ISSN 1068-3674, Russian Agricultural Sciences, 2007, Vol. 33, No. 2, pp. 130–132. © Allerton Press, Inc., 2007.
Original Russian Text © M.N. Borisevich, 2007, published in Doklady Rossiiskoi Akademii Sel'skokhozyaistvennykh Nauk, 2007, No. 2, pp. 49–50.

MODELING

Automation of Modeling and Forecasting in Veterinary Medicine

M. N. Borisevich

Vitebsk State Academy of Veterinary Medicine, Vitebsk, 210619 Russia

Recommended by D.S. Strebkov, Academician of the Russian Academy of Agricultural Sciences
Received August 22, 2005

Abstract—A modeling and forecasting automation system created for veterinary medicine needs and based on principles of internet technologies with the use of a three-tier client/server architecture with a dedicated application server is briefly described.

DOI: 10.3103/S1068367407020206

Decisive success in eliminating infectious diseases of animals is impossible without knowledge of all regularities of the epizootic process. Modeling and forecasting are the most important links of theoretical investigations in this direction. The modeling and forecasting automation system (MFAS) is intended for creating a bank of mathematical models and forecasts used directly for the needs of veterinary medicine. Structurally, it consists of the following units: initial data bank (IDB), applications, mathematical models bank (MMB), optimal mathematical model and output information block, prepared models bank (PMB) and prepared forecasts bank (PFB), and PMB and PFB warehouses.

The system is realized in three variants: (I) functional blocks, including control application, located in an individual computer, the user can work with the system in an offline mode; (II) the system is located on a dedicated server of the local computer network (corporate network), access to the units is organized from any work place; (III) operations with the MMB are realized in the Internet network, the user can work with data regardless of where he is.

The initial data bank (IDB) is an aggregate of several databases, each of which is intended for arranging numeric arrays constructed on materials of statistical veterinary records. Selection of the arrays for constructing the mathematical model and forecast is done by the user in one of the following ways: either by marking the necessary fragment with a manipulator or by manually filling out a specially created screen form.

The data selected for modeling is processed by the application with the enlistment of the mathematical models bank, which is the warehouse of more than 100 known analytical relations. The computer automatically looks through each model for the given array and calculates the coefficient of determinacy for it. Then the maximum value is selected from the set of values of this coefficient; it determines the optimal mathematical model for which the necessary parameters are calculated.

Graphs of the optimal model and initial data (with the calculated parameters) are kept in the prepared models bank (PMB), and the forecast constructed by means of the optimal model is kept in the prepared forecasts bank (PFB). When necessary, they can be called for by the user for the purpose of analysis or practical use.

The mathematical models bank (MMB) is the warehouse for three groups of models—linear, polynomial, and nonlinear—which can be used in veterinary medicine for a correct description of epizootological processes. The components of the MMB approximate the data of the veterinary records with calculated statistical parameters. Mathematical formation of the problem of approximation for components of the MMB consists in the following. The dependence of the quantity (numerical value) of a certain property of a random process or physical phenomenon Y on another variable property or parameter X, which in the general case can also be referred to a random variable, is recorded on a set of points x_k by a set of values y_k, in which case at each point the recorded values y_k and x_k represent the real values $Y(x_k)$ with random error V_k distributed, as a rule, normally. On the basis of the set of values y_k, it is required to select such a function $f(x_k, a_0, a_1, \ldots, a_N)$ for which the dependence $Y(x)$ would be represented with minimum error. From here follows the conditions of approximation:

$$y_k = f(x_k, a_0, a_1, \ldots, a_N) + V_k.$$

The MMB models provide for assigning the form of the function $f(x_k, a_0, a_1, \ldots, a_N)$ and determining the numerical values of its parameters $a_0, a_1, \ldots, a_N$ ensuring a minimum error of approximation to the set of values y_k. The error of the approximation is calculated by the least squares method. For this purpose the function of the squares of the residual errors is minimized:

$$F(a_0, a_1, \ldots, a_N) = \sum_k [f(x_k, a_0, a_1, \ldots, a_N) - y_k]^2.$$

To determine parameters $a_0, a_1, \ldots, a_N$, the residual error function is differentiated with respect to all parameters, the partial differential equations obtained

are equated to zero and solved altogether for all values of the parameters.

The simplest method of LSM approximation of arbitrary data s_k is by means of a linear polynomial, i.e., a function of the form $y(t) = a + bt$. With consideration of the discreteness of the data with respect to points t_k, for the residual error function we have:

$$F(a, b) = \sum_k [(a + bt_k) - s_k]^2.$$

We differentiate the residual error function with respect to arguments a, b, equate the equations obtained to zero, and form two normal equations of the system:

$$\sum_k (a + bt_k) - s_k \equiv a\sum_k 1 + b\sum_k t_k - \sum_k s_k = 0,$$

$$\sum_k ((a + bt_k) - s_k)t_k \equiv a\sum_k t_k + b\sum_k t_k^2 - \sum_k s_k t_k = 0.$$

The solution of the given system of equations in an explicit form for K readings:

$$b = \left[K\sum_k t_k s_k - \sum_k t_k \sum_k s_k\right] / \left[K\sum_k t_k^2 - \left(\sum_k t_k\right)^2\right],$$

$$a = \left[\sum_k s_k - b\sum_k t_k\right] / K.$$

We use the values of the coefficients obtained in the regression equation $y(t) = a + bt$. The coefficients of any other forms of regression, differing only in the awkwardness of the corresponding expressions, are calculated by a similar method.

Linear regression in the system is done with respect to vectors of the argument X and readings Y by the functions: *intercept*(X, Y), which calculates parameter a, the vertical displacement of the regression line; *slope*(X, Y), which calculates parameter b, the slope of the regression line. The arrangement of the readings with respect to argument X is arbitrary. Pearson's correlation coefficient can be calculated additionally by function *corr*(X, Y). The closer it is to 1, the more accurately the data correspond to a linear dependence.

Univariate polynomial regression with an arbitrary degree n of the polynomial and with arbitrary coordinates of readings in the system is performed by the functions: *regress*(X, Y, n), which calculates vector S for the function *interp*(…), in the composition of which are coefficients k_i of a polynomial of degree n; *interp*(S, X, Y, x), which restores the values of the approximation function on the x coordinates. The function *interp*(…) realizes calculations by the formula:

$$f(x) = k_0 + k_1 x^1 + k_2 x^2 + \ldots + k_n x^n \Sigma_i k_i x^i.$$

The values of coefficients k_i can be extracted from vector S by the function *submatrix*(S, 3, *length*(S), 0, 0).

Polynomial regression with the use of polynomials of degree 1, 2, 3, 4, 5, 6, 7, and 8 is calculated in the system. The degree of the polynomial is usually set not more than 4–6 with its successive increase while checking the standard deviation of the approximation function from the actual data. It should be noted that, as the degree of the polynomial increases, the approximation function approaches the actual data, and for a degree of the polynomial equal to the number of readings of the data—1, it turns into the data interpolation function, which does not correspond to the tasks of regression.

The function *regress* with respect to the entire set of points creates one approximating polynomial. For large coordinate intervals with a large number of readings and sufficiently complex dynamics of the change in data, it is recommended to use sequential local regression by segments of polynomials of small degrees. In the system this is performed by segments of second-degree polynomials by the function *loess*(X, Y, *span*), which forms a special vector S for function *interp*(S, X, Y, x). The argument *span* > 0 in this function (of the order 0.1–2) determines the size of the local region and is selected with consideration of the character of the data and necessary degree of their smoothing (the greater *span*, the greater the degree of data smoothing). When modeling any random processes and signals at a high noise level, the optimal value of the parameter *span* can be determined from minimum of the mean-square approximation.

In the system it is possible to perform regression with approximation to a function of a general form as the weighted sum of functions $f_n(x)$:

$$f(x, K_n) = K_1 f_1(x) + K_2 f_2(x) + \ldots + K_N f_N(x).$$

In this case, the functions $f_n(x)$ themselves can be of any type, including nonlinear. On one hand, this markedly increases the possibility of an analytical representation of the regression functions and, on the other, requires from the user certain skills in approximating experimental data by combinations of rather simple functions.

Generalized regression with respect to vectors X, Y, and f is realized by the function *linfit*(X, Y, f), which calculates the values of the coefficients K_n. Vector f should contain the symbol notation of functions $f_n(x)$. Coordinates x_k in vector X can be any, but arranged in increasing order of the values of x (with corresponding readings of the values of y_k in vector Y.

A second form of nonlinear regression is realized by fitting parameters k_i to the given approximation function with the use of the function *genfit*(X, Y, S, F), which restores coefficients k_i providing the minimum root-mean-square error of the approximation of the regression function to the input data (vectors X and Y of the coordinates and readings). The symbol expression of the regression function and symbol expressions of its derivatives with respect to parameters k_i are written in vector F. Vector S contains the initial values of coeffi-

cients k_i for solving a system of nonlinear equations by the iterative method.

A number of regression functions, in which the parameters of the functions are selected by the system independently, is provided for simple standard approximation formulas. They include the following functions: exp*fit*(X, Y, S), which restores the vector containing coefficients a, b, and c of exponential function $y(x) = a\exp(bx) + c$. Into vector S are introduced the initial values of coefficients a, b, and c of the first approximation *lgsfit*(X, Y, S), ditto for expression $y(x) = a/(1 + c\exp(bx))$; *pwrfit*(X, Y, S), ditto for expression $y(x) = ax^b + c$; *sinfit*(X, Y, S), ditto for expression $y(x) = a\sin(x + b) + c$, which fits the coefficients of the sinusoidal regression function; *logfit*(X, Y), ditto for expression $y(x) = a\ln(x + b) + c$, assignment of the initial approximation is not required; *medfit*(X, Y), ditto for expression $y(x) = a + bx$, that is, for the linear regression function, assignment of the initial approximation also is not required, the graph is a straight line.

The materials given in the article are used for solving a number of problems of veterinary medicine related to forecasting infectious diseases of farm animals. The results obtained attest in behalf of the proposed approach, which on the whole provides a satisfactory description of real epizootic processes.