Vol. XXX, no. 1 (2023), pp. 165 - 177
ISSN 2587-3474
eISSN 2587-3482

https://doi.org/10.52326/jes.utm.2023.30(1).14
UDC 633.34:631.52

MODEL OF STATISTICAL DATA ANALYSIS ON NITROGEN CONTENT IN SOYBEANS (GLICINE MAX MERRILL) IN CLAVERA VARIETY

Ion Ganea*, ORCID: 0000-0002-9346-2575

Moldova State University, 60, Mateevici Street, Chisinau, Republic of Moldova
*Corresponding author: Ion Ganea, *ganea.ion@usm.md*

Received: 12. 23. 2022
Accepted: 02. 17. 2023

Abstract. Climate change, drought and high temperatures lately have led to the need to increase the adaptation capacities of plants to these changes. The purpose of the research is to gain knowledge regarding the influence of the biologically active substances *Reglalg* (compound of algal nature) and *Biovit* (compound of humic nature) on plant development, productivity and adaptation of plants to new climatic conditions. The research was based on decision support systems, machine learning and graph databases. These systems allow in-depth processing of unstructured data and making the necessary decisions. In this sense, an intelligent model was developed for the processing of biological data as part of a Decision Support System. For data analysis and knowledge generation, a graph database was developed to determine relationships and connections between entities, phenomena or events. Graphs allow the representation of complex information in a simple and intuitive way, which makes it easier to perform and analyze them. The use of Machine Learning methods allows the highlighting of laws and interactions between different elements, which can lead to the discovery of patterns and trends that can be easily identified. The paper presents the structure and functions of some components of the decision support system for the study of nitrogen content in soybean.

Keywords: *Decision Support Systems, biostatistics, Wolfram Mathematica, soy, Biovit, Reglalg.*

Rezumat. Schimbările climatice, seceta și temperaturile ridicate din ultima vreme au dus la necesitatea creșterii capacității de adaptare a plantelor pentru a face față acestor schimbări. Temperaturile ridicate au un impact semnificativ asupra proceselor fiziologice care au loc în plante. Problemele studierii influenței biostimulatorilor asupra compoziției chimice a boabelor unor culturi sunt slab structurate. Cercetarea s-a bazat pe sisteme de sprijinire a deciziilor, învățare automată și baze de date graf. Aceste sisteme permit prelucrarea în profunzime a datelor nestructurate și luarea deciziilor necesare. Printre problemele analizate se numără obținerea unei producții de înaltă calitate, care să satisfacă nevoile tot mai mari ale populației: problema mediului și cea a alimentelor. A fost dezvoltat un model inteligent pentru prelucrarea datelor biologice. Permite o înțelegere cuprinzătoare a mecanismelor de acțiune ale biostimulatorilor asupra calității boabelor de soia (*Glycine Max* Merrill). Scopul cercetării este de a dobândi cunoștințe privind influența substanțelor biologic active asupra dezvoltării plantelor, productivității și adaptării plantelor la anumite condiții climatice. Lucrarea prezintă structura și funcțiile unor componente ale sistemului de suport decizional pentru studiul conținutului de azot din boabele de soia. Plantele au fost crescute în condiții de câmp prin administrarea de compuși obținuți din surse naturale.

Cuvinte cheie: SSD, biostatistică, *Wolfram Mathematica*, soia, *Biovi*, *Reglalg*.

1. Introduction

Agriculture is a field highly dependent on natural conditions, climate changes caused by increasing temperatures and reducing precipitation, factors with a significant role in obtaining high quality crops. One of the global problems at the present time is to provide the population with high quality food products and to a sufficient extent to ensure both the health of the population and the protection of the environment. The continuous growth of the population leads to an increase in consumption and, implicitly, in the quality of food products.

The main objective in agriculture is to obtain a high yield of agricultural crops, to adapt them to climate change and to develop their own responses as a result of adaptive reactions. This yield is due to several factors: the amount of precipitation, the temperature, the quality of the soil. Each of these factors contributes significantly to obtaining a high yield.

In addition to the mentioned factors, biologically active compounds play an essential role in the growth and development of plants, as well as obtaining high quality products. To determine the quantity and quality of agricultural production, the efficiency of biologically active compounds, we must take into account all the factors that interact with each other, each having a certain impact and influencing to a certain extent the quantity and quality of the products. In order to achieve the set objectives and make the necessary decisions, a decision support system was developed for the purpose of analyzing and solving problems through a trans-disciplinary approach on problem families.

The application of efficient and complex technologies to reduce the use of chemical fertilizers and their replacement with ecological compounds (biostimulants), optimizing production costs by optimizing the use of resources and obtaining consistent and stable incomes, as well as improving their quality and efficiency, obtaining a high production and quality soybean. Thus, it is necessary to implement advanced technologies for data processing and knowledge generation in order to solve decision-making problems.

A general definition of the problem is: a matter that presents unclear, debatable aspects, that requires clarification, that lends itself to discussions, an important matter that constitutes a task, a (major) concern and that requires an (immediate) solution, a difficulty that must be solved to achieve a certain result; weight, impasse, thing hard to understand, hard to solve or explain; mystery, enigma [1]. „A question proposed for solution; a matter stated for examination or proof; hence, a matter difficult of solution or settlement; a doubtful case; a question involving doubt" [1]. Research Problem is a situation or circumstance that requires a solution to be described, explained, or predicted [2].

The decision-making problems faced by managers from various fields of activity and research are: ***structured***, ***semi-structured*** or ***unstructured***.

A classification given by H. Simon and A. Newell, adapted by C. Gaindric reveals aspects of systems modeling and decision making [3].

1) ***Well-structured problems*** are formulated quantitatively, in which the essential dependencies are highlighted so well that they can be expressed by numbers and symbols, which ultimately receive numerical evaluations.

2) ***Unstructured problems*** or problems expressed qualitatively, are those problems that contain only the description of the most important resources, characteristics and properties, the quantitative dependencies between which are absolutely unknown.

3) ***Poorly structured problems*** are those problems, which contain both quantitative and qualitative elements. Qualitative moments tend to predominate [3, 4].

To solve these problems, Decision Support Systems (DSS) are developed and implemented. These systems can be used for all types of decision problems, but lend themselves best to loosely structured and unstructured ones.

The field of science, which deals with decision-making, is an interdisciplinary field, in training, in which methods of management science, optimization, information technologies, psychology, etc. are accumulated.

The decision-making process has the following components:

1) ***The object*** (system, process examined).

2) ***The subject*** (the decision maker) - the person or the intelligent system that makes the decision and determines that the system needs to be transformed to obtain new qualities and characteristics. The decision-maker formulates the objectives and establishes the criteria by which the degree of achievement of the objectives is measured, determines the internal parameters and restrictions, imposed by the technology and the operation of the system [5].

The problems researched in the doctoral project are loosely structured problems and can effectively be modeled within a *decision support system.*

An DSS is an interactive software system used to help decision makers extract useful information from a set of raw data, documents and knowledge, identify and solve problems in order to make optimized decisions.

The Academician Florin Gheorghe Filip defines the concept of DSS as "an anthropocentric, adaptive and evolving computer system intended to implement some of the functions of a possible "human support system" (or team) that would otherwise be needed to help the decision-maker overcome his limits and constraints they might face when approaching a problem of decision that matters” [5].

DSS integrates a knowledge base. The model materializes and harmonizes the decision concept with the user criteria and the user interface. The main advantages of using an DSS include examining more alternatives, better understanding of processes, identification of contingencies, improved communication, cost efficiency, and better use of data and resources [4].

The concept of decision is defined as "Decision taken after examining a problem, a situation, etc. adopted solution (among several possible)" [4, 6].

F. Gh. Filip defines the concept of "decision" as follows: "The decision represents the result of conscious activities of choosing a course of action and engaging in it, which usually involves the allocation of resources. The decision results from the processing of information and knowledge and belongs to a person or a group of persons, who have the necessary authority and who are responsible for the effective use of resources in certain given situations" [5, 7]. Depending on the degree of structuring of problems and decisions, they can be classified into [7]:

1. **Structured decisions** are made following a known and explicit process that allows the processing of input information to choose alternatives for the adoption of which there are predetermined procedures.
2. **Unstructured (non-programmable) decisions** are those decisions whose elements are more qualitative, the aims and objectives are not precise and there is no known solution algorithm. They are important decisions, innovative and often atypical. There are no set procedures for their adoption.
3. **Semi-structural decisions** are those decisions in the adoption of which only partially known procedures can be used. The decision has predominantly quantitative elements and the objectives and goals are not precise and the solution algorithm does not cover all the elements of the problem.

In the field of biology and agriculture we often face unstructured problems. The application of DSS in agriculture and the environment has grown rapidly in recent times, which allows the rapid assessment of agricultural production systems, knowledge acquisition and decision-making both at the farm level and at the regional or national level.

One of the important applications of DSS in agriculture is the management of the use of compounds to stimulate the growth and development of plants, as well as to increase the quality of the fruit, the aim being to reduce or eliminate the use of chemical products in agriculture and to replace them with biological compounds.

Knowledge decisions are related to evaluating ideas for new information products and services, methods of communicating new knowledge, and disseminating information in the organization.

In order to solve the problems regarding the quality assurance of food products, an DSS was developed. Its purpose is to assist research to determine:

(a) the influence of certain biostimulants on plants and

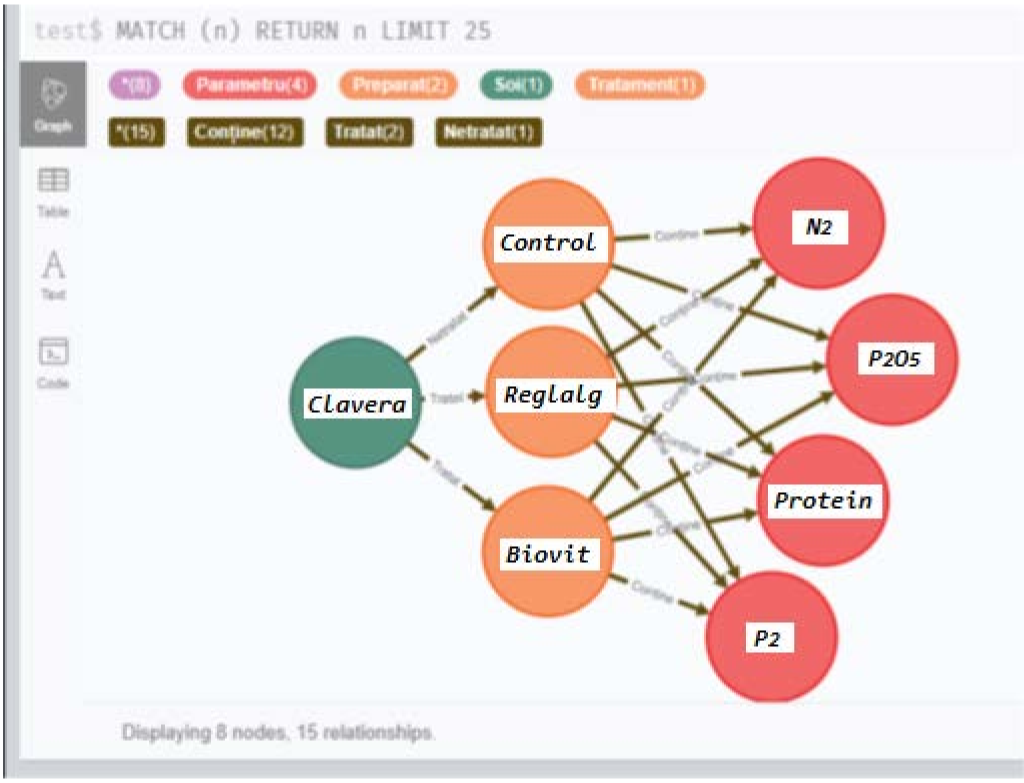
(b) determining the degrees of influence of these biostimulators.

The goals of DSS are:

a) Research assistance to determine the influence of certain biostimulants on plants and their degrees of influence.

b) Complex data analysis and extraction of information/knowledge/opinions with applicability in the field of biology and agriculture, starting from the model made on soybean plants (*Glicine Max* Merrill) and processing them for the discovery of patterns, correlations, classifications, groupings, or new predictive models as well as data mining using data processing techniques based on biostatistics.

c) *Graph database* development to improve data access by providing customized search services, normalization and distributed access to content and multiple data sources. Information technologies equipped with graph databases make it possible to store relationships and connections as first-class entities. The architecture of graph databases is based on graph theory. A graph database can be defined as a database that uses semantic query graph structures with nodes, relationships, and properties to represent and store data [8,9]. This type of database presents data as it is conceptually visualized. This ability is achieved by transferring data in nodes and the relationships between them in edges, being an optimized solution for modelling and querying large volumes of closely related data, representable by graph-type structures, which allows a high degree of adaptability to real models. Graph databases demonstrate the advantages of efficiently storing relationships between data points and are flexible in adding new types of relationships or adapting new data models for new research requirements. Stakeholders can use graph databases to generate competitive insights and meaningful value from connected data. These databases are that technology solution that allows data professionals at all levels to exploit the potential of their data relationships rather than just individual data points and the imagination of the database is the only limitation of how these relationships could be harnessed by the user [9]. A subgraph reflecting the content of *Nitrogen*, *Phosphorus* and *Protein* in the samples untreated and treated with compound *Prep1* and *Prep2* at soil *S1* is represented in figure 1.



```
test$ MATCH (n) RETURN n LIMIT 25
```

Displaying 8 nodes, 15 relationships.

Figure 1. Subgraph content of elements.

The created model allows determining the influence of humic (Biovit) and algal (Reglalg) biostimulant compounds on the quality of soybeans. Research can help researchers and agricultural producers strengthen their efforts in order to mitigate the consequences of climate change and increase the adaptability of crops to new development conditions [10]. The research was carried out on several soybean genotypes. In this work, the results obtained on the Clavera variety are reproduced.

The Clavera variety is an indigenous variety obtained by hybridization and selection of the combination Timpurie × Nordic 138. Grain production within the limits of 22.3 - 32.5 q/ha. The protein content in the grains is 36.9% and 20.5% oil. The mass of 1000 grains (MMB) is 161 g. The vegetation period is 122 days. The plant has a height of 80-90 cm. The medium-sized grain has an oval-elongate shape, a thin yellow skin, with an elongated, light yellow hilum. The variety is resistant to falling, shaking and drought, with tolerance to diseases and pests [11].

2. Materials and Methods

The biostatistics method was used in our research [12]. The objectives of the software components, carried out at the moment, were to assist research to determine the degrees of influence of the biostimulators *Reglalg* (compound of algal nature) and *Biovit* (compound of humic nature) on the quality of soybeans. The results of the statistical processing of the data with this model are presented.

The main problem of the research is the verification of the statistical hypothesis regarding the difference of means. Eq. (1), [13]:

$$h_0 : \bar{d} = 0 \text{ or } h_a : \ \bar{d} \neq 0, \tag{1}$$

where: h_0 - null hypothesis.

h_a - alternatival hypothesis.

$\bar{d}$ - the difference of means.

The meaning of the first two concepts is as follows:

- ***Null hypothesis h₀*** - determines the probability distribution of one or more variables by which it finds that there is no difference or relationship between two measured phenomena or association between two groups. Thus, a fact is assumed to be true until proven otherwise [14].
- ***Alternative hypothesis - h_a* -** the data are not related to each other, the compared values differ from each other.

One of the parameters for statistical analysis are the *degrees of freedom*. The degrees of freedom (*df*) in statistics indicate the number of independent values that can vary in an analysis without breaking any constraints. It is an essential idea that appears in many contexts throughout statistics including hypothesis tests, probability distributions, and linear regression [14-17].

To determine the degree of influence on the researched parameters based on a small sample, the *t test* (*Student)* was used.

We note, that the data sets are independent, with different dispersions Eq. (2):

$$(\sigma_1^2 \neq \sigma_2^2), \tag{2}$$

where: σ_1^2 – the dispersion of the first sample.

σ_2^2 – the dispersion of the second sample.

and equal number of repetitions Eq. (3):

$$(n_1 = n_2 = 3), \tag{3}$$

where: n_1 – the number of items in the first sample.

n_2 - the number of items in the second sample.

Thus the difference of the means will be calculated [13].

2.1. Data processing

The component's model former realized using the biostatistics method [12].

Testing stages/ In the current research we have different dispersions ($\sigma^2 \neq \sigma^2$) and the number of repetitions is equal *($n_1 = n_2$)* , the mean of the differences will be calculated [17].

a) Functions have been developed to calculate ***the mean's difference* .** Eq. (4):

$$\bar{d} = | \bar{x}_1 - \bar{x}_2 |, \text{ where} \quad (4)$$

$\bar{d}$ - the mean's difference;
$\bar{x}_1$ – the mean of the first sample;
$\bar{x}_2$ – the mean of the second sample.

Example of a function made in Wolfram Mathematica:

medCN2MR = Abs[Mean [cN2M]] – Abs[Mean[cN2R]], where:

- *medCN2MR* - The variable that determines the difference in the average *nitrogen* content between the sample treated with *Reglalg* and *Control* in the *Clavera* variety;
- *Abs[Mean [cN2M]]* - The absolute value of the average nitrogen content in the control sample of the Clavera variety;
- *Abs[Mean[cN2R]]*- The absolute value of the average nitrogen content in the sample treated with *Reglalg* of the *Clavera* variety.

b) Degrees of freedom (*df*) for the independent test are determined according to the following formula Eq. (5):

$$df = n_1 + n_2 - 2 \quad (5)$$

c) The error of the average of the differences is calculated according to the formula Eq. (6):

$$s_{\bar{d}} = \sqrt{s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2}, \quad (6)$$

where: $s_{\bar{d}}$ - The error of the mean of the differences.
$s_{\bar{x}_1}^2$ - Standard error of the first sample.
$s_{\bar{x}_2}^2$ - Standard error of the second sample.

Having 3 repetitions in each sample, we have *df = 3 + 3 – 2 = 4* (degrees of freedom).

d) *t -test (Student)* is determined as the ratio between the mean of the differences and the error of the mean of the differences according to the formula Eq. (15,16):

$$t = \frac{| x_1 - x_2 |}{s_{\bar{d}} = \sqrt{s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2}}. \quad (7)$$

The elements of this formula being exposed above.

If the experimental value is greater than the theoretical value ($t_{exp} \geqslant t_{teor}$), the null hypothesis of no significant differences is accepted, the differences are within the confidence interval.

Example: The theoretical *t*-value (t_{teor}) for 4 degrees of freedom at a significance threshold of $t_{0.05}$ is 2.13 and at a significance threshold of $t_{0.01}$ is 3.75 for the one-sided *t*-test. The *t*-value obtained is 34.2, therefore the null hypothesis is rejected and a significant positive influence of the treatment on grain quality is found.

The result can also be confirmed by using analysis ***lsd*** - *Least Significant Difference* Eq. (8), Eq. (9) [15, 18]:

$$lsd_{0.05} = t_{0.05} * \bar{d} \quad (8)$$

$$lsd_{0.01} = t_{0.01} * \bar{d} \quad (9)$$

When the difference of means ($\bar{d})$ is greater than the value *lsd* calculated, the null hypothesis is rejected.

Examples:

$$lsd_{0.05} = t_{0.05} * \bar{d} = 2.13 * 0.657 = 0.0409;$$
$$lsd_{0.01} = t_{0.01} * \bar{d} = 3.75 * 0.657 = 0.072.$$

difference of means represents $\bar{d} = 0.657$, which is higher than the *lsd value* obtained, therefore *confirms* the previous result.

An alternative to the point of rejecting the null hypothesis (H_0) is the *p - value*, which provides the lowest significance threshold below which the alternative hypothesis (H_a) is proven that there is sufficient evidence to reject the null hypothesis (H_0). It is a measure of statistical significance. The result *p* of the test, provided as a number between 0 and 1, and represents *the probability of error* if we reject the hypothesis H_0 . If *p* is lower than the significance threshold α chosen (usually α = 0.05), we reject the hypothesis $H\ _0$ and accept as true the hypothesis H_a [19]. The interpretation of *p* values is done in most statistical tests as follows:

- ✓ $p < 0.05$ *(5%)*, the statistical link is significant (*S*, 95% confidence).
- ✓ $p < 0.01$ *(1%)*, the statistical link is significant (*S*, 99% confidence).
- ✓ $p < 0.001$ *(0.1%)*, the statistical link is highly significant (*HS*, 99.9% confidence).
- ✓ $p \geqslant 0.05$, the statistical relationship is not significant (*NS*).

2.2. Statistical analysis

The statistical analysis and the results obtained are presented below. The exposed tasks were carried out by means of the Wolfram Mathematica software system [20]. This is a high-level language that allows automation of the software development process.

To create and use the calculation functions, the Wolfram Mathematica libraries are imported:

- **Needs["Hypothesis Testing`"]** is a hypothesis testing package. The package contains functions for computing confidence intervals from data, *p-values* and confidence intervals for distributions related to the normal distribution.
- **Needs["ComputerArithmetic`"]** represents an arithmetic calculation package.

The experimental data are presented in Table 1.

Table 1

Nitrogen content (experimental data)

	Control	***Reglalg***	***Biovit***
N_2 (%)	5.35	6.0	6.17
	5.39	6.05	6.19
	5.37	6.03	6.26
$\bar{x}$	**5.37**	**6. 02**	**6.20**

Note. ***N_2*** **–** nitrogen; $\bar{x}$ - values' mean.

3. Results and Discussion

3.1. Nitrogen content (N_2)

Functions were developed to create the diagram that determines the mean nitrogen content with the standard error based on the experimentally obtained data. The resulting diagram is shown in figure 2.

The *MeanAround* function provides an Around object that describes the mean of x_i and its uncertainty. By means of this function, the mean of the analyzed values and the standard error are calculated which will be indicated on the bar.

The chart is made with the *BarChart* function with formatting parameters such as style, labels, legend, appearance, etc.

ChartElementFunction is an option for chart functions. In the present case *BarChart* provides a function to use to generate the rendering primitives of each chart element.

"GlassRectangle", "Pastel" is the display style of the chart and bars.

ChartStyle - option for chart functions to specify styles in which chart elements should be drawn.

ImageSize - the size of the chart.

ChartLabels - the labels of each bar.

ChartLegends - chart legend. Specifies the legends used for chart elements.

Comments are shown between symbols: (* *Comment* *).

```
ClaveraN2 (*Declare the variable used in the calculation*).
={ MeanAround [{5.35,5.39,5.37}],
MeanAround [{6,6.05,6.03}],
MeanAround [{6.17, 6.19, 6.26}]};
BarChart [ClaveraN2,
ChartElementFunction → "GlassRectangle",
ChartStyle → "Pastel",
ImageSize → 400,
ChartLabels →{"1","2","3"},
ChartLegends → {"Control", " Reglalg ", " Biovit "}].
```

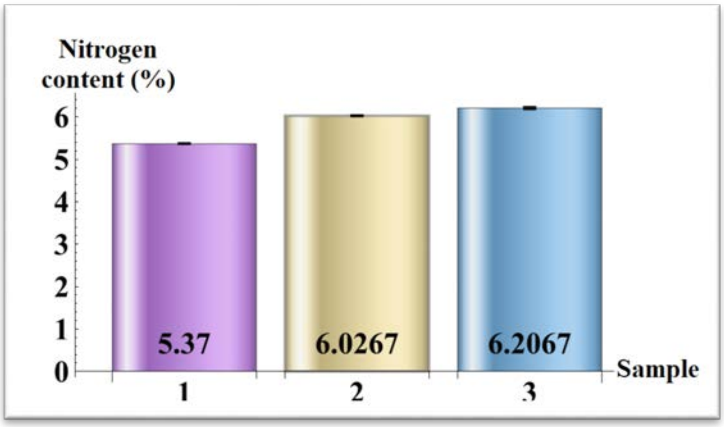


Figure 2. Variation of Nitrogen Content: 1– control, 2 – reglalg, 3 – biovit.

3.2. Variation of nitrogen content (N_2)

Functions have been developed to analyze the variation in nitrogen content. They materialize in a table, using the following elements:

- *Grid* function - is an object that is formatted with $_{an}$ *expression* arranged in a two-dimensional grid.

The statistical functions for each parameter (*Control* , biostimulators *Reglalg* and *Biovit*):

1) *Sample mean* with standard error - *MeanAround* (described previously).

2) *Dispersion - Variance* (- σ_2).

3) *The standard deviation - StandardDeviation* (σ).

4) *The coefficient of variation* expressed as a percentage, which is calculated as the ratio between *the standard deviation* and *the mean of the sample values* * 100, according to the formula Eq.(10):

$$\frac{\sigma}{\bar{x}} * 100, \tag{10}$$

where: σ - *standard deviation.*

$\bar{x}$ - *the mean of the sample values.*

5) ***Confidence Interval*** - *MeanCI*. A confidence interval provides boundaries within which the value of a parameter is expected to lie with a certain probability. Interval estimation of a parameter is often useful in observing the accuracy of an estimator as well as making statistical inferences about the parameter in question [16]. The *MeanCI* function provides confidence intervals of means and differences of means based on the central limit theorem.

The functions receive as parameters the variables declared for each studied parameter (*Control, Reglalg , Biovit*). The results are presented in table 2.

```
    Grid[{{"Parameters", " x̄(%)", " σ 2 ", " σ ", "CV(%) * ", "C I * * "},
    {"Witness", MeanAround [cN2M],
    Variance [cN2M], StandardDeviation [cN2M], StandardDeviation [cN2M]/ Mean
[cN2M]*100, MeanCI [cN2M]},{ " Regallg ",
    MeanAround [cN2R],
    Variance [cN2R],
   Standard Deviation [cN2R],
   StandardDeviation [cN2R]/ Mean [cN2R]*100,
   MeanCI [cN2R]},
   { " Biovit ",
   MeanAround [cN2B],
   Variance [cN2B],
   Standard Deviation [cN2B],
   StandardDeviation [cN2B]/ Mean [cN2B]*100,
   MeanCI [cN2B]}} //N]]
```

Table 2

Variation of nitrogen content (N_2)

Parameters	$\bar{x}$(%)	σ^2	σ	Coefficient of variation (%)	Interval trustworthy
Control	5.37 ± 0.012	0.0004	0.02	0.372	{5.32, 5.42}
Reglalg	6.027 ± 0.015	0.000633	0.025	0.418	{5.96, 6.09}
Biovit	6.207 ± 0.027	0.00223	0.047	0.761	{6.09, 6.32}

Notes. $\bar{\mathbf{x}}$ - mean's differences; $\boldsymbol{\sigma}^2$ – sample's dispersion; $\boldsymbol{\sigma}$ – sample's standard deviation.

3.3. Results of treatment influence on nitrogen content (N_2)

Functions were developed to analyze the influence of treatment on nitrogen content. The influence of a compound is compared with the control sample (control) and we find:

The difference in the values of the samples expressed in percentage (Δ) according to the formula Eq. (11):

$$dif = 100\ \frac{Mean[param_1]}{Mean[param_2]}, \tag{11}$$

where: *dif* is the difference between the mean values expressed as a percentage

Mean [ ] is the function for determining the average of the first parameter [*Control*] and the second parameter the group treated with the biostimulator [*Reglalg*] or [*Biovit*]:

1) Difference of sample means ($\bar{d}$).
2) Standard error of differences ($s_{\bar{d}}$).
3) The result of the *t*-test (student).
4) Least significant difference (*lsd*).

The formulas used [2-5] were previously presented in the paper. The results are presented in table 3.

```
Grid [{{" Comparison ", " Δ (%)", "d̄ ", "s d̄ ", "t", " lsd 0.05 ", " lsd 0.0 1 "},
{"Witness vs Reglalg ", dif =10 0 -( Mean [cN2M]*100/ Mean [cN2R], medCN2MR = Abs
[ Mean [cN2M] – Mean[cN2R]], ercN2MR=√ 0.12^2 +0.15^2 , t=(medCN2MR/ erCN2MR )
* , lsd 0.05  =t 0.05  * erCN2MR * ,
lsd 0.01  = t 0.01  * erCN2MR * },
{"Witness vs Biovit ", dif =10 0 -( Mean [cN2M]*100/ Mean [cN2B], medCN2MB = Abs [
Mean [cN2M] – Mean [cN2B]], ercN2MB=√ 0.12^2 +0.27^2 , t=(medCN2MB/ erCN2MB )
* , lsd 0.05  =t 0.05  * erCN2MB * ,
lsd 0.01  = t 0.01  * erCN2MB * }]
```

Table 3

Results of treatment influence on nitrogen content (*N*$_2$)

Comparison	Δ (%)	$\bar{d}$	$S_{\bar{d}}$	*t*	$lsd_{0.05}$	$lsd_{0.01}$
Control vs Reglalg	10.9	0.657	0.0192	34.2^{*}	0.0409^{*}	0.072^{**}
Control vs Biovit	13.5	0.837	0.0295	28.3^{*}	0.0629^{*}	0.111^{**}

Notes. * Significant influence at the 95% level. ** Significant influence at 99% level. **Δ** - the absolute difference in percentages; $\bar{d}$ - the mean's differences of samples; $s_{\bar{d}}$ - the standard error of differences; ***t*** – rezult of test ***t*** (Student); ***lsd*** – least significant diffenences.

The influence of biostimulators on protein content is determined. Table 3 shows that *the* t_{exp} value obtained from the data analysis is 34.2 when using *Reglalg* and 28.3 when using *Biovit*, which is higher than the theoretical value which is 2.13 at the 95% confidence level and 3.75 at the 99% confidence level. Thus, **the null hypothesis is rejected**.

The result can also be confirmed by using least significant difference analysis. The differences of means $\bar{d}_{reglalg}$ = 0.657, $\bar{d}_{biovit}$ = 0.837 are greater than the *lsd values* obtained according to table 2.3 ($lsd_{0.05}$ =0.04 and $lsd_{0.01}$ = 0.072 when using *Reglalg* and $lsd_{0.05}$ =0.0629 and $lsd_{0.01}$ = 0.111 when using *Biovit*), therefore the previous result **is confirmed** , the null hypothesis is rejected.

An additional confirmation for rejecting the null hypothesis is the ***p***-value [16] is used:

Wolfram Mathematica function is used for this purpose *StudentTPValue* which has as parameters the value of the test ***t*** and the degrees of freedom *(df):*

StudentTPValue [34.2, 4] → $\boldsymbol{2.1843*10^{-6}}$

StudentTPValue [28.3, 4]→ $\boldsymbol{4.6273*10^{-6}}$

The result of this test is less than the significance threshold. Thus, the result obtained by the ***t*** and ***lsd*** tests is confirmed.

4. Conclusions

The evaluation of the quantity and quality of agricultural production, the efficiency of the compounds involved is done by analyzing their interaction and the impact on plant development. The model helps solve the problems of eliminating synthetic chemicals and using ecological products, which will subsequently contribute to obtaining high-quality products and increasing production volumes.

Wolfram Mathematica systems and graph databases. Representing linked data is simple. Data analysis is done quickly across a large number of entities. No complex connections are required to retrieve connected data for analysis. Solutions to complex problems can be obtained quickly. Any type of problem can be solved: structured, semi-structured and unstructured.

The statistical analysis of the data shows that the experimental preparations substantially influence the chemical composition of soybeans. The biostimulants used influence the nitrogen content, the values of the researched parameters increase substantially in the Clavera variety, the results demonstrating significant influence both at the 95% and 99% level of veracity. Thus, these biostimulants are recommended to be widely used in agriculture to increase plant resilience. This system helps researchers, manufacturers of plant preparations and agricultural producers to solve their research and activity problems aimed at obtaining high-yielding soybean crops and plant adaptation to climate change.

The work was carried out as part of the Doctoral Project "Models, Techniques and Program Products for Intelligent Data Analysis in Plant Physiology".

Acknowledgments: I would like to thank the Associate professor, PhD. Ana Bârsan, Head of the Department of Biology and Ecology, Faculty of Biology and Geoscience, Moldova State University for the data obtained during the experiments and the valuable consultations in the field of plant physiology, and professor, PhD. Eng. Gheorghe Capațână for methodological assistance related to research and development of the intelligent support system for plant genetics.

Conflicts of Interest: The author declares no conflict of interest.

References

1. biologyonline.com. Available online: https://www.biologyonline.com/dictionary/problem. (accessed on 1.12.2022)
2. ndl.ethernet.edu.et. Available online: http://ndl.ethernet.edu.et/bitstream/123456789/87834/2/Chapter%202_Scientific%20Research%20Methods%20-%20Defining%20the%20Research%20Problem.pdf. (accessed on 1.12.2022)
3. Gaindric, C. *Decision making. Methods and technologies*. Science, Chisinau, Moldova, 1998; 21 p.

4. Gaindric, C. *Systemic approaches in decision-making (course support)*. University of the Academy of Sciences of Moldova, Chisinau, Moldova, 2017, 26 p.
5. Filip, F.G. *Decision Support Systems*. Technical Publishers, Bucharest, Romania, 2004, 16 p.
6. Rinaldi, M.; Zhenli, H. Chapter Six - Decision Support Systems to Manage Irrigation in Agriculture. In *Advances in Agronomy*, Academic Press, Foggia, Italy, 2014, 123, pp. 229-279, https://doi.org/10.1016/B978-0-12-420225-2.00006-6.
7. DEX.RO. Definition Decision, Available online: https://www.dex.ro/decizie (accessed on 1.12.2022).
8. TowardsDataScience.com. Morgante, V. What is a graph database? Available online: https://towardsdatascience.com/what-is-a-graph-database-249cd7fdf24d. (accessed on 03. 12. 2022).
9. Robinson, R; Webber J.; Eifrem, E. *Graph Databases*. O'Reilly, Sebastopol, California, USA, 2015, pp. 19-24.
10. Hasan, N; Suryani, E.; Hendrawan, R. Analysis of Soybean Production and Demand to Develop Strategic Policy of Food Self Sufficiency: A System Dynamics Framework. *Procedia Computer Science* 2015, 72, pp. 605-612. https://doi:10.1016/j.procs.2015.12.169.
11. Bîrsan A.; Jigău G.; Armaş A. The influence of the humic compound "biovit" on the growth and development of soybean plants grown on an aqueous nutrient medium. *Studia Universitatis Moldaviae* 2019, 6, 18 p.
12. McDonald, J. H. *Handbook of Biological Statistics*. University of Delaware. Sparky House, Newark, USA, 2014, pp. 45-48.
13. StatisticsByJim.com Available online: https://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/ (accessed on 3.12.2022).
14. Dodge Y. *The Concise Encyclopedia of Statistics*. Springer, Berlin, Germany, 2008; 25 p.
15. Dospekhov B. A. *Methods of field experience (with the basics of statistical processing of research results)*. 5th ed., Agropromizdat, Moscow, Russia, 1985, pp. 193-195. [in Russian].
16. Gataulin, A.; Lica, D.; Pomohaci, C. *Biostatistică intuitive*. Ceres, București, România, 2002, 147 p.
17. Singpurwalla, D. *A handbook of statistic: An overview of statistical methods*. 1st edition; Bookboon, Copenhaga, Denmark, 2013, pp. 64-66.
18. Brian E. *The Cambridge Dictionary of Statistics*. Cambridge University Press, Cambridge, United Kingdom, 1998, pp. 91,245
19. StatisticsHowTo.com. Available online: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/p-value/ (accessed on 3.12.2022).
20. References.Wolfram.com. Avalable online: https://reference.wolfram.com/language/guide /Statistics.html (accessed on 12.12.2022).

Citation: Ganea, I. Model of statistical data analysis on nitrogen content in soybeans (*Glicine Max* Merrill) in Clavera variety. *Journal of Engineering Science* 2023, 30 (1), pp. 165-177. https://doi.org/10.52326/jes.utm.2023.30(1).14.

Publisher's Note: JES stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submission of manuscripts: jes@meridian.utm.md