Modeling of technological processes for processing vegetable raw materials

N A Grechishnikova¹, 4, E A Strupan², N N Tipsina¹, O A Sizyh² and S K Manasyan³

¹Krasnoyarsk State Agrarian University, 90 Mira Prospect, 660049, Krasnoyarsk City, Russia
²Siberian Federal University, 82 Svobodnyy Prospect, 66004, Krasnoyarsk City, Russia
³Krasnoyarsk technical school of social technology, 70 Motoristov street, 660046, Krasnoyarsk City, Russia

4E-mail: nadusha_01@mail.ru

Abstract. To build a model of technological processes, linear and nonlinear regression equations are used to solve the problems of developing new food products. At the initial stage, the analysis was conducted of the study of the most real technical and technological economic system. As an object of study, depending on the purpose of the study, a particular enterprise, or group of enterprises, as well as an entire branch of the economy of a region or country, or perhaps the gross regional product of a certain subject of the Russian Federation or the country's GDP, can be chosen. Gooseberry processing complex is considered as an example. The correctness of the choice of model type can be proved very conditionally, as well as from the point of view of the researcher conducting the analysis, his system of preferences regarding models and modeling methods. As an example, consider some conditional company VRM, for which we simulate the initial data on sales for the last 4 years.

Models of technological processes (TP) for processing vegetable raw materials (VRM) in the form of linear and nonlinear regression equation systems or time series are necessary for solving the problems of developing new food products, optimizing their composition and structure, forecasting and planning economic processes [1,2].

The process of constructing this class of models includes a number of stages [2,3]. The initial stage is analysis, research of the most real technical and technological economic system. It is necessary to highlight its main structural elements and identify the main parameters that determine the dynamics of the studied object [4].

As an object of study, depending on the purpose of the study, a particular enterprise, or group of enterprises, as well as an entire branch of the economy of a region or country, or perhaps the gross regional product of a certain subject of the Russian Federation or the country's GDP, can be chosen. Gooseberry processing complex is considered as an example. Below is a hardware diagram (figure 1) of the production of a semi-finished product — product-puree gooseberries using microwave heating. This scheme of the complex allows rationally and with minimal losses to obtain a semi-finished product — product-puree with high physical and chemical parameters [5].
Figure 1. Machine-hardware diagram of the line for obtaining product-puree from gooseberries: 1 – receiving tray; 2 – inspection conveyor; 3 – washer; 4 – Bosch microwave oven; 5 – vertical wiping machine; 6 – vacuum sealing machine.

It should be borne in mind that for the same research object the construction of different models is possible. Moreover, the “correctness” of choosing the type of model and the most specific mathematical model can be proved very conditionally and with respect to many factors, as well as from the point of view of the researcher conducting the analysis, its system of preferences both regarding models and modeling methods, and regarding the choice of the degree of detail of real technological processes and selection of described (and accordingly ignored) properties, attributes and variables and parameters by which these selected properties (in the accepted degree of detail) can be described.

Ultimately, the “correctness” of the choice, construction, identification and use of models of this class depends on the formulation of the modeling problem and the suitability of the tools used (from informational data taken as a basis to models and methods for obtaining predictive estimates based on them and algorithms for evaluating correctness and adequacy).

Thus, the sequence of construction, configuration, implementation and use of the model can be represented as follows:

- Qualitative analysis of the studied TP VRM.
- Allocation of the economic system.
- Setting the goal of modeling.
- Selection of the initial data, their obtaining and preliminary analysis.
- Simplifying the system and replacing it with a mathematical model that can describe the dynamics of TP and give the required forecast values from the point of view of the goal.
- Numerical implementation of the economic and mathematical model.
- The reverse transition from the model to the system and from the system to the object of study and the analysis of the “degree of confidence” in the obtained predicted values.
- Raising the question of model verification, the possibility of refinements and / or additions to the studies performed in the previous stages 1., 2., 3., 4. and 5., 6., 7.
- Note that in all cases it is advisable to start the study with the simplest model. As such, we recommend a trend model with the smallest possible number of input exogenous variables and the only resultant output endogenous predicted variable (for example, “product quality”, “productivity”, “sales volume”) in the form of a simple linear regression. It can give some useful non-obvious prior to the calculations of the quantitative results. Based on the preliminary results obtained in this way, in the future it is possible to pose more complex and closer to reality economic problems for more detailed and accurate forecasting with the construction and application of more complex and more advanced economic and mathematical models of the TP VRM.

We emphasize that the need to conduct research in this sequence with the implementation of the selected stages is associated with the peculiarity of technological complexes VRM, which are uncertain probabilistic open systems in which there are not fully deterministic partially blurred random processes (in the probabilistic and statistical sense), which in principle can not be unambiguously defined.
To build trending models in the economy, it is necessary to have representative initial data in the form of time series.

Let's consider what issues and problems arise when modeling TP VRM. First, these are the requirements for the source data. They must meet certain conditions and requirements: not only the accuracy of the numerical data, but also their "appropriateness" of inclusion in a particular sample. In fact, this will depend not only on the type of trend model (linear or nonlinear), but also the closeness of the relationship between the studied indicators (and hence the very possibility of building a model and the degree of its adequacy). Second, there is the choice of model type. The dynamic model can be presented as a regression equation of various types (linear, polynomial, exponential, fractional-rational, logarithmic, logistic type with different number of coefficients) or as a differential equation (linear or nonlinear, first or second order, with the presence or absence of lagging arguments, ordinary or partial derivatives, with initial and (or) boundary conditions [6]). In this paper, we will limit ourselves to the consideration of the first type of models and start with the simplest of them — the linear regression model.

As an example, consider some conditional company VRM, for which we simulate the initial data on sales for the last 4 years. Thus, the task was reduced to the construction of a linear trend equation for sales volumes and short-term forecast based on it for one or two years on this indicator, the most important for the operation of the enterprise. The initial data for the variable Y — Sales volume is presented in column 3 of table 1 and in figure 2.

Table 1. Initial data and results of calculations on the model of linear regression.

| year | n | Y | T | Ŷ | \((y-\bar{Y})^2\) | \((\hat{y}-\bar{Y})^2\) | \(e=y-\hat{y}\) | \(e^2\) |
|------|---|---|---|---|----------------|----------------|----------------|----|
| 2014 | 1 | 1 | 1 | 0.300 | 7.563 | 11.903 | 0.700 | 0.490 |
| 2015 | 2 | 2 | 2 | 2.600 | 3.063 | 1.323 | -0.600 | 0.360 |
| 2016 | 3 | 3 | 3 | 4.900 | 0.063 | 1.323 | -0.900 | 0.810 |
| 2017 | 4 | 4 | 4 | 7.200 | 18.063 | 11.903 | 0.800 | 0.640 |
| the amount | - | 15 | 10 | - | 28.750 | 26.450 | 0 | 2.300 |
| average | - | 3.75 | 2.50 | - | - | - | - | - |
| stand. off | - | 2.681 | 1.118 | - | - | - | - | - |
| dispersion | - | 7.188 | 1.250 | - | - | - | - | - |
| forecast | 5 | 9.5 | 5 | for 2018 | | | | |
| 6 | 11.8 | 6 | for 2019 | | | | | |

Figure 2. Initial data for constructing the equation of the trend of product quality.
According to the results of table 1, the linear regression equation (1):
\[ Y = b_0 + b_1 \times T + e \]  

in which the regression coefficients are calculated by formulas (2) and (3):
\[ b_1 = r_{yt} \times S_Y / S_t, \]  
\[ b_0 = \bar{Y} - b_1 \times \bar{T} \]

Thus, we get \( b_1 = 2.3; \ b_0 = -2; \ \hat{Y} = b_0 + b_1 \times T = -2 + 2.3 \times T \)

The correlation coefficient between the output variable \( Y \) (sales volume, in notional monetary units from the level of 2014) and the input variable \( T \) (time, notional value from 1 to 4 with the possibility of extrapolation to 6) is equal to
\[ R_{YT} = 0.959 \]

The forecast values for 2018 and 2019 are equal respectively:
\[ Y_5 = -2 + 2.3 \times 5 = 9.5, \]
\[ Y_6 = -2 + 2.3 \times 6 = 11.8. \]

We determine the coefficient of determination by the formula
\[ R^2 = \frac{\sum (\hat{Y} - \bar{Y})^2}{\sum (y - \bar{Y})^2} \]

and check the ratio using the calculated value of the correlation coefficient (5) and the data of the intermediate calculations given in the table for the sums of the sixth and seventh columns 26,450 and 28,750:
\[ R^2 = (r_{xy})^2 = 0.9197 = (0.959)^2 = 26.450 / 28.750 = 0.920 \]

Given that the verification results in (6) coincided with an accuracy of 3 * 10^{-4}, we can conclude that the calculations were mathematically correct.

We calculate the estimate of the residual variance and the standard error of the regression equation, and on this basis we draw a conclusion about the quality of the trend equation:
\[ Y(T) = b_0 + b_1 \times T = -2 + 2.3 \times T \]

Have
\[ S_{oct}^2 = (\sum e^2)^{-(n-p-1)} = (\sum (y - \hat{y})^2)^{-(n-p-1)} = 1.150 \]

where \( p = k = 4 - 2 = 2 \) is the number of degrees of freedom.
\( Se = \sqrt{S_{oct}} = 1.072. \)

Most (92%) of the variation \( v \) of \( Y \) is explained by the constructed regression equation. The standard error (1.072) is not great, but is almost 30% of the average \( Y \): \( ve = se / Y_{average} = 1.072 / 3.75 \). Therefore, the constructed trend equation is of good (but not high) quality.

At the next stage of the study, the significance of the regression equation should be checked at the level of 5%.
\[ F = R^2 * (n - p - 1) / [(1 - R^2) * p] \]

The calculated value according to formula (9) gives an estimate of 23, while the critical value, in accordance with the standard Fisher distribution table \( F \) with a confidence probability of 95% and the numbers of degrees of freedom 1 and 2 is 18.51. Therefore, given that there is inequality
\[ 23 = F > F_{5\%}(4-1-1) = 18.51, \] it can be argued that the equation is significant at 5% (in other words, with a 95% probability, it can be argued that the coefficient of determination \( R^2 \) is not zero).
At the final stage, we will find out the possibilities of practical use of the results. First of all, we will give an economic interpretation of the regression coefficients in the forecast trend model (6). The b1 coefficient shows how many units Y will change on average if T changes by 1 unit. Thus, it is possible to predict that in case of invariance of external conditions of functioning of the enterprise annually values of Y on the average increase by 2.3 units. In a linear trend, the coefficient b1 is the average absolute gain of Y.

In conclusion, we make a forecast Y for 2018:

\[ \hat{Y} = b_0 + b_1 \times T = -2 + 2.3 \times 5 = 9.5 \]  

The average Y level in 2018 will be 9.5 units.

Using this technique, an assessment was made of the quality of the final product (mashed gooseberry fruit) and the determination of the performance of individual technological processes of gooseberry processing.

The methodology of modeling on the basis of time series is explained by the example of assessing the persistence of physical and chemical parameters of berries and gooseberry powder depending on the shelf life. This technique, along with the above approximation approach, uses the interpolation approach as an intermediate.

One of the approaches to solving this problem is to construct an interpolation polynomial, the values of which will be at the nodal points will coincide with the corresponding values from the source data table (table 1). But this approach does not take into account what character the described function has and whether it coincides with the character of the interpolating polynomial. Based on this, we have proposed a data processing technique that is as close as possible to solving the problem of modeling the persistence of the main indicators of gooseberries of different varieties during 15 months of storage, depending on the processing technology.

Table 2. Initial data for time series model calculation.

| Indicator                  | Contents after N months of storage | 0   | 1  | 2  | 3  | 6  | 9  | 12 | 15 |
|----------------------------|-----------------------------------|-----|----|----|----|----|----|----|----|
| Dry basis, %               |                                   | 95.00 | 94.98 | 94.95 | 94.91 | 94.84 | 94.79 | 94.56 | 94.41 |
| Total sugar, %             |                                   | 20.35 | 20.32 | 20.24 | 20.12 | 19.93 | 19.84 | 19.76 | 19.67 |
| Acidity (titrated), %      |                                   | 2.41  | 2.41  | 2.40  | 2.39  | 2.31  | 2.27  | 2.18  | 2.09  |
| Total pectin, %            |                                   | 3.34  | 3.31  | 3.29  | 3.26  | 3.18  | 3.04  | 2.97  | 2.81  |
| Ascorbic acid, mg per 100 g|                                   | 68.54 | 68.41 | 67.29 | 67.23 | 65.34 | 64.72 | 63.49 | 61.48 |
The basis of any technique is the formulation of the problem. We will set and solve the interpolation problem of fitting the formula with numeric data, and not approximation the task of approximation of the data to the formula, and a task description formula data so that, given the true nature of the underlying data (e.g., a concavity/convexity in the subject's projections and bands at the first stage, or more subtly, the nature of convexity/concavity, in a second step):

\[ y(x), y(x_i) = y_i \]  \hspace{1cm} (12)

\[ (y_2 - y_1)/(x_2-x_1) \quad (y_3 - y_2)/(x_3 - x_2) \quad \ldots \quad (y_8 - y_7)/(x_8 - x_7) \]  \hspace{1cm} (13)

\[ (y_3 - 2*y_2 +y_1)/(x_3 - x_1)^2 \quad \ldots \quad (y_8 - 2*y_7 +y_6)/(x_8 - x_6)^2 \]  \hspace{1cm} (14)

it had a certain type identified at the first stage and predetermined at the second stage.

In other words, the problem is not to satisfy the conditions (12) exactly (interpolation) or approximately (approximation), since these two problems give completely different formulas, but to bring these two opposite approaches closer together in order to maximize their agreement and reduce the disadvantages of each of them, leaving the advantages of both methods.

The generalized point of the result in the phase plane must be close to n available points (n=8) in terms of the available shelf life (i = 0, 1, 2, 3, 6, 9, 12, 15):

\[ F(x_i, y_i, n) = \sqrt{\sum \left(y(x_i) - y_i\right)^2} \rightarrow \min \]  \hspace{1cm} (15)

and which at the same time in the basis points (5) x_i, i = 1, 2, ..., 8 took values as close as possible to the identified values yi.

Note that the proposed approach is an alternative compromise (between the formulation of interpolation and approximation problems) version of the formulation of the data processing problem.

The requirement of proximity of table values Y and values (15) can be interpreted as follows. We will consider the set of values of the function Y from table 1 and the set of model values (15) as coordinates of two points of n-dimensional space. With this in mind, the function approximation problem can be reformulated as follows.

Verification of the adequacy of the obtained formulas was carried out by the following method, which is reduced to checking the mutual correlation of the values of x and y.

Pearson sample correlation coefficient r calculated from pairs of observations from a two-dimensional distribution: (x_1, y_1), ..., (x_n, y_n) were tested by the criterion of materiality of its difference from zero for each of the studied dependencies.

As the null hypothesis H0 we assume the absence of correlation. It is clear that a sufficiently large modulo the value of r will force to refute the null hypothesis.

The student's t-test is used to test the hypothesis

\[ t = r \sqrt{\frac{k}{(1 - D)}} \]  \hspace{1cm} (16)

where D = r^2 is the coefficient of determination, k = n-1 is the number of degrees of freedom.

To accept the alternative to H0 hypothesis H1 on the materiality of the relationship between the values of x and y – in other words, the significant difference between the calculated correlation coefficient from zero, compared calculated by the formula (15) the actual value of the Student's criterion in each case x_i with the theoretical boundary of the standard Student distribution table: \( \alpha = 0.05 \) is the level criterion (it corresponds to the value of confidence) \( p = 1 - \alpha = 1 - 0.05 = 0.95 \); and k = 6-2 =4 is the number of degrees of freedom.

From the standard Student distribution table for these parameters we find the lower critical boundary of the t-criterion:

\[ t_q(k, \alpha) = 2.132. \]

At the next step, we calculate the calculated value of the t-criterion by the formula (15) and compare it with the table one. If the condition is met
\[ t_{\text{history}} > t_{\text{crete}} = t_{\alpha}(k, \alpha) \]

we accept the alternative hypothesis; otherwise, we accept the null hypothesis.

For example, for the coefficient \( r = 0.75 \) obtained by analyzing 6 pairs of data

\[ t_{\text{history}} = 0.75 \times [4/(1 - 0.75 \times 0.75)]^{1/2} = 2.442 > 2.132 = t_{\alpha}(k, \alpha). \]

Hence, we accept the H1 – hypothesis-the correlation is reliable with 95% reliability.

References
[1] Fedorenko I Ya 2014 Designing technical devices and systems: principles, methods, procedures: textbook for universities
[2] Manasyan S K 1997 Modeling and system identification of the structure of non-entropic empirically complete objects Materials of the international scientific conference Krasnoyarsk state University pp 18-22
[3] Manasyan S K 2019 Matrix model of the state of a complex technical system of resource Bulletin of KrasSAU pp 135-40
[4] Manasyan S K 2003 Algorithm for assessing the quality of the technological process in many parameters in conditions of incomplete awareness Agricultural science at the turn of the century: Abstracts. All-Russian scientific Conf. Krasnoyarsk State Agrarian University
[5] Tipsina N N and Grechishnikova N A 2015 Use of gooseberry puree in sugar confectionery Vestnik KrasSAU pp 143-8
[6] Manasyan S K 2003 Construction of the dynamic multidimensional models of machines and units, their linearization Bulletin of KrasSAU pp 118-22