Advances in Intelligent Systems Research, volume 167

International Scientific and Practical Conference "Digitization of Agriculture - Development Strategy" (ISPC 2019)

Technologies of intellectual analysis of the data in agricultural research

Viktor Buyarov
Department of small animal science and farm animal breeding
Orel State Agrarian University named after. N.V. Parakhin
Orel, Russia
bvc5636@mail.ru

Vadim Shumetov
Department of information technologies and mathematics
Orel State Agrarian University named after. N.V. Parakhin
Orel, Russia
shumetov@list.ru

Aleksandr Buyarov
Department of economics and management in the agroindustrial complex
Orel State Agrarian University named after. N.V. Parakhin
Orel, Russia
buyarov_aleksand@mail.ru

Yulia Mikhaylova
Department of foreign and Russian languages, pedagogics and phycology
Orel State Agrarian University named after. N.V. Parakhin
Orel, Russia
julia_michailova@mail.ru

***Abstract*—An important factor in increasing the efficiency of decisions made on the results of agricultural research is the use of advanced digital economy technologies. However, in practice, the analysis of empirical results is often limited to a pairwise comparison of means on impact options, without taking into account the realities of the plan of laboratory or field experiments. The possibility of obtaining new knowledge using information technologies for the analysis of empirical data that support algorithms for multiple comparison of means and multivariate analysis of variance is discussed. To implement these algorithms, you do not need to access expensive software products of the Data Mining class (DM); it is enough to have relatively inexpensive statistical data analysis packages such as early versions of SPSS (Statistical Package for the Social Sciences) starting from version 8.0. An example is given to illustrate the effectiveness of using information systems of the Knowledge Discovery class, focused on the data knowledge search.**

Keywords—digital economy, agricultural research, data mining, multiple comparison of means, multivariate analysis of variance.

I. Introduction

One of the high-impact directions of the development of Russia in the nearest future should be the formation of digital economy [1, 2], an important component of which is data mining (DM). Two directions of software evolution are closely associated with data mining, known as Knowledge Discovery in Databases (KDD) and Data Mining. The first one of them suggests an analytical approach to the problem of knowledge search "from the data", due to which it is characterized by mathematical correctness and reliability of results, while the second direction is more focused on the practical significance of the conclusions and judgments made [3, 4]. However, it is necessary to admit reluctance of agricultural specialists to the widespread usage of data mining technologies: in the first direction due to the insufficient level of mathematical and statistical training, in the second one because of the high cost of software products that support Data Mining technologies.

II. Literary review

Various authors pointed out disadvantages in the use of new data analysis technologies in the practice of agricultural research [5, 6]. For example, V.M. Kuznetsov noted that "most of the works of Russian researchers in livestock breeding contain the analysis of experimental and "field" data ... limited to calculating mean values and their standard errors at the best case. Only a small number of studies use single-factor analysis of variance and very rarely multifactor generalized linear models" [5, p. 27]. The situation in crop production is not better [7].

A number of agrarian researchers see the resolution of this problem in the transition from software products of the KDD class to Data Mining technologies [3, 8]. However, it is obvious that both areas of the digital economy need to be developed in parallel, since not entirely well-formed mathematical and statistical setting of the problems can lead to incorrect conclusions. Let's see an example from the work [9] on the effect of microbiological preparations on the hemoglobin contents in the blood of chickens.

III. Results and discussion

The object of the study was broilers of the cross "ROSS-308". For each experiment two experimental and one control groups of broiler chickens with 100 birds in each group were formed. The chickens of the experimental groups received microbiological preparations in addition to the basic ration: experimental group 1 received preparation "URGA", experimental group 2 received preparation "Baikal EM-1". Each experiment lasted 39 days; blood for the research was collected from three chickens of each group on the 14th, 28th and 38th days.

The results of the hemoglobin measurements in the blood of chickens are presented in Table I and in the Fig. 1. and Fig.2.

TABLE I. Hemoglobin contents (g/l) in the blood of broiler chickens [9]

Age (days)	Group		
	Control	***Experimental group 1***	***Experimental group 2***
14	79,53±1,16	87,73±2,02	80,72±2,03
28	111,10±8,20	113,46±2,33	118,13±1,16
38	102,90±3,11	108,80±2,02	109,96±2,33

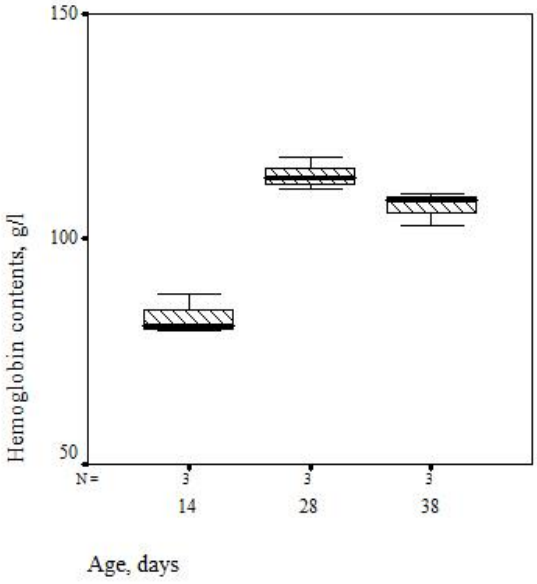

Fig. 1. Hemoglobin contents in the blood of chickens of different age

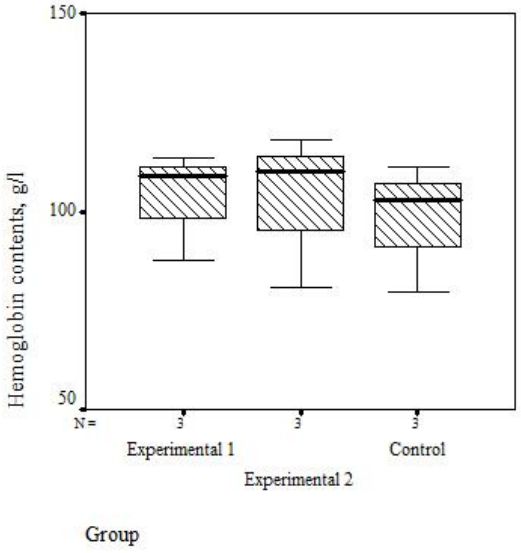

Fig. 2. Hemoglobin contents in the blood of chickens of the experimental and control groups.

Figure 1 shows the mean difference by the factor "age", whereas the distribution of the indicator over the groups is overlapped and more sophisticated analytical tools are required.

The analysis of the mean difference of the indicator "hemoglobin contents" in the "Generalized linear model" procedure of one of the early versions of the statistical software package for analyzing social science data SPSS Base 8.0 [10] showed the significance of the influence of the chicken age on the hemoglobin contents, but the p-level value for the “group” factor exceeded the norm of 0.05, and this factor cannot be considered statistically significant.

TABLE II. TEST OF BETWEEN-SUBJECTS EFFECTS OF THE HEMOGLOBIN CONTENTS INDEX

Source of variability	**Sum of squares**	**Degree of freedom**	**Mean square**	***F-test***	**Signifi-cance**
Corrected model	1705,199	4	426,300	45,677	0,001
Intercept	92482,892	1	92482,892	9909,412	0,000
GROUP	56,200	2	28,100	3,011	0,159
AGE	1648,999	2	824,499	88,344	0,000
Error	37,331	4	9,333		
Total	94225,422	9			
Corrected total	1742,530	8			

The latter fails in agree with the author’s conclusions [9] about the statistical significance of the means differences in the factor “group”, drawn under the assumption of three replications of each experiment. The reason for this contradiction is fallible interpretation of the notion “replication”: in the work cited three replications does not refer to the true experimental unit, which role in the experiment was represented by a group of chickens of 100 heads, but to the samples from the groups of three chickens, the dimensions of which were taken as “replications”, which in reality are “ pseudo-replications ”.

Such errors are not rare, and in the opinion of Finnish ecologist M.V. Kozlov, they are associated with an unjustified generalization of the particular (for a given method of sampling action) results for the general population [11, p. 294]. The phenomenon of imaginary replication is also discussed in the work [12], which is a compilation of several articles. One should agree with the authors of one of them - V.K. Shitikova, N.A. Tseitlina and V.N. Yakimov that “pseudoreplication should be considered only as a situation when nonstatistical argument is replaced by a statistical one. If the researcher clearly outlines the limits of his statistical analysis, there can be no complaints about it ... ”[12, p. 109].

In the specific context described above, the way out is simple: taking into account that preparation “URGA” differs from preparation “Baikal EM-1” only in some supplements that have little effect on the hematological parameters of the blood of the chickens, experimental groups 1 and 2 should be combined. As a result, the factor “group”: p-value for the Fisher criterion p = 0.041 is less than the critical value 0.05. At the same time, the model with one experimental group is quite informative - it explains 97.8% of the total variance, almost as much as the model with two experimental groups (97.9%).

TABLE III. TEST OF BETWEEN-SUBJECTS EFFECTS OF THE HEMOGLOBIN CONTENTS INDEX (MODEL WITH ONE EXPERIMENTAL GROUP)

Source of variability	**Sum of squares**	**Degree of freedom**	**Mean square**	***F-test***	**Signifi-cance**
Corrected model	1704,967	3	568,322	75,648	0,000
Intercept	80783,241	1	80783,241	10752,919	0,000
GROUP	55,968	1	55,968	7,450	0,041

AGE	1648,999	2	824,499	109,748	0,000
Error	37,563	5	7,513		
Total	94225,422	9			
Correc-ted total	1742,530	8			

The results obtained make it possible to recognize a two-factor model of analysis of variance with one experimental group as adequate.

$$Y_{ij} = \mu_0 + \alpha_i + \beta_j + \varepsilon_{ij} \quad (1)$$

and evaluate its parameters (Table IV), where Y_{ij} is the observed value of the output signal Y at the i-th level of the factor "group" and the j-th level of the factor "age"; μ_0 - estimation of the free coefficient of the model; α_i and β_j – assessments of the main effects; ε_{ij} is a random error.

TABLE IV. LSM (LEAT SQUARES METHOD) ESTIMATION OF PARAMETERS OF A TWO-FACTOR MODEL OF THE INFLUENCE OF THE GROUP AND AGE OF CHICKENS ON HEMOGLOBIN CONTENTS

Parameter	**B (coefficient)**	**Standard error**	**Student t-test**	**Significance**	**95% Confidence interval**	
					lower bound	**upper bound**
Intercept	103,693	2,043	50,756	0,000	98,442	108,945
[GROUP=1]	5,290	1,938	2,729	0,041	0,308	10,272
[GROUP=2]	0	,	,	,	,	,
[AGE=14]	-24,560	2,238	-10,974	0,000	-30,313	-18,807
[AGE=28]	7,010	2,238	3,132	0,026	1,257	12,763
[AGE=38]	0	,	,	,	,	,

We will explain Table IV. In it, the constant μ_0= 103.7 g/l, the effects of group 2 (control) and the age of 38 days are taken as zero. The effects of the experimental group 1 (combined) and age are counted from this level; thus, the hemoglobin contents in the blood of chickens of the combined experimental group at the age of 28 days is characterized by additives α_1= 5.29 g/l and β_{28} = 7.01 g/l. The values of the 95% confidence interval of all effects does not include zero, which indicates the statistical significance of all the parameters of the model.

Additional information about the significance of the mean difference is presented in the Table V of multiple comparisons and Table VI of homogeneous subgroups of chicken age. obtained when using the Tukey multiple comparison test [13].

TABLE V. A POIST-HOC PAIRWISE COMPARISONS BY TUKEY CRITERION (DIFFERENCES OF THE HEMOGLOBIN CONTENTS; G/L)

(I) Age (days)	**(J)Age (days)**	**Mean difference (I-J)**	**Standard error**	**Significance (2-way)**	**95% Confidence interval**	
					lower bound	**upper bound**
14	28	-31,5700	2,23796	0,000	-38,8521	-24,2879
	38	-24,5600	2,23796	0,000	-31,8421	-17,2779
28	14	31,5700	2,23796	0,000	24,2879	38,8521
	38	7,0100	2,23796	0,057	-0,2721	14,2921
38	14	24,5600	2,23796	0,000	17,2779	31,8421
	28	-7,0100	2,23796	0,057	-14,2921	0,2721

TABLE VI. HOMOGENEOUS SUBGROUPS OF CHICKEN AGE ACCORDING TO TUKEY TEST (THE LEVEL OF SIGNIFICANCE OF THE TEST FOR THE DIFFERENCE BETWEEN SUBGROUPS P = 0.05)

Age, days	**N**	**Hemoglobin contents, g/l**	
		1	***2***
14	3	82,6600	
38	3		107,2200
28	3		114,2300
Difference test significance		1,000	0,057

In column 3 of Table V the mean differences are given, that is"effects" caused by the age of chickens, and in column 5 p-values of differences are given. These values exceed the critical level only for the difference in hemoglobin contents by 28 and 38 day-old chickens, 95% confidence interval for the difference in these subgroups include zero. Therefore, the null hypothesis is accepted that there is no difference in the index for these chicken subgroups. On the contrary, p-values of hemoglobin differences by 14 day-old chickens, on the one hand, and 28 and 38 day-old chickens on the other hand, are less than the critical level (not worse than 0.005), 95% confidence interval of the difference between these subgroups does not include zero, therefore, the null hypothesis of the absence of a difference in the indicator for these subgroups of chickens is rejected in favour of the alternative, i.e. with a probability of 95%, it can be stated that there is a difference in the hemoglobin contents in the blood of "young" chickens in comparison with the older ones.

From Table VI of homogeneous subgroups it also follows that groups 28 and 38 day-old chickens are statistically indistinguishable: they form a homogeneous subgroup 2. At the same time, the two-sided significance level of the criterion for differences in subgroup 2 is only slightly higher than the standard value of 0.05; by the transition to one-sided p-level it can be stated that the maximum hemoglobin contents is observed by the 28 day-old chickens.

The visual result of the generalized linear model procedure is the graphical representation of the influence of factors on the studied indicator generated by the program - Figure 3.

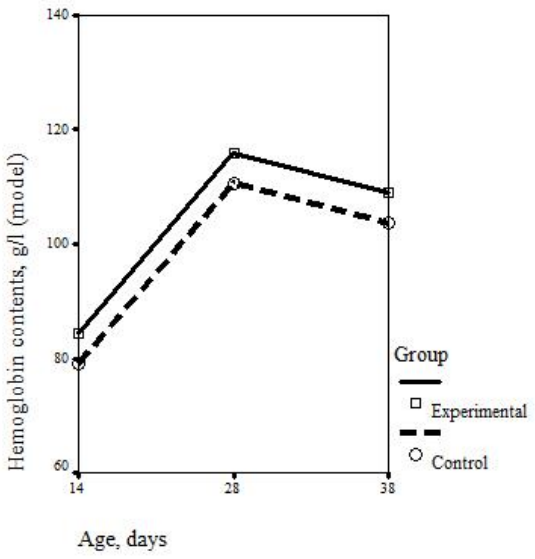

Fig. 3. Hemoglobin contents in the blood of chickens of the combined experimental and control groups (two-factor model)

It is seen that the age of chickens affects the hemoglobin contents more strongly than the addition of microbiological preparations to the basic diet, and the "symbathical" graphs - polygonal lines for the experimental and control groups of chickens are parallel to each other. This is the result of the linearity of the model: the real dependencies look different (Figure 4).

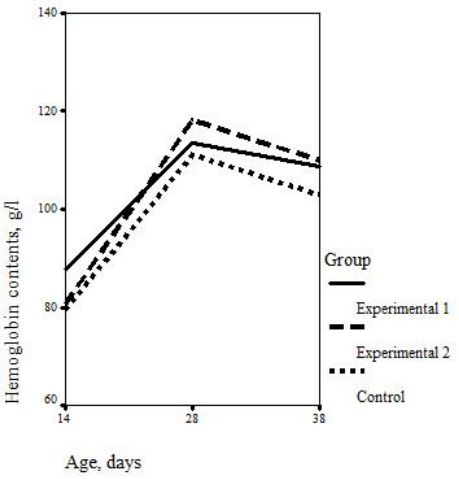

Fig. 4. Hemoglobin contents in the blood of chickens of the experimental and control groups (actual data)

Comparing Figures 3 and 4, it can be concluded that the model reflects the patterns in the observed data "more accurately", since its parameters are determined by the least squares method and thus "cleared" from random fluctuations.

A combination of two experimental groups into one common group looks logical against the background of fluctuations: according to the empirical data in Figure 4 the polygonal lines are not simbathical to each other and even intersect. But as for the maximum at the 28-week mark, this fact is most likely not accidental, but rather logical, since the local maximum manifests itself on both empirical and model data.

IV. Conclusion

We gave only one example, but it is already clear from it that the use of intelligent technologies oriented towards the search for knowledge in data allows us to expand and deepen the analytical capabilities of the researcher. In particular, the use of the generalized linear model procedure makes it possible to estimate the statistical significance of the effects of factors and their confidence limits, which ensures the reliability of the findings from the experimental data. It is also essential that for the construction of factor models it is possible to limit oneself to the analysis not of all the measurements, but only averages over their subgroups. This allows you to build models based on the materials of publications, which, as a rule, contain only the average experimental data, while the measurement results for replicates are not available.

References

[1] E.V. Popov, K.A. Semyachkov, "Problems of economic security of digital society in the context of globalization," Economy of the region, vol.14, Issue 4, pp.1088-1101. 2018.

[2] E.A. Skvortsov, E.G. Skvortsova, I.S. Sandu, G.A. Iovlev, "The transition of agriculture to digital, intellectual and robotic technologies," Economy of the region, vol.14; Issue 3, pp.1014-1028. 2018.

[3] E.P. Vasiliev, V.I. Oreshkov, "Improving the decision-making process in the economy and business through the use of data mining," Basic Research. № 9, pp.965-971. 2012.

[4] J. Han, M. Kamber, J. Pei, "Data mining: concepts and techniques," 3rd ed., Morgan Kaufmann, Elsevier, p. 744. 2012.

[5] V.M. Kuznetsov, "Line breeding and Holsteinization: assessment methods, status and prospects," Problems of the biology of productive animals, №. 3, pp. 25–79. 2013.

[6] V.G. Shumetov, A.S. Kolomeychenko, V.S. Buyarov, S.Yu. Metasova, "Multiple contrast of means in the empirical agricultural research," Bulletin of agrarian science, №. 4 (67), pp. 113-122. 2017.

[7] A.F. Melnik, V.G. Shumetov, B.S. Kondrashin, "Using the procedure of a generalized linear model for analyzing the results of agricultural research," Success of modern natural science, №. 2, pp. 23-29. 2019.

[8] I.A Katsko, "Intellectual data analysis and modeling the dependence of grain yield on costs," Scientific journal of the Kuban State Agrarian University, №. 36 (2), pp. 213-222. 2008.

[9] M.A. Zyablitseva, "The productivity of broiler chickens when using microbiological preparations "URGA" and "Baikal EM-1"," Diss. ... to the farm sciences, Troitsk, p. 153, 2018.

[10] SPSS Base 8.0 for Windows. Application Guide. Translation, Copyright 1998 SPSS Rus, p. 397.

[11] M.V. Kozlov, "Imaginary replications (pseudoreplication) in environmental studies: a problem not noticed by Russian scientists," Journal of General Biology, vol. 64, №. 4, pp. 292–307. 2003.

[12] "Problems of the environmental experiment (Planning and analysis of observations)," ed. Corr. RAS G.S. Rosenberg and Dr. Sc. D.B. Gelashvili; draf. and com. Dr. Sc. V.K. Shitikova, Tolyatti: SamSC RAS; Kassandra,. p. 274, 2008.

[13] A.M. Grzhibovskiy, "Analysis of three or more groups of quantitative data," Human Ecology, №. 3, pp. 50-58. 2008.