NON-PARAMETRIC APPROACH FOR PRELIMINARY PROCESSING OF EARTH REMOTE SENSING DATA

Maksim A Denisov[1] , *Ekaterina* A Chzhan[1] and *Anna* A Korneeva[1]

[1]SibFU Institute of Space and Information Technologies, Siberian Federal University, 660074 Krasnoyarsk, Russian Federation

Abstract. Researches described in the paper are aimed at studying the methods of data preprocessing from a sample of observations of a system characterized by input-output values of variables. We consider the data containing omissions and outliers. Algorithms for leveling outliers in a sample of observations, as well as algorithms for filling data gaps are presented. In addition, it is implemented a data repair algorithm that is able to recover lost values (outliers) after their exclusion. Our studies are useful in geographic information systems or in the analysis of information received from satellites during remote sensing of the earth.

1 Introduction

The sample of the input-output observations of the system (object) under study plays a valuable role in solving the problem of identification. The data in the tables with the values of observations of the object can be both quantitative and qualitative. Qualitative data can be, for example, geographical name, species composition of vegetation, soil characteristics, etc. Paper [1] describes how to work with such type of data obtained from satellite sensors. Author in paper [2] explores the potential of using geographical information systems (GIS) and paper [3] considers the applied side of working with quality data in such systems. In this article, the sample is presented in the form of numbers – quantitative data type.

Most often, in practical problems, the data may contain defects of various kinds: outliers or omissions.

Omissions arise, for example, during the process of shooting or transmitting data from satellites in cases when the transfer process was interrupted or simply due to some technical malfunction of the shooting device. All of this may interfere with the further processing of the image. Algorithms for filling such gaps using nuclear functions are developed under the condition of parametric identification [4], which differs from the approach proposed in this article.

The occurrence of outliers in a sample of observations may be due to an inaccurate mathematical model, malfunction or improper calibration of instruments, mistaken readings, gross recording, and calculation and execution faults. The impact of outliers can be leveled using the robust identification algorithms described in books [5, 6] and papers [7, 8] or using data censoring methods, one of which is described in [9]. In particular, in [10] it is described outliers detection in the field of hyperspectral imagery which is

connected with the task of locating pixels with spectral signatures that deviate significantly from the local background. As a result, such significantly deviate values in data affects the final accuracy of the object of study approximation.

The paper discusses two methods of data preprocessing, where the first one is a method of censoring a sample of observations to remove outliers, and the second one is restoring omissions using a non-parametric identification algorithm. Previously, the analysis of these algorithms cumulatively was not considered, which underlines the relevance of this work. In addition, after removing the outliers, it was decided to review and implement the data repair algorithm, which additionally increase the accuracy of object modeling.

2 The problem statement

An object of a discrete-continuous type is examined, whose scheme in its general form is shown in the figure below:

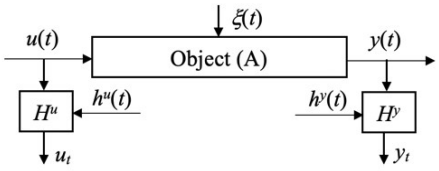

Fig. 1. Discrete-continuous process flowchart

The notation of the figure 1: A is unknown object functional; $y(t)\in\Omega(y)\subset R^1$ is an output process variable; $u(t)=(u_1(t), u_2(t),\ldots,u_m(t))\in\Omega(u)\subset R^m$ is a vector of input signal, where m is the number of input signals; $\xi(t)$ is the vector random noise; t is the continuous time; H^u, H^y are communication channels; $h^u(t)$, $h^y(t)$ are stochastic noise measurements; $\{u_{ji}, y_i, i=1,\ldots,s, j=1,\ldots m\}$ is a training sample, where s is sample size.

Most often, priori information about the object of study is not sufficient to build a model with an accuracy of the vector of parameters (in other words it is difficult or impossible to built a parametric model), therefore, a non-parametric approach is used. In this paper, the Nadaraya-Watson non-parametric estimation is used for conducting computational experiments. It is described by the following equation:

$$\hat{y}(u)=\sum_{i=1}^{s} y_i \prod_{j=1}^{m} \Phi\left(\frac{u_j-u_{ji}}{c_s}\right) \Bigg/ \sum_{i=1}^{s}\prod_{j=1}^{m}\Phi\left(\frac{u_j-u_{ji}}{c_s}\right), \quad (1)$$

where c_s – bandwidth parameter, and $\Phi[(u_j-u_{ji})/c_s]$ is a kernel function.

Due to the fact that we do not have a sample of observations with outliers and omissions, it is necessary to generate it by yourself as part of a computational experiment.

Omissions are generated as follows: $t=3\Delta t$, where t is the time step. Let us suppose, there is a sample with the volume s = 300. In order to obtain a sample with omissions according to the method described above, it is supposed to leave every third value in the data table. Thus, for a sample containing, for instance, s = 300 objects, omissions equal 200 values but other 100 values form remaining sample of observations.

Below, in Table 1, the initial sample is presented in general form (a “clean” sample without outliers or omissions); Table 2 contains a sample with omissions generated according to the rule described above.

Table 1. Initial sample

u_1	u_2	…	u_n	y
u_{11}	u_{12}	…	u_{1n}	y_1
u_{21}	u_{22}	…	u_{2n}	y_2
u_{31}	u_{32}	…	u_{3n}	y_3
u_{41}	u_{42}	…	u_{4n}	y_4
u_{51}	u_{52}	…	u_{5n}	y_5
…	…	…	…	…
u_{s1}	u_{s2}	…	u_{sn}	y_s

Table 2. Sample with omissions

	u_2	…	u_n	y
u_{11}	u_{12}	…	u_{1n}	y_1
u_{21}	u_{22}	…	u_{2n}	–
u_{31}	u_{32}	…	u_{3n}	–
u_{41}	u_{42}	…	u_{4n}	y_4
u_{51}	u_{52}	…	u_{5n}	–
…	…	…	…	…
u_{s1}	u_{s2}	…	u_{sn}	y_s

Outliers are generated for the remaining sample values (y_1, y_4, ..., y_s from Table 2). The value number that will be an outlier is a pseudo-random. After the number has been selected, the value in the sample is changed by the following formula:

$$y_i = y_i \cdot k \cdot c \qquad i \in [1; s], \tag{2}$$

where k = 0.99 is noise coefficient and c is normally distributed random value in the interval [-1;1].

3 Algorithm for filling omissions in data

Filling the missing values in the observation matrix is performed using nonparametric estimation (1). The application restoration algorithm of this type for objects of different mathematical description is considered in [11].

At the first stage, non-parametric identification of the object of study is performed using (1) from a sample of observations, in which omissions were removed in advance (cells with gaps from Table 2), that is a sample: $\{u_{ji}, y_i, i=1,\ldots,s_2; j=1,..,m\}$, where $s_2<s$ is sample size after removing all omissions. Next, the gaps in the observation matrix are filled using the estimate (1) obtained in the previous step. In those cases, where observations of the output of the object y are omitted, the known values of the input of the object u are substituted into the estimate. Thus, the values of the omissions are restored from the available input data about the object of study. The formula for calculating missing values is presented below:

$$\hat{y}_1(u) = \sum_{i=1}^{s-s_2} y_i \prod_{j=1}^{m} \Phi\left(\frac{u_j - u_{ji}}{c_s}\right) \Bigg/ \sum_{i=1}^{s-s_2} \prod_{j=1}^{m} \Phi\left(\frac{u_j - u_{ji}}{c_s}\right), \tag{3}$$

where u – the value of the input variables for which the omissions of the object's output is restored. As a result, the nonparametric estimate (1) is calculated anew for the elements of the reduced matrix.

4 Data censoring algorithm

As a method for eliminating outliers from a sample of observations, a data censoring algorithm will be used.

In the first stage using a sample of observations without gaps $\{u_{ji}, y_i, i=1,\ldots,s_2; j=1,..,m\}$ a non-parametric model of the form (1) is constructed. After that, the following condition is checked for all sample values:

$$\left|y_i - \hat{y}_i\right| > \mu \cdot \Delta, \tag{4}$$

where y_i is output values of the object, $\hat{y}_i$ is output values of the model (1), μ is customizable parameter, Δ is parameter, whose value is defined as:

$$\Delta = s^{-1}\sum_{i=1}^{s}\left|y_i - \hat{y}_i\right|. \tag{5}$$

If a sample point satisfies condition (4), then it is marked as an outlier and is removed from the sample of observations. As a result, the new sample is of the form: $\{u_{ji}, y_i, i=1,\ldots,s_3; j=1,..,m\}$, where $s_3 < s_2$ – volume of censored sample.

After applying the operations described above, the non-parametric non-outliers model is reconstructed from s_3-volumed sample of observations, the mathematical description of which is presented below:

$$\hat{y}_2(u) = \sum_{i=1}^{s_3} y_i \prod_{j=1}^{m} \Phi\left(\frac{u_j - u_{ji}}{c_s}\right) \Bigg/ \sum_{i=1}^{s_3}\prod_{j=1}^{m}\Phi\left(\frac{u_j - u_{ji}}{c_s}\right). \tag{6}$$

Note that the number of outliers found depends on the parameter μ set by the user. The result of the most accurate detection of an outlier value will depend on how this parameter has been configured.

5 Data repair

Outliers exclusion from a sample of observations increases the final accuracy of the object approximation, but some of the data after censoring is lost. It is known, as a sample size decreases, the accuracy of modeling declines. In this regard, it was decided to repair the data. The term "data repair" refers to the identification and subsequent replacement of rough measurements (outliers) with values of the robust model [12].

The description of repair algorithm is as follows: the values that were labeled using the data censoring algorithm as outliers will be replaced with values of the non-parametric model (1).

6 Computational experiment

Let us assume that the mathematical description of the object under investigation is as follows:

$$y = 0.25u_1^2 - 0.25u_2^2 + \xi, \tag{7}$$

where ξ is noise imposed on the output of an object that is generated as:

$$\xi = y \cdot c \cdot k, \tag{8}$$

where c is normally distributed random value in the interval [-1;1] and k is noise coefficient.

The model will be calculated using the non-parametric estimate which is represented by the following mathematical expression:

$$\hat{y}(u) = \sum_{i=1}^{s} y_i \prod_{j=1}^{2} \Phi\left(\frac{u_j - u_{ji}}{c_s}\right) \Bigg/ \sum_{i=1}^{s}\prod_{j=1}^{2}\Phi\left(\frac{u_j - u_{ji}}{c_s}\right). \tag{9}$$

The kernel $\Phi[(u_j - u_{ji})/c_s]$ used in the estimate (9) has the form of a parabolic bell-shaped:

$$\Phi(z) = \begin{cases} 0.75 \cdot (1 - z^2), & |z| \le 1, \\ 0, & |z| > 1, \end{cases} \tag{10}$$

where $z = (u_j - u_{ji})c_s^{-1}$.

The bandwidth c_s is configuring using cross-validation based on the minimum of the quadratic criterion:

$$F = s^{-1} \sum_{i=1}^{s} (y_i - \hat{y}_i)^2 \to \min_{c_s}. \tag{11}$$

It is accepted a relative approximation error in the paper, which evaluates the accuracy of object modeling:

$$W = \sqrt{(s-1) \sum_{i=1}^{s} (y_i - \hat{y}_i)^2 \Big/ s \sum_{i=1}^{s} (\hat{m}_y - y_i)^2}. \tag{12}$$

At the first stage of the computational experiment, we generate a sample consists of s = 300 observations $\{u_{ji}, y_i, i=1,\ldots,s; j=1,2\}$ using (7) with noise coefficient equals 5%. Further, create the gaps and outliers using the methods described above.

At the first stage of the computational experiment, we remove the observations containing omissions in the data and leave the outliers. Thus, we obtain a new sample of observations with a volume of s_2 = 100, after that we construct a non-parametric model (9). The accuracy of the modelling calculates using (12).

Next, filling the omissions using a mathematical expression (3). Combine them with the rest of the sample and get the original data volume s = 300. After that, we construct a non-parametric model (9), estimate the accuracy of which by (12).

In the next experiment, it is proposed to censor the data according to the method described earlier with a sample of volume s_2. The model (6) is constructed from a sample of volume s_3 (obtained after censoring), the accuracy is estimated using (12). The penultimate computational involves constructing a model (9) using a censored sample with filled gaps.

Finally, at the last stage of the computational experiment, for a censored sample with filled gaps, we apply the procedure of data repairing according to the algorithm described earlier. Thus, after conducting the entire series of experiments, the following results were obtained.

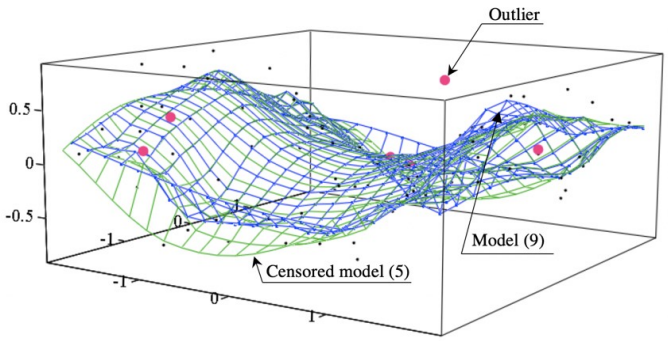

Fig. 2. Surface object (6) graph with omissions removed, model (9) and censored model (5)

Figure 2 shows the results of the computational experiment, when the omissions from the data were removed, after which, the data were censored of outliers. The graph shows that the censored model approximates the object more accurately than the model with outliers.

Table 3. The values of the relative error of approximation for all computational experiments

W_1	W_2	W_3	W_4	W_5
58.5%	35,2%	28,2%	16,6%	16,9%

In table 3, the following notation:

W_1 – relative approximation error for a sample with outliers and removed omissions;

W_2 – relative approximation error for a sample with outliers and filled omissions;

W_3 – relative approximation error for a censored sample with removed omissions;

W_4 – relative approximation error for a censored sample filled omissions;

W_5 – relative approximation error for repaired and censored sample with filled omissions.

Based on the results presented in the table above, it can be seen that when we fill omissions, as well as censor data, the accuracy of approximation increases by approximately one and a half to two times, which confirms the efficiency of the algorithms described in the work. In addition, the data repair algorithm after censoring also somewhat increases the accuracy of the final approximation of the object with the model.

7 Conclusions

The effectiveness of the algorithms described in the work confirmed an increase in the accuracy of approximation in the process of identifying the object of study. The computational experiment showed that sampling censoring gives more accurate simulation results than filling omissions in the data, however, if both of these methods are used together, then the final approximation results show the highest accuracy.

References

1. J. A. Engel-Cox, C. H. Holloman, B. W. Coutant, R. M. Hoff, Atmospheric Environment **38**, 2495-2509 (2004)

2. M. Pavlovskaya, Environment and Planning A **38**, 2003-2020 (2006)

3. J. Cidell, Area **42**, 514-523 (2010)

4. E. A. Lupyan *et al*., Sovr. Probl. DZZ Kosm. (in Russ) **8**, 190-198 (2011)

5. P. J. Huber, *Robust Statistics* (2011)

6. R. Maronna, *Robust Statistics: Theory and Methods* (2006)

7. W. Zhong, H. Lu, M. H. Yang, *Computer vision and pattern recognition (CVPR)* 1838-1845 (2012)

8. T. Ding, M. Zhang, D. He, *International Conference in Communications, Signal Processing, and Systems* 1229-1236 (2017)

9. C. E. Brodley, Journal of artificial intelligence research **11**, 131-167 (1999)

10. G. Camps-Valls, L. Bruzzone *Kernel methods for remote sensing data analysis* 170-172 (2009)

11. A. A. Korneeva, N. A. Sergeeva, Vestnik SibGAU (in Russ.) **45** 49-54 (2012)

12. E. S. Kirik, Vychislitelnye tehnologii (in Russ.) **6** 351-355 (2001)