The non-parametric algorithm of omissions filling in stochastic data

A A Korneeva[1], **E A Chzhan**[1], **M A Denisov**[2] , **A V Medvedev**[1], **V V Kukartsev**[1,2] **and V S Tynchenko**[1,2]

[1]Siberian Federal University, Institute of Space and Information Technologies, Krasnoyarsk 660074, Russian Federation
[2]Siberian State University of Science and Technology, Krasnoyarskiy Rabochiy Ave., 31, Russian Federation, Krasnoyarsk, 660037

E-mail: ekach@list.ru

Abstract. The paper presents the results of an algorithm for data processing. In the initial data omissions may occur, due to different control discreteness of input and output variables. The paper proposes a non-parametric algorithm for filling gaps. The basic idea is to calculate the non-parametric estimate of the regression function from observations obtained from the object. This allows to use all available measurements. Numerous computational experiments have shown that the use of the proposed algorithm has improved the quality of the resulting model several times. The algorithm is influenced by such parameters as the total number of omissions in the sample of observations, measurement interference in communication channels, and the type of object. It should be noted that the developed algorithm is universal and does not depend on the type of equation of the object.

1. Introduction

One of the main fields of cybernetics is the modeling and identification of stochastic processes. A special role in the formulation of identification and control tasks is the level of a priori information, which largely depends on controls «input-output» variables of the process under study. In many practical tasks there is the situation when discreteness of control «input-output» variables are significantly different. This is due to the fact that measurements of some variables are carried out electrically, and others – by laboratory tests, physical and mechanical tests, etc.

The discreteness of the measurement of these variables is significantly different, which leads to omissions in «input-output» variables of the process observation matrix. The general scheme of the process under investigation is presented in Fig.1.

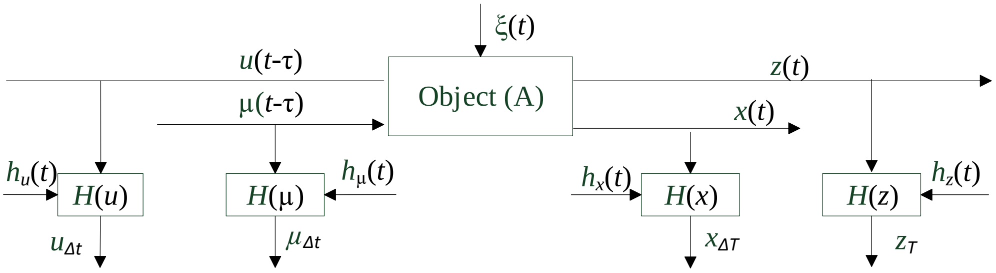

Figure 1. The general scheme of investigated object

The following notation is used in the figure: A is an unknown object operator; $x(t) \in \Omega(x) \subset R^1$, $z(t) \in \Omega(z) \subset R^1$ are output process variables; $u(t) = (u_1, u_2, ..., u_m) \in \Omega(u) \subset R^m$ is a control action; $\mu(t) = (\mu_1, \mu_2, ..., \mu_m) \in \Omega(\mu) \subset R^n$ is the vector input process variable; $\xi(t)$ is a vector random effect; (t) is continuous time; $H(u)$, $H(\mu)$, $H(x)$, $H(z)$ are communication channels corresponding to various variables and including controls; $h_u(t)$, $h_\mu(t)$, $h_x(t)$, $h_z(t)$ are random interference measurements of the corresponding process variables with zero expectation and limited variance.

The feature of these processes is that the input variables $u(t)$ and $\mu(t)$ are controlled at the time interval Δt, the output variable $x(t)$ is controlled at the interval ΔT, and the output variable $z(t)$ is controlled at interval T, moreover, the relation $\Delta t \le \Delta T \le T$ is satisfied. The difference in the discreteness of measurement of variables characterizing the state of the process under study is due to the available means of control. In particular, the measurement of the variable $x(t)$ can be carried out by electrical means and be carried out fairly quickly, and for the measurement of the variable $z(t)$, for example, chemical analysis is required, which requires much more time. Note, however, that the variable $z(t)$ is the most important for the process, since it characterizes the quality of the finished product.

Output process variables depend on all input variables:

$$x(t) = A\left(u\left(t-\tau\right), \mu\left(t-\tau\right), \xi\left(t\right), t\right) \tag{1}$$

where τ is the delay in the various channels of the process, which should not be confused with the delay in measuring certain process variables. In the described conditions, it is advisable to use the entire set of variables that influence the forecast $z(t)$. In this case, the model will be as follows:

$$z(t) = A\left(u\left(t-\tau\right), \mu\left(t-\tau\right), x\left(t\right), \xi\left(t\right), t\right) \tag{2}$$

The observations matrix of input-output variables of the described process can be presented, for example, in the form of the table 1.

Table 1. The observation matrix with omissions.

u_1	u_2	…	u_m	μ_1	μ_2	…	μ_n	x	z
u_{11}	u_{21}	…	u_{m1}	μ_{11}	μ_{21}	…	μ_{n1}	x_1	z_1
u_{12}	u_{22}	…	u_{m2}	μ_{12}	μ_{22}	…	μ_{n2}	–	–
u_{13}	u_{23}	…	u_{m3}	μ_{13}	μ_{23}	…	μ_{n3}	–	–
u_{14}	u_{24}	…	u_{m4}	μ_{14}	μ_{24}	…	μ_{n4}	x_4	–
…	…	…	…	…	…	…	…	…	…
u_{1s}	u_{2s}	…	u_{ms}	μ_{1s}	μ_{2s}		μ_{ns}	x_s	z_s

In table 1 it is assumed that the measurement resolution of the output variable $x(t)$ is three times the measurement resolution of the input variables $u(t)$ and $\mu(t)$, i.e. $\Delta T = 3\Delta t$, the output variable $z(t)$ is measured with a resolution of $T = 2\Delta T = 6\Delta t$, where s is the size of the original sample.

The task of filling omissions in the input-output process variables in order to improve the quality of the model is interesting. Of course, in solving the problem of identification, only completely filled rows of the matrix of observations can be used. It does not take into account the rows of the matrix of observations with omissions. However, there is a loss of information. This is unacceptable from a practical point of view. In addition, to solve the problem of identification, it is preferable to have a larger sample size. The search for new methods for filling omissions in the input-output variables of the observation matrix is an actuality task. To solve it we can use the methods of parametric and non-parametric identification**.**

2. Methods of filling the matrix of observations

As noted earlier, in practice there are often cases when the discreteness of measurement of the input and output variables of the process under study may not coincide. As a result, the observation matrix will consist of incomplete rows (table 1).

This article proposes to give estimates x_s and z_s of the output variables $x(t)$ and $z(t)$ in the blank rows of the observation matrix for known values of the input variables $u(t)$, which make use of a sample consisting of the results of the filled rows of the observation matrix (table 2).

Table 2. The observation matrix.

u_1	u_2	…	u_m	μ_1	μ_2	…	μ_n	x	z
u_{11}	u_{21}	…	u_{m1}	μ_{11}	μ_{21}	…	μ_{n1}	x_1	z_1
u_{12}	u_{22}	…	u_{m2}	μ_{12}	μ_{22}	…	μ_{n2}	x_{2s}	z_{2s}
u_{13}	u_{23}	…	u_{m3}	μ_{13}	μ_{23}	…	μ_{n3}	x_{3s}	z_{3s}
u_{14}	u_{24}	…	u_{m4}	μ_{14}	μ_{24}	…	μ_{n4}	x_4	z_{4s}
…	…	…	…	…	…	…	…	…	…
u_{1s}	u_{2s}	…	u_{ms}	μ_{1s}	μ_{2s}		μ_{ns}	x_s	z_s

Parametric [1; 2] and non-parametric methods could be used to get estimations $x(u,\mu)=M\{x\,|\,u,\mu\}$, $z(u,\mu)=M\{z\,|\,u,\mu\}$, $z(u,\mu,x_s)=M\{z\,|\,u,\mu,x_s\}$ [3; four]. The proposed technique turns out to be quite justified, since the identification problem in the latter case (table 2) is solved more accurately than in the case when rows with omissions are excluded from the observation matrix (table 1).

3. Nonparametric estimation of the regression function from observations

Let us give observations $\{u_i,\mu_i,x_i,z_i,i=\overline{1,s}\}$ of input and output variables, distributed with unknown probability densities $p(x,u)$, $p(u)>0\forall u\in\Omega(u)$. To restore the estimation $x(u,\mu)=M\{x\,|\,u,\mu\}$ non-parametric method could be used [4, 5]:

$$x_s(u,\mu)=\sum_{i=1}^{s}x_i\prod_{j=1}^{m}\Phi\left(\frac{u_j-u_{ji}}{c_s}\right)\prod_{k=1}^{n}\Phi\left(\frac{\mu_k-\mu_{ki}}{c_s}\right)\Bigg/\sum_{i=1}^{s}\prod_{j=1}^{m}\Phi\left(\frac{u_j-u_{ji}}{c_s}\right)\prod_{k=1}^{n}\Phi\left(\frac{\mu_k-\mu_{ki}}{c_s}\right),\quad (3)$$

here the kernel function $\Phi(\cdot)$ and the coefficient of blurring the core of c_s have certain properties of repeatability:

$$\begin{aligned} &c_s>0; && \Phi(c_s^{-1}(u-u_i))<\infty; && \lim_{s\to\infty}c_s^{-1}\Phi(c_s^{-1}(u-u_i))=\delta(u-u_i);\\ &\lim_{s\to\infty}c_s=0; && c_s^{-1}\int\limits_{\Omega(u)}\Phi(c_s^{-1}(u-u_i))dx=1; && \lim_{s\to\infty}sc_s^m=\infty. \end{aligned}\quad (4)$$

In this case, the kernel function has a triangular shape and is described by the following system of equations:

$$\Phi\left(\frac{u-u_i}{c_s}\right)=\begin{cases}1-\left|c_s^{-1}(u-u_i)\right|, & \left|c_s^{-1}(u-u_i)\right|\le 1;\\ 0, & \left|c_s^{-1}(u-u_i)\right|>1.\end{cases}\quad (5)$$

The blur parameter c_s is determined by solving the problem of minimizing the quadratic index of conformity of an object's output and the model's output based on the sliding exam method, when the *i*-th measurement pair is not taken into account when building the model:

$$R(c_s)=\sum_{k=1}^{s}\left(x_k-x_s\left(u_k,c_s\right)\right)^2=\min_{c_s},k\neq i. \quad (6)$$

If each component of the vector $u(t)$ corresponds to a component of the vector c_s, then in practice c_s can be taken as a scalar value. To do this, it is necessary bring the components of the vector $u(t)$ from a sample of observations to the same interval, using, for example, the centering and normalization operations.

To restore the dependence $z(u,\mu)=M\{z\,|\,u,\mu\}$ the non-parametric estimates is used.

$$z_s(u,\mu)=\sum_{i=1}^{s}z_i\prod_{j=1}^{m}\Phi\left(\frac{u_j-u_{ji}}{c_s}\right)\prod_{k=1}^{n}\Phi\left(\frac{\mu_k-\mu_{ki}}{c_s}\right)\Bigg/\sum_{i=1}^{s}\prod_{j=1}^{m}\Phi\left(\frac{u_j-u_{ji}}{c_s}\right)\prod_{k=1}^{n}\Phi\left(\frac{\mu_k-\mu_{ki}}{c_s}\right). \quad (7)$$

As noted earlier, the resulting estimates of the output variable $x(t)$ can be used to reconstruct the output variable $z(t)$. In this case, the following non-parametric estimates are used to restore $z(u,\mu,x_s)=M\{z\,|\,u,\mu,x_s\}$:

$$z_s(u,\mu,x_s)=\frac{\sum_{i=1}^{s}z_i\Phi\left(\frac{x_s-x_{si}}{c_s}\right)\prod_{j=1}^{m}\Phi\left(\frac{u_j-u_{ji}}{c_s}\right)\prod_{k=1}^{n}\Phi\left(\frac{\mu_k-\mu_{ki}}{c_s}\right)}{\sum_{i=1}^{s}\Phi\left(\frac{x_s-x_{si}}{c_s}\right)\prod_{j=1}^{m}\Phi\left(\frac{u_j-u_{ji}}{c_s}\right)\prod_{k=1}^{n}\Phi\left(\frac{\mu_k-\mu_{ki}}{c_s}\right)}. \quad (8)$$

Further it will be shown that using the estimate (8) gives a better result compared to the estimate (7).

4. Non-parametric algorithm of filling omissions in the matrix observations

Let us consider the non-parametric algorithm by way of example of of restoring the output variable $x(t)$.

At the first stage, the regression function x_s (3) is restored by the observations of $u(t)$, fully represented in the original measurement matrix, i.e. along the lines completely filled in the result of the experiment (in table 1 these are the first, fourth, seventh lines, etc.). Rows with missing $x(t)$ output values are not taken into account at this stage, due to which the sample size decreases. The optimal value of the blur coefficient c_s is determined in accordance with (6).

At the second stage, omissions are filled in the observation matrix using the x_s estimate obtained at the previous stage. Where observations $x(t)$ are omitted, the measured values $u=(u_1,u_2,...,u_m)$ are substituted into the estimate $x_s(u_1,u_2,...,u_m)$ and the corresponding estimate x_s is calculated, which makes up for the missing observation $x(t)$ (for example, the missing estimate x_2 in the observation matrix presented in Table 2 is filled with the value x_{s2}).

At the final stage, the nonparametric estimation of $x_s(u_1,u_2,...,u_m)$ is built over the entire existing (filled) observation matrix. Similarly, the omissions in the observation matrix for the output variable $z(t)$ are filled. As a result, non-parametric estimates (7) and (8) are built along the elements of the restored matrix.

5. Computational experiment

For the experiment was selected object presented in Fig. 1, for which the vector control action $u=(u_1,u_2,...,u_m)\in[0;3]$, and the effect of $\mu(t)$ is absent. The measurement resolution ΔT of the output variable $x(t)$ is three times the discreteness of measurement Δt of the input variable $u(t)$, i.e. $\Delta T=3\Delta t$. The output variable $z(t)$ is measured with an even greater discreteness $T=2\Delta T=6\Delta t$.

Output process variables are described by the following dependencies:

$$x = 0.5u_1^2 + \sin u_2 + 2\sqrt{u_3} \tag{9}$$

$$z = u_1 + \sin u_2 + u_3 + 0.5x \tag{10}$$

that are necessary for obtaining in a computer experiment the corresponding initial samples $\{u_i, x_i, z_i, i = \overline{1,s}\}$. In the following, the character of the dependences $x(u)$ and $z(u,x)$ is assumed to be unknown.

To build a model of the process $x(u)$ under study, the classical nonparametric regression estimate (3) is used. At the first stage, the estimate (3) is based on the initial observation matrix with omissions in the output variable $x(u)$ (table 1), at the second stage – on the matrix filled with the above method (table 2). Evaluation of the quality of models is carried out in accordance with the following formula:

$$W = \left((s-1)\sum_{i=1}^{s} \left(x_i - x_s(u_i)\right)^2 \Big/ s\sum_{i=1}^{s} \left(m_x - x_i\right)^2 \right)^{1/2}, \tag{9}$$

where W is the relative modeling error; m_x is estimation of the mathematical expectation of the object output: $m_x = s^{-1}\sum_{i=1}^{s} x_i$.

According to the results of the experiment, graphs were obtained of the dependence of the relative modeling error W on the size of the initial sample s with interference at the object output of 5% (Fig. 2). In the figure the solid line corresponds to the case of estimation by the observation matrix with omissions, and the dotted line corresponds to the case of estimation by the filled observation matrix. And since the process under consideration is stochastic, averaging was performed over the results of ten experiments.

Analysis of the graphs allows us to conclude that the application of the proposed technique leads to an increase in the accuracy of modeling by 5-10%, and on small sample sizes (s = 100-300) the accuracy increases by 15-20%.

To estimate the output variable *z*(*t*), nonparametric estimations (7) and (8) were used (fig. 3 and 4).

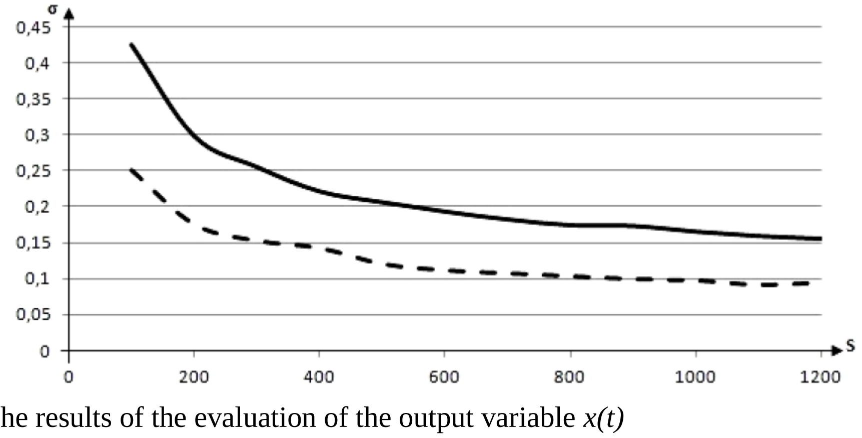

Figure 2. The results of the evaluation of the output variable *x(t)*

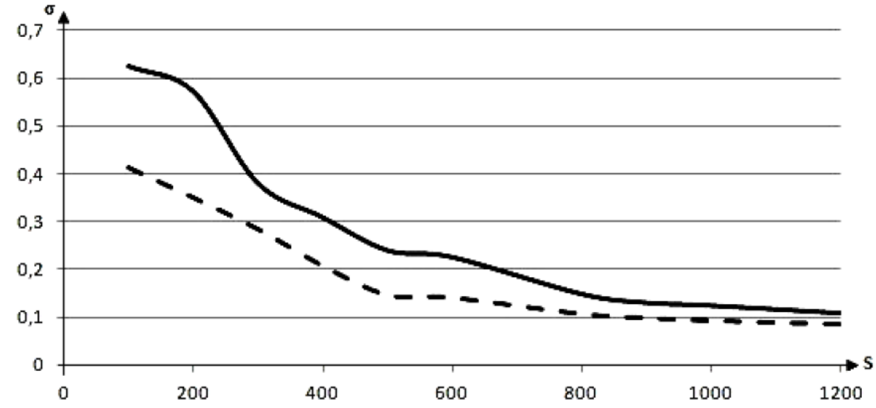

Figure 3. The results of the evaluation of the output variable $z(t)$

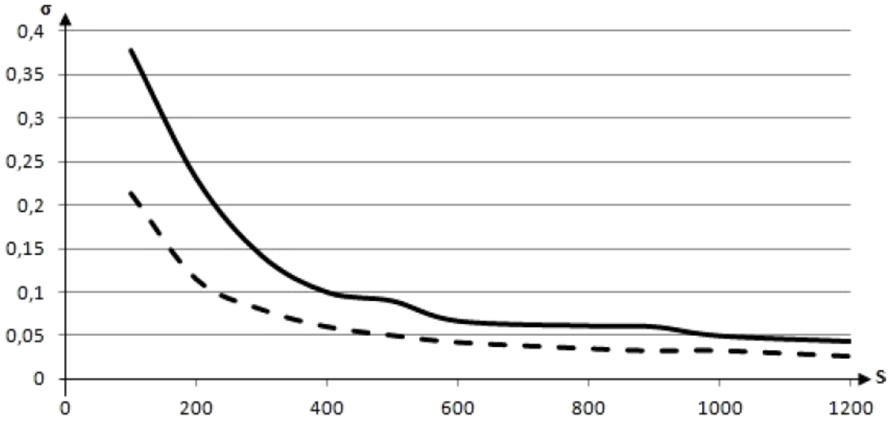

Figure 4. The results of the evaluation of the output variable $z(t)$ taking into account the values of the output variable $x(t)$

And in this case, the estimate for the filled observation matrix turned out to be more accurate. Thus, taking into account the output variable with a smaller discreteness $x(t)$ when estimating and restoring the output variable with a greater discreteness $z(t)$ makes it possible to significantly reduce the relative simulation error (fig. 3 and 4).

6. Conclusion

The authors proposed a technique for recovering omissions in the matrix of observations of input-output variables. The corresponding algorithms for filling these omissions are given. It is shown that the filling of the matrix leads to an increase in the quality of the model. In this regard, the problem of reconstruction of the matrix of observations with omissions was considered to solve the problem of identifying stochastic, static objects, including objects with delay.

It seems quite interesting to use when building a model for the main output $z(t)$ of another output variable $x(t)$, controlled with a smaller time resolution than $z(t)$. In this case, the accuracy of the model $z_s(u, x_s)$ will be much higher.

[1] Tsypkin Ja Z 1984 *Fundamentals of Identification Information Theory* (Moscow: Nauka)

[2] Eykhoff P. 1975 *Fundamentals of Identification Control Systems* (Moscow: Mir)

[3] Medvedev A V 1983 *Non-parametric adaptation systems* (Novosibirsk: Nauka, Siberian Branch)

[4] Nadaraya E A Non-parametric estimates of probability density and regression curve (Tbilisi: Publishing University of Tbilisy)

[5] Denisov M A Chzhan E A Korneeva A A Kukartsev V V 2018 About algorithm of robust nonparametric estimation of regression function of observation *IOP Conference Series: Materials Science and Engineering*, vol.450, issue 4, №042001