Available online at www.elixirpublishers.com (Elixir International Journal)

Applied Mathematics

Elixir Appl. Math. 61 (2013) 16879-16884

ISSN: 2229-712X

Theoretical Basis of Identification of Genotypes by Their Phenotypes in Process of Selection in Segregating Generations

I.M. Mikhailenko[1] and V.A. Dragavtsev[2]

[1]Scientific Deputy Director, Head of Lab of Informatic-Measuring Systems, Russia.

[2]Plants Ecologic Physiology Lab Agrophysical Institute, St.-Petersburg, 195220, Grazhdansky prospect 14, Russia.

ARTICLE INFO

Article history:

Received: 1 June 2013;

Received in revised form:

24 July 2013;

Accepted: 2 August 2013;

Keywords

Genotypes,
Phenotypes,
Mathematical Models,
Identification,
Breeding Technologies,
Genetic-Physiological Systems,
Evaluation.

ABSTRACT

The formalized theory of the identification of genotypes by their phenotypes in modern breeding technologies is offer. It erect on the mathematical model of the interaction "genotype-environment".

1. Introduction

In our first paper has presented the main principles of the simulation system for interaction "genotype-environment", and shows the possibility of the proposed models for a number of the solution of basic problems of modern genetics and breeding [1]. Of all the problems solved with the help of such models, we select the most important (Fig. 1): 1) assessment of the mechanisms of transgressions and the selection of parental pairs for a desirable result mating, 2) evaluation of the contributions (in productivity) of genetic-physiological systems of parental pairs, 3) prediction of transgressions of breeding traits in the F_2 generation and obtaining a population mating F2, 4) identification and selection of genotypes by their phenotypes. These problems in their indissoluble unity represent stages in the same general problem - the strict control of genetic-breeding process. Development of the theory to address this problem is one of the key perspectives of the modern genetics. Figure 1 shows the scheme of relations between tasks.

As can be seen from the scheme, the problem forms a closed loop control by the selection process. Here, the initial step is the selection of the parent pairs for a desirable result mating. To optimize this selection is used, the prediction of the results of parental mating pairs, and actual results are analysed on the crossing point for the identification of genotypes by their phenotypes. They are also used to correct the models predict the results of crossing.. At each of these stages are used mathematical models of the"genotype-environment" interaction.

2. The General Scheme of the Problem of Identification of Genotypes by Phenotypes

In content, the selection of genotypes by phenotypes is quite complicated in terms of scientific classification of the information problem. Its aim is to find (or formation) genotype, which includes the maximum number of positive shifts given signs of breeding. Therefore, the algorithm is based on the principle of background characters [2] and the principle of differently-directional shifts of quantitative trait of individual genotype under the influence of genetic and environmental causes in the two-dimensional character's coordinates [3].

The ideal background character has zero genetic variation, so that it "writes" by own variation only the ecological fluctuation of limiting factor [2]. Respectively, the individual, which have a deviation of background character from average of population - this is a plus-modification, got the better micro-ecological niche. At the same time, if the breeding character of this individual is shifted in the positive direction from population's average, then this is a common modification and it does not make sense to select. If the other individual's background character is expressed at the level of average population value and breeding character shifted in the positive direction from average in the population, this is recombination (or mutation), and it is necessary to select for productive breeding.

3. The Informational Sense of the Problem

Here is more information situation in which we need to solve this problem. In view of the concepts discussed problem of identification of genotype by phenotype is essentially reduced to the identification of breeding character on which this individual can be selected or not for further breeding. We have a mathematical model of the "genotype-environment" interaction that allows us to predict quantitative traits of individuals or populations [1]. In addition, we have data for monitoring of all environmental factors affecting, as controllable and uncontrollable throughout the growing season of plants, as well as data on the actual growth and development of individuals and populations, ranging from planting until the end of quantitative results, considered by us as signs of.

Tele:
E-mail addresses: dravial@mail.ru

That's the end results we have to classify individuals. The result of this classification will be the division of the whole set of phenotypes new generation of fissile subset of genotypes that have different sets of quantitative traits, some of which are economically valuable. The number of individuals in certain subsets may be very small or even be a few units. Given that the formation of these subsets will require modelling of states of each individual, without which it is impossible to correct classification, after the formation of subsets of genotypes is expedient to find the boundary separating them, which will further simplify and speed up the class ification of the individual of other generations, not resorting to their modelling. With this approach, the first stage of classification is the training, where as a "non-ideal (real) teachers' use mathematical models of individuals.

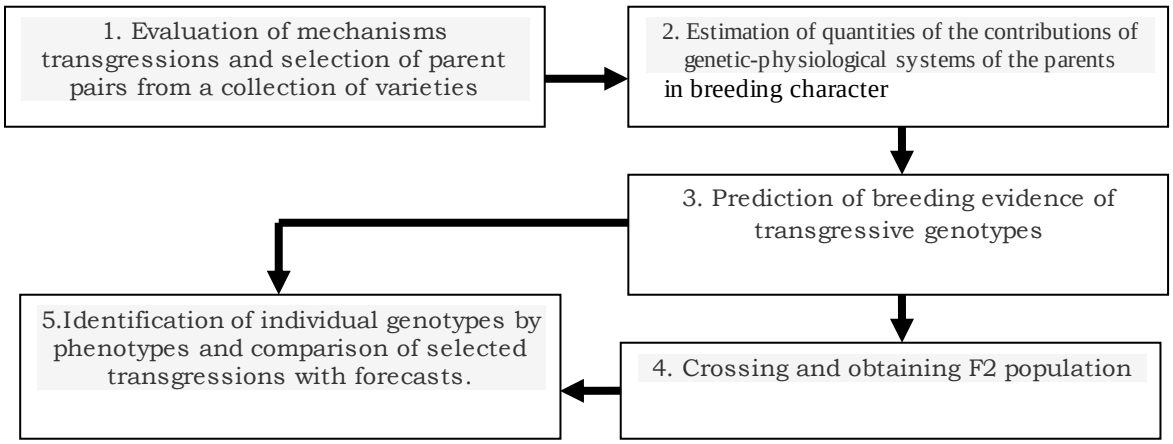

Figure 1. Block diagram of the relationship management tasks genetic-breeding process

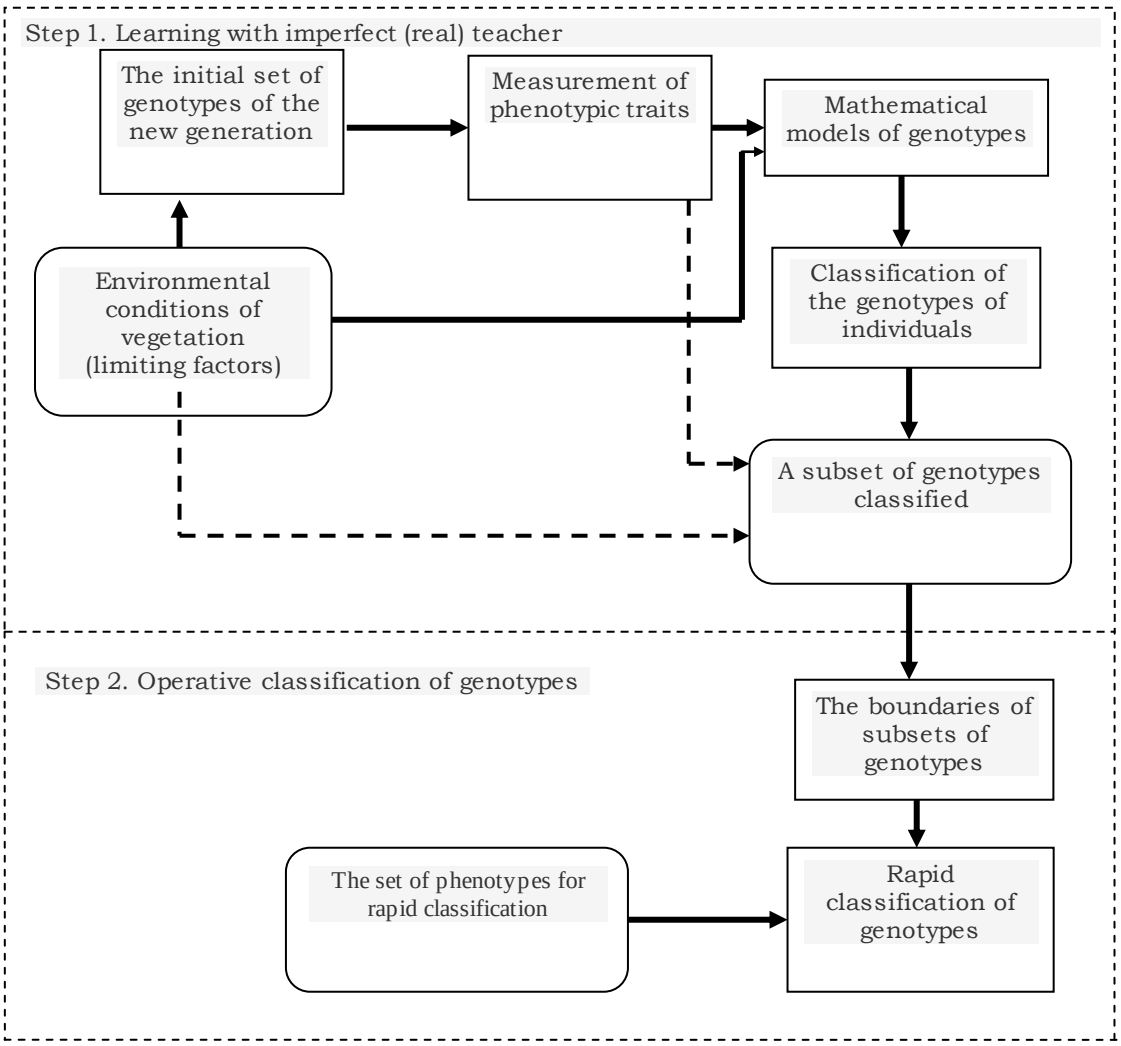

Figure 2. Block diagram of a general algorithm for the classification of genotypes by phenotypes

The result of training will determine the number of possible classes and identification of the boundaries of subsets of classes, and the second phase of the task itself is the operative classification of the genotypes of individuals according by their phenotypic characteristics. Figure 2 shows the structural scheme of the classification problems of identification of genotype by phenotype.

In developing the algorithm for the general classification of genotypes based on the above principles will give a brief description of the evolution of the models used in modern genetics. So in 1984 the two existing models that describe the relationships genes - characters (the first model of Mendel [4], the second – R. Fisher, C. Mather, S. Wright [5]), added a third - a model of eco-genetic organization of quantitative traits (MEGOQT)[6]. In the period from 1984 to 2011 been theoretically predicted and experimentally confirmed 23 issues of this model. The most important are - interpreting of nature and prediction: transgression, environmentally dependent heterosis, changes the signs and levels of genotypic and genetic (additive) correlations, effects of the interaction"genotype-environment", changes in the numbers of genes and genetic variation of the amplitude characteristics of productivity, genetic homeostasis, and others [7]. In 2008, the performance model has been fully confirmed at the molecular level, together with the German's geneticists [8], which translate the model in 1984 to the rank of the theory of eco-genetic organization of quantitative traits (TEGOQT). This theory has led to a change in the classical model of R. Fisher

$$\Psi i = \mu + \gamma_i + \pi_i \qquad (1)$$

where Ψ_i - phenotypic value of a quantitative trait in the i-th individual, μ-average trait in the population, γ_i - genotypic deviation from the mean trait values of i- individuals, π_i – i- individual environmental deviation from the mean of population.

The new model proposed in [9], describes the efficiency integral property of the i-th individual:

$$\Psi i = \mu \text{ (plant's productivity)} = \gamma_{attr,i} + \gamma_{mic,i} + \gamma_{ad,i} + \gamma_{imm,i} + \gamma_{ef,i} + \gamma_{tol,i} + \gamma_{ont,i} + \gamma_{com,i} + \pi_{com,i} + \pi_{ont,i} + \pi_i , \qquad (2)$$

where Ψi - phenotypic trait value of productivity at the i-th individual; μ - average productivity of the population; γ_{attr} - deviation of attraction products of photosynthesis from the stems and leaves in the ear; γ_{mic} - deviation of the distribution of products of attraction between the grains and chaff in the ear; γ_{ad} - the influence of the deviation system adaptability to the products effectiveness, measured by total dry biomass of plants, γ_{imm} - impact on the productivity of horizontal resistance; γ_{ef} - payment by biomass of limiting factors of soil nutrition; γ_{to}l - a deviation tolerance to density; γ_{ont} - deviation of the genetic variability in the duration of the phases of ontogenesis ; γ_{com} - deviation of the genetic competition of plants for moisture, food, light, etc.; π_{com}-deviation of not the genetic competition caused by the unequal growth of the initial conditions, π_{ont} - deviation caused by the change of limitative factors in ontogeny between bookmarks and development mark, π_i - deviation caused by the influence of the environment.

Decipher each component of the model in the form of specific states of genetic-physiological systems and the modular structure of the model:

1) The system of attraction - the mass of stem and ear φ_{11} φ_{12} (commodity and non-market part of the plant);

2) Microdistribution. - weight of the grain φ_{21} and φ_{22} of non-grain spike (chaff, awns, etc.);

3) Adaptation (resistance to environmental stressors and chemical environment) - the degree of deceleration of growth processes under the influence of adverse environmental factors (stressors), speed and recovery time course of the normal growth of the processes;

4) Polygenic immunity - plant resistance to pests and pathogens with an array of diseases, development of plant protection substances and machinery;

5) The sensitivity (response) at doses of soil nutrition elements - parameters of the sensitivity characteristics of the productive to the doses of nutrients;

6) The tolerance to density - the parameters of sensitivity productive of indicators to the density of phytocenosis;

7) The variability of the periods of ontogenesis - is used in the selection for the "withdrawal" of the critical phases of the ontogeny of the stressor, "beating" in a critical phase.

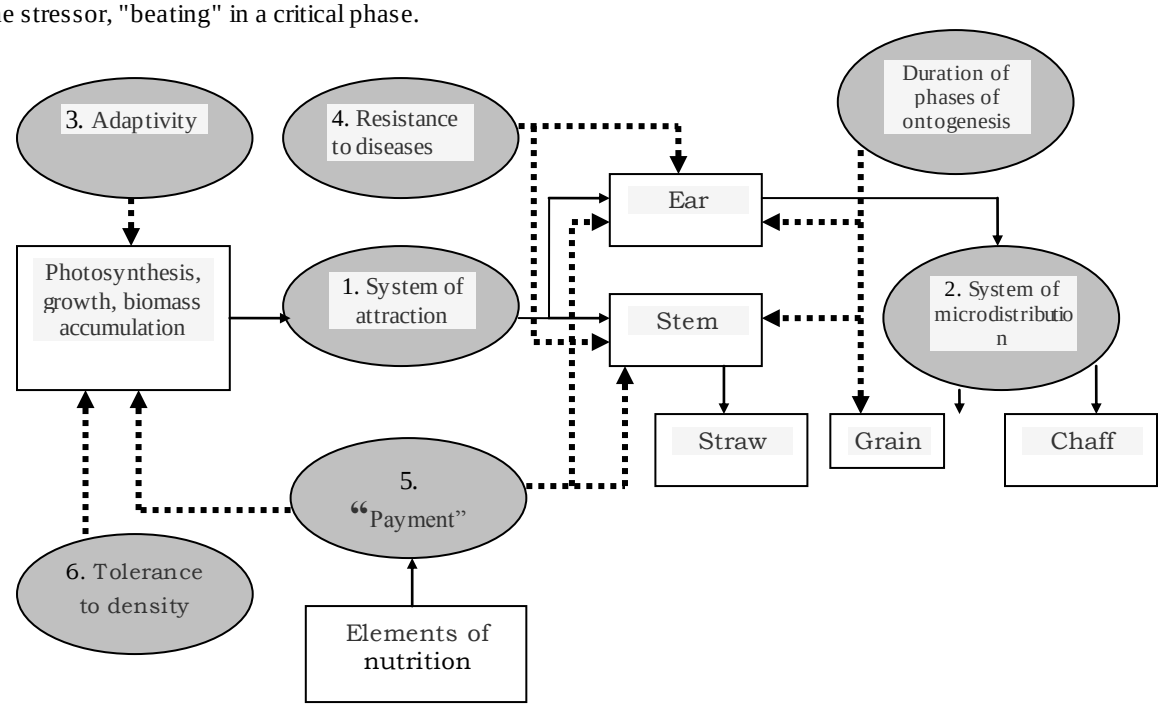

Figure 3. Scheme of the model of "genotype-environment" for crops

4. The Basic Mathematical Model

Reflect the given characteristics of the genetic-physiological systems of the first on the modular structure of the model of "genotype-environment" for the crops (Fig. 3)

This scheme corresponds to the mathematical model of the main (output) module [4]

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix}_{ij}=\begin{bmatrix}a_{11}(\varphi_3) & a_{12}(\varphi_2) & a_{13}(\varphi_1)\\ a_{21}(\varphi_2) & a_{33}(\varphi_3) & 0\\ a_{31}(\varphi_1) & 0 & a_{33}\end{bmatrix}\begin{bmatrix}x_1(t)\\ x_2(t)\\ x_3(t)\end{bmatrix}_{ij}+\begin{bmatrix}b_1(\varphi_5)\\ b_2\\ b_3\end{bmatrix}[u(t)]+$$

$$+\begin{bmatrix}0 & c_{12}(\varphi_3) & c_{13}(\varphi_3)\\ 0 & 0 & c_{23}\\ c_{31} & c_{32} & c_{33}\end{bmatrix}\begin{bmatrix}f_1(t)\\ f_2(t)\\ f_3(t)\end{bmatrix}+\begin{bmatrix}0 & 0 & 0 & d_{14} & 0 & d_{16} & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & d_{34} & 0 & d_{36} & 0\end{bmatrix}\begin{bmatrix}\Delta\varphi_1\\ \Delta\varphi_2\\ \Delta\varphi_3\\ \Delta\varphi_4\\ \Delta\varphi_5\\ \Delta\varphi_6\\ \Delta\varphi_7\end{bmatrix}+\begin{bmatrix}\xi_1(t)\\ \xi_2(t)\\ \xi_3(t)\end{bmatrix}, \quad (3)$$

$$t\in[t_0(\varphi_7);T(\varphi_7)]$$

where the following notation: x_{1i} - the mass of grain in the ear i-th individual, x_{2i} - the mass of chaff in the ear, x_{3i} - the mass of straw in the ear, u-provision (control) of nitrogen nutrition; f_1 - luminous efficiency factor, f_2 - thermal efficiency factor products, f_3 - moisture as a factor ofproductivity; $\Delta\varphi_1$ ... $\Delta\varphi_7$ - the influence of genetic- physiological systems; ξ_1, ξ_2, ξ_3 - random perturbations, reflecting the uncertainty in the information model; a_{kj}, b_k, c_{kj}, d_{kj} - the dynamic parameters of the model.

We represent the model (2) in a more compact vector-matrix form in which all variables and parameters are combined into the corresponding vectors and matrices.

$$\dot{X}_i = A(j_1,j_2,j_3)X(t)+b(j_5)u(t)+C(j_3)F(t)+D^*[j_4(t)j_6(t)]+\mathbf{x}(t), \quad (4)$$

$$t\hat{I}\,[t_0(j_7);T(j_7)]$$

Model (4) determines the state of the i-th individuals, and the effect of limiting factors, the differential for all individuals, as the action of genetic-physiological systems, leading to perturbations of the states of individuals and the emergence of environmental and genetic variance. These perturbations can be represented as follows

$$DX_i = X - X_i = U_E DE_i + U_j D\mathbf{j}_i, \quad (5)$$

$$U_E = \frac{\P X_i}{\P E},\quad U_j = \frac{\P X_i}{\P \mathbf{j}},$$

where the UE, Uφ-state vectors of the sensitivity functions of the module, respectively, to environmental and genetic perturbations; ΔEi, Δφi - vectors of the observed variations in environmental factors and unobservable genetic effects.

Equation (5) reflects the modelled contributions of environmental and genetic factors. However, the breeder usually has to do with the observed variations in the recognition of magnitude, which we denote as ΔY_i. In this case, the meaning of the classification of genotypes is to establish the causes of observed variations in characteristics of individuals compared with average of the po pulation values. In the event that such causes are environmental factors, then we are dealing with modifications of the same genotype, and in establishing the genetic basis - with a new genotype.

We introduce the quadratic functional classification quality

$$J_i = \int_{t_0}^{t}[(DX_i(t)-DY_i(t))^T(DX_i(t)-DY_i(t))]dt, \quad (6)$$

whose meaning is "balancing" of the simulated and observed shifts of breeding traits by unobserved effects of seven genetic-physiological systems.

Required to minimize the criterion (6) of unobservable influences of genetic factors for the known variations $\Delta\varphi_i$ environmental factors $\Delta E_i(t)$, which will assess the "contributions" of each of the genetic-physiological systems.

Now, having regions of admissible values of the impacts of genetic-physiological systems for individual genotypes Ω_k, k = 1,2,3 .... K-indices of the genotypes (classes), we can determine the decision rule of classification

$$k_i = k^*, if\ Dj_i \hat{I}\ W_k^* \cdot \quad (7)$$

With background information on deviations of environmental factors $\Delta E_i(t)$ and the observed deviations of quantitative traits i-th individual $\Delta Y_i(t)$ we have by the procedure (2) - (7) may determine that it belongs to the genotype with the specified tag-mi. Fixing at

the time of this procedure for each class of sets (ΔE_k (t), ΔY_k (t)), we can construct a boundary between the genotypes in the space of environmental deviations and variation of quantitative traits

$$L_k = F_{k,k+1}(DE, DY), \quad (8)$$

Where F (.) - special function approximation bounds.

In this case, the decision rule is as follows

$$k_i = k^*, \; if \; F_{k,k+1}(DE, DY) - c \le 0, \quad (9)$$

$$k_i = k^* + 1, \; if \; F_{k,k+1}(DE, DY) - c > 0,$$

Here, c- the threshold value, which is one of parameters of the decision rule.

Thus, we examined the whole scheme of identification of genotypes for the observed phenotypes shown in Figure 1. This entire procedure is prior separation of the genotypes of individuals and modifications to the stage of teach "teacher" more than a s imple decision rule (8) (9). Due to the fact that this algorithm can make mistakes, then that teacher is "imperfect," or, more accurately, a "real teacher".

5. Algorithm of Identification

Consider the Hamiltonian of the system

$$H_i = (DX_i(t|DE) - DY_i(t))^T (DX_i(t|DE) - DY_i(t)) + \quad (10)$$

$$+ l^T [A(j_1, j_2, j_3) DX_i(t) + b(j_5) D_i u(t) + C(j_3) DF_i(t) + D^*[Dj_4(t) Dj_6(t)]$$

Where: the λ-vector of conjugate variables, which is a solution of the system in reverse time

$$\dot{l}_i = -\frac{\partial H}{\partial DX} = -2[(DX_i(t|DE) - DY_i(t)) + A^T(j_1, j_2, j_3) l_i], \quad (11)$$

$$t \in (t, t_0), \; l(t) = 0.$$

Given the introduction of new auxiliary variables identification procedure is to minimize the criterion (5) on no observed effect of genetic-physiological systems will look like the following multi-step procedure

$$Dj_{i, j+1} = Dj_{i, j} - g_j \frac{\partial H_i}{\partial Dj_{i, j}}, \quad (12)$$

Where: the j-number of iterations of the process of working to minimize the criterion (5).

Upon reaching the iterations (12) breakpoint conditions the estimates of impacts genetic-physiological systems in the future will be denoted by - $\Delta\varphi_i$*. With the obtained values of the separation vectors in a subset of the classes according to the rule (7), their boundaries are convenient to specify the system of inequalities

$$W_k : j_{lk\min} \le j_{lk} < j_{lk\max}, \; l = \overline{1,7}, \quad (13)$$

Where: l - the indices of genetic-physiological systems.

The system of inequalities (13), whereby the space of seven influences of genetic-physiological systems by separation of the individuals in the future we will be called "eco-genetic portrait of the genotype, bearing in mind that in the light of developing my theory (TEGOKP), he is the only possible representation of the differences in genotype.

Note that the vectors of the effects of genetic-physiological systems $\Delta\varphi_i$* we are only "label" or guidelines for the formation of subsets of a causal relationship

$$W_{kEX} : (\bar{DE}_{ki}, DX_{ki}), \; i = \overline{1, I_k}. \quad (14)$$

Here - the average value of the vector of environmental variations on the final phase between the periods.

That's it for these sets we construct a simple decision rules. To do so, combine the vector of environmental causes and consequences of the vector ΔY in the unit vector of $Z^T = [\bar{DE}, \Delta Y]^T$. Then we define the basic statistical characteristics of the classes on the sets (14) - vectors of expectation and covariance matrix M_{Zk} K_{Zk}, as well as the same probability of occurrence of classes, which represent the ratio of estimates of numbers of individuals who have fallen into many separate classes I_k to the total number of individuals studied

$$p_k = \frac{I_k}{\sum_k I_k}. \quad (15)$$

For these characteristics, it is easy to construct a function separating the classes F (.)[10]

$$F_{k,k+1}(Z) = Z^T (K_{Zk} - K_{Zk+1}) Z + 2(M_{Zk} K_{Zk} - M_{Zk+1} K_{Zk+1}) Z \quad (16)$$

and the threshold number **c** of rules (9)

$$c = 2\ln\frac{p_{k+1}}{p_k} + \ln\frac{|K_{Zk}|}{|K_{Zk+1}|} + M_{Zk}^T K_{Zk}^{-1} M_{Zk} - M_{Zk+1}^T K_{Zk+1}^{-1} M_{Zk+1} \quad (17)$$

Where: $|K|$ - the norm of the matrix K.

Obviously, in accordance with rule (9) for each new implementation of the causes and effects, but the signs of $Z^T=[\Delta\varphi, \Delta Y]^T$ we will need to compare the pairs all the possible genotypes.

As mentioned above, the vectors of the impacts of genetic-physiological systems $\Delta\varphi_i^*$ serve us only "labels" for the formation of subsets of causal relations. However, to solve the following problems of selection, we need a static version of the model of "ecological disturbance-response genetic-physiological systems". To do this we need to form a set of identification, which we estimate the parameters W of the desired model

$$\Delta j = W^T \Delta\varphi \quad (18)$$

Here, as an illustration for the solution of the problem we have considered only one of the modules of the general model of the "genotype-environment". In the case of the need to include in the genotypes of characteristic other quantitative trait dimension of the problem can be increased without changing the essence of the approach. In this important feature of the developed theory of identification of genotypes is that it provides a solution to this problem during the whole period of ontogenesis, starting from first phases of ontogenesis, namely, the modules of the lowest level of the hierarchy, completing her final product output modules.. This greatly improves the reliability of solving the problem and allows for more productive use of all the genotypic variability in the possession of the breeder-geneticist.

6. Conclusions

We propose a formalized theory of the identification of genotypes by their phenotypes, including:

Assessment (using the mathematical model and a special algorithm optimizing unobserved variables sevencontributions of genetic-physiological systems in theproductivity of the individual);

Classification of individuals with a given system of inequalities in the levels of contributions received by the genetic-physiological systems in the productivity of individual;

Formation for each of the classes of individual's subsets of variations of environmental factors and variations in quantitative traits with simultaneous estimation of multidimensional statistical characteristics of these subsets combined;

Determination (on the statistical characteristics of the environmental factors variability and variability ofquantitative traits) boundaries of individual classes ofgenotypes, which can be implemented in a simplified algorithm for identification of genotype by phenotype.

References

Mikhailenko I.M., Dragavtsev V.A. (2010) The basic principles ofmodeling mathematical system"genotype-environment"interaction / /Agricultural Biology (rus). 2010, № 3, p.31-34.

Dragavtsev V.A. The method of estimation the role of heredity and environment in the development of plant characters that does not require a change of generations./ /The Botanical Magazine (rus) 1966, № 7, pp. 939-946.

Dyakov A.B., Dragavtsev V.A. Differently directional shifts of individual quantitative trait of an organism under the influence of genetic and environmental causes in the two-dimensional system of character's coordinates. In: Dragavtsev V.A. Eco-genetic algorithms for the inventory of the gene pool, and methods for creating of plant varieties for yield, stability and quality (Methodic recommendations (new approaches). St. Petersburg, VIR, 1994, pp. 22-47.

Mendel G. Versuche uber Pflanzen Hybriden. Verhandlungen des naturforschenden Vereins in Brunn, 1865, Bd 4, S. 3-47.

Wright S. The genetics of quantitative variability.Quantitative inheritance. Edinburgh, 1950, Ed. London, 1952.

Dragavtsev V.A. Litun P.P., Shkel, N.M., Nechiporenko N.N. The model of eco-genetic control of quantitative traits of plants. / /Reports of AS USSR (rus). 1984, T. 274, № 3, pp. 720-723.

Kocherina N.V., Dragavtsev V.A. The book "Introduction to the theory of eco-genetic organization of quantitative traits and theory of plant breeding indexes", AFI, St. Petersburg, 2008, 87 pp.

Chesnokov Y.V., Pochepnya N.V., A. Berner, W. Lovasser, Goncharova EA, Dragavtsev V.A. Eco-genetic organization of quantitative traits of plants and mapping loci determining agronomically important characters in common wheat. // Reports of the Academy of Sciences (RAS), 2008, T. 418, № 5, pp. 1-4.

Dragavtsev V.A. Eco-genetic screening of the gene pool and methods of creating plants varieties for yield, stability, quality, (new approaches), VIR, St. Petersburg, 1998, pp. 52 .

M. de Groot. Optimal statistical decisions. Springer-Verlag, 1974.