is important to remember that the data processing method choice should be based on data nature understanding, its distribution, and the study objectives. Only then reliability and objectivity of the results can be guaranteed. Ignoring these aspects can lead to serious errors in conclusions and decision-making based on incorrect data analysis.

Let X_1, X_2, ..., X_n, … be the observed sample and the variation series constructed from it. The being tested hypothesis $H0$ is that all X_1, X_2, ..., X_n, ... belong to the same general population. When checking for the largest sample value outlier, the competing hypothesis $H1$ is that, for example, the values up to X_n belong to one law, and the X_n values and the subsequent ones belong to another law, significantly shifted, for example, to the right. When checking for the X_n outlier, the Grubbs criterion statistics is:

$$\frac{\left|X_j - \overline{X}\right|}{S} > 3, \qquad (1)$$

where X_j is the observed sample and the variation series constructed from it; $j = \overline{1 \ldots n}$; S is the quadratic error.

When checking for the smallest sample value outlier, the competing H_1 hypothesis assumes that some of the sample elements under study belong to some other law that is significantly shifted to the left [1, 2, 12]. In this case, the calculated statistics take the form of (1). The statistics distribution significantly depends on the sample size n.

3. Results and discussions

In this work, more than 40 sugar production quality parameters which affect the varietal yield [1, 10, 13] were considered. The process quality study was conducted on 147 values for two quality parameters: molasses quality (Db molasses), % and molasses quantity per day, t. Statistical characteristics of some of the studied variation series are presented in Figs. 1 and 2. The Excel system was used as a statistical analysis system.

Preparing the experimental data process is constructing a regression model initial stage [14]. The module for constructing a regression model is included in many mathematical statistics and specialized statistical processing packages. Fig. 2 shows the graph of the per day amount of molasses and the parameter of the regression model for 118 days. This packages advantage is that the user without delving into the process specifics and relying on the specialists' (technologists') opinion can get an idea of the behavior of the most important parameters of the process quality [15]. The mathematical package allows constructing change in the analyzed sample graph with the subsequent determination of the structure type of the proposed model and the numerical values of the coefficients for this model. The model structure is selected and the coefficients are calculated based on averaged experimental data and statistical results of the yielded good (the correlation coefficient R varies in the range from 87% to 99%). All developed templates have their own set of limited structures, and such templates cannot be modernized [16]. Therefore, such packages use is convenient for obtaining an initial idea of the process model structure.

Statistical packages and modules where arbitrary model structures are available to the user have the greatest interest. The software modules for determining the numerical values of the model coefficients are based on the least squares method.

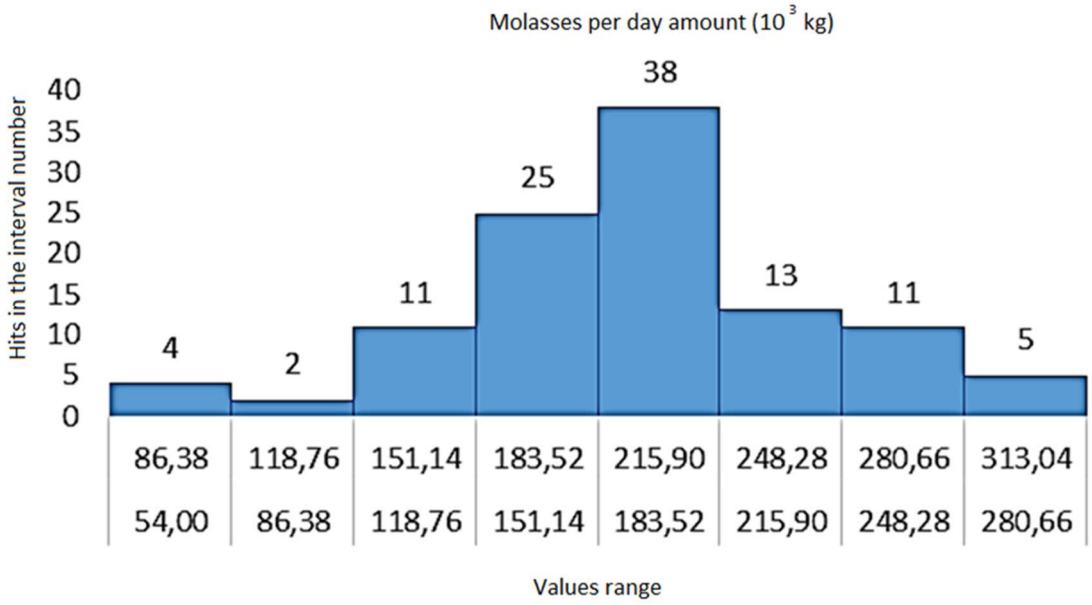

Fig. 1. The example of interval analysis of the process parameter "Amount of molasses per day"

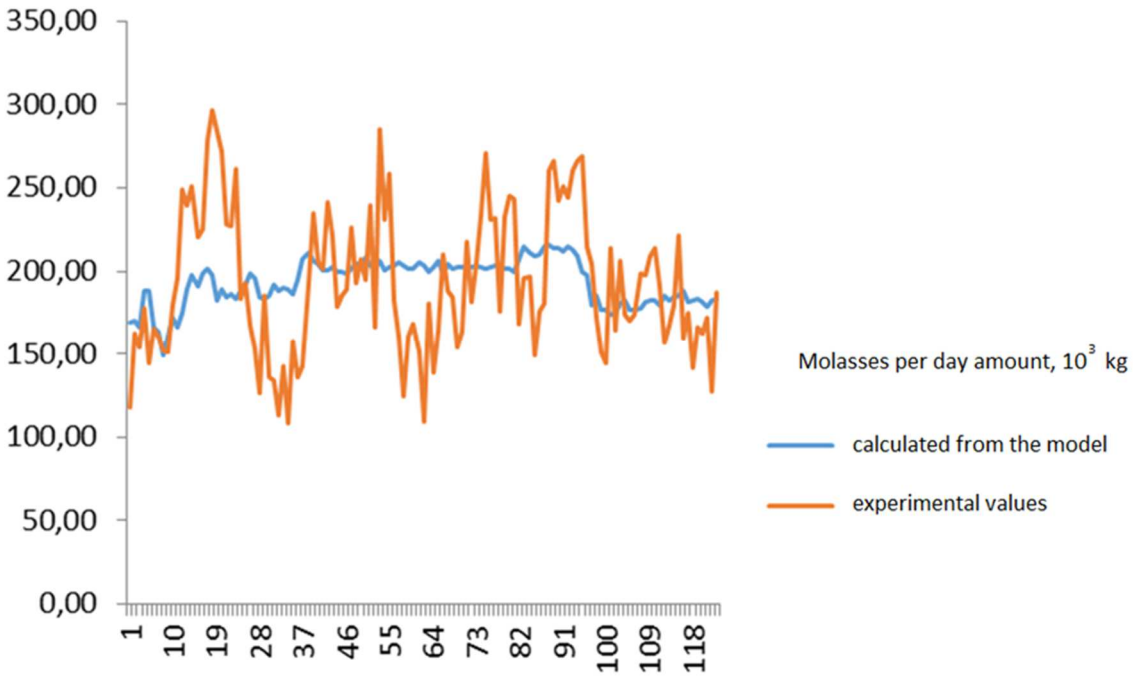

Fig. 2. The regression model of the parameter "Amount of molasses per day"

4. Conclusion

More than 40 sugar quality parameters influencing the varietal yield have been studied [10, 11, 16]. The analysis has shown that the obtained characteristic samples generally obey the normal distribution law. But this is not typical of all parameters. An important feature is the identification of the consistency of the obtained experimental data and the determination of correlations between the quality parameters. Regression equations have been constructed for the per day indicators of molasses quality and molasses quantity. Linear and power models have been constructed. Mathematical statistics and specialized statistical processing packages were used to process the data. The correlation coefficient for the developed models ranges from 0.85 to 0.97.

Figure 1 clearly shows that the average value of per day molasses quantity is in the center of the measurement range. This indicates that the process under study is stable. Analysis of the graph in Figure 2 allows us to conclude that despite the picket-shaped appearance of the experimental data graph the third-order power model yields a result with a correlation coefficient of 0.95. Mutual peaks and valleys of the experimental data are suppressed and as a result we obtain an adequate model.

References

1. S. Chernyaeva, L. Korobova, M. Ivliev. I. Tolstova, B. Nikitin, I. Matytsina Implementation of the Extrapolation Method of Expert Assessments in Selection Problems, High-Performance Computing Systems and Technologies in Scientific Research, Automation of Control and Production, **1304** (2020)
2. Yu.V. Bugaev, L.A. Korobova, I.Yu. Shurupova On the statistical stability of the optimal solution found by the regression equation, VSUIT Bulletin, **86** (2024)
3. N. Gusyatinskaya, T. Nechipor Eureka: Life Sciences 5 (2018)
4. S. Chernyaeva, L. Korobova, M. Ivliev, I. S. Tolstova, B. E. Nikitin, I. A. Matytsina, High-Performance Computing Systems and Technologies in Scientific Research, Automation of Control and Production, **1304** (2020)
5. N.V. Zueva, G.V. Agafonov, T.I. Romanyuk, A.E. Chusova, A.N. Yakovlev, S.A. Veretennikov, IOP Conference Series: Earth and Environmental Science, **640** (2021)
6. V. I. Golik, V. Yu. Konyukhov, Mining Informational and Analytical Bulletin. **11-1,** 175-189 (2023). DOI: 10.25018/0236_1493_2023_111_0_175
7. I.M. Zharkova, Yu.A. Safonova, V.G. Gustinovich, T.L. Ilyeva, Storage and processing of agricultural raw materials, **1** (2020)
8. K. A. Bashmur, V. V. Kondratiev, Mining Informational and Analytical Bulletin. **11-1,** 239-251 (2023). DOI: 10.25018/0236_1493_2023_111_0_239
9. N.G. Kulneva, Yu.I. Bulletin of VSUET, **86** (2024)
10. L. A. Korobova, I. S. Tolstova, I. A. Matytsina, M. S. Mironova, J. Phys, **1902** (2021)
11. A. I. Trunova, S. O. Kurashkin, R. B. Sergienko, Metals, 13(7), 1234 (2023). DOI: 10.3390/met13071234
12. N.G. Kulneva, M.V. Zhuravlev Bulletin of the Voronezh State University of Engineering Technologies, 82 (2020)
13. L.N. Putilina, N.A. Lazutina Storage and processing of agricultural raw materials, **1** (2021)
14. N. Dolgopolova, A. Kaluzhskikh, M. Kotelnikova IOP Conference Series: Earth and Environmental Science, **666** (2021)
15. Yousef, A. Zohri, A. Darwish, A. Shamseldin et al. Microbiol, **14** (2023)
16. O.S. Korneeva, S.F. Yakovleva, N.A. Matvienko, L. N. Frolova et al. Bulletin of the Voronezh State University of Engineering Technologies, **85** (2023)