Retrieving Missed Values of the Time Series Methods

Boris Bocharov, Maria Voevodina
Department of Applied Mathematics and Information Technologies
O. M. Beketov National University of Urban Economy in Kharkiv
Kharkiv, Ukraine
boris.bocharov@kname.edu.ua
voevodina@kname.edu.ua

Nataliia Braterska, Anastasiia Dashkovska
Students of CS 2016-1 group
O. M. Beketov National University of Urban Economy in Kharkiv
Kharkiv, Ukraine
nataliia.braterska@kname.edu.ua
anastasiia.dashkovska@kname.edu.ua

Методи Отримання Пропущених Значень Часових Рядів

Борис Бочаров, Мария Воєводіна
кафедра прикладної математики і інформаційних технологій
Харківський національний університет міського господарства імені О.М. Бекетова
Харків, Україна
boris.bocharov@kname.edu.ua
voevodina@kname.edu.ua

Наталія Братерська, Анастасія Дашковська
Студенти групи КН 2016-1
Харківський національний університет міського господарства імені О.М. Бекетова
Харків, Україна
nataliia.braterska@kname.edu.ua
anastasiia.dashkovska@kname.edu.ua

***Abstract*—The article researches the restoring missing values methods of knowledge testing results time series. To restore the values, the method of singular spectral analysis (SSA) is used. The effectiveness of the method is determined by simulation.**

***Анотація*—У статті досліджуються методи відновлення пропущених значень часового ряду результатів тестування знань. Для відновлення значень використовується метод сингулярного спектрального аналізу (ССА). Ефективність методу визначена за допомогою імітаційного моделювання.**

Keywords—distance learning, the effectiveness of educational impacts, educational site, singular spectral analysis

Ключові слова—дистанційне навчання, ефективність навчальних впливів, освітній сайт, сингулярний спектральний аналіз

I. Introduction

One of the great features of the SSA method for applying to the analysis of time series without spaces is that it can be used to study the structure of a series (the identification of trend, harmonic and noise components) without assumptions about the model series. However, the forecast of the components allocated by the method is possible only within a certain, rather broad, but nonetheless model of these components. It is assumed that the predicted component is close to the finite rank [1-5].

The idea of filling the passages within the SSA method is largely similar to the idea of the forecast and is to continue the allocation of the series and their structure to the places of missed observations. Accordingly, the theoretical results concerning the conditions and methods for the exact restoration of missed values in the components of the observed series with spaces are in the finite rank series.

Like the basic SSA method, the analysis method is used to study the time series with spaces, which gives accurate results in rather rigorous assumptions. This method also applies to real series with spaces, in which case we obtain approximate results [1-5, 9].

II. Restoring Missing Values of Time Series

In [6-9] we consider the application of the basic SSA method for time series of student test results, statistical estimations of the effectiveness of educational influences and the efficiency of using the library of the hybrid library. The time series of the test results were considered (35 attempts during the semester without passes). Knowledge of students who missed one or two attempts for any reason was checked through a survey using traditional methods. The results of students who missed more attempts were not considered.

Still, getting a time series that is the mainstream of data for such a study is not always possible. For example, it's hard to imagine a student group with 100% attendance at classes,

exams, tests, and the like. As a consequence, we usually get the missing elements to study.

The section is devoted to determining the effectiveness of filling in the algorithms in the time series of knowledge testing results using the SSA method. With the help of simulation, statistical estimations of the effectiveness of the algorithms under study were obtained.

The result of the basic algorithm of the "Track" method - SSA is the expansion of the observed time series into additive components. Let's consider the modification of the method for analyzing series with spaces. The overall structure of the algorithm is the same, but the steps of the stages will be slightly different [10].

Let the initial time series $F_N = (f_0,..,f_{N-1})$ consists of N elements, some of which are unknown. We describe the scheme of the algorithm for the case of recovery of the first component of the series $F_N^{(1)}$ on the basis of the sum of two:

$$F_N = F_N^{(1)} + F_N^2$$

The first stage: decomposition.

We fix the length of the window $L: 1 < L < N$. The attachment procedure translates the output time sequence into a sequence L - dimensional vectors of an embedding

$$\{X_i\}_{i=1}^K$$

where $K = N - L + 1$. Some attachment vectors may have spaces. From vector of embedding without spaces $X_i, i \in C$ create a matrix $\tilde{X}$, which in the absence of passes coincides with the trajectory matrix of a row F_N.

Let

$$\tilde{S} = \tilde{X} \cdot \tilde{X}^T$$

Let's denote $\lambda_1,...\lambda_L$ as the own numbers of the matrix $\tilde{S}$, taken in the not-for-going order $(\lambda_1 \geq ... \lambda_L \geq 0)$ and $U_1,...U_L$ - as an orthonormal system of eigenvectors of the matrix $\tilde{S}$, corresponding to their own numbers,

$$d = max\{i : \lambda_i > 0\}.$$

In [10], a formal fill-in option is proposed. It consists in replacing the scalar product of vectors with a similar formal procedure, applicable to vectors with spaces. We ask two vectors

$$A = (a_1,...,a_n)^T,$$

and

$$B = (b_1,...,b_n)^T,$$

and their set of missing components and, accordingly, to the same

$$|A \cup B| < n.$$

Enter the operation "*" so that

$$(A,B)^* = A^T * B = \frac{n}{n - |A \cup B|} \sum_{k:k \notin A \cup B} a_k b_k .$$

When multiplication of vectors without spaces, the result of the operation coincides with the scalar product, and for vectors with spaces, the scalar product will be numerically replaced.

As a matrix $\tilde{S}$ you can choose a matrix, which is calculated by the formula:

$$\tilde{S} = X * X^T,$$

where X – the trajectory matrix of a series F_N, which contains spaces.

The above method is generalized as follows: consider the value τ, $0 \leq \tau \leq L$, which we call the threshold of the number of missing components. Next, create a matrix $\tilde{X}_{(\tau)}$, Next, we create a matrix consisting of attachment vectors that contain no more τ missed components, then

$$\tilde{S} = \tilde{X}_{(\tau)} * \tilde{X}_{(\tau)}^T.$$

Note that the matrix $\tilde{X}_{(0)}$ coincides with the matrix $\tilde{X}$, consisting of vectors without spaces, and

$$\tilde{X}_{(L)} = X.$$

Second stage: recovery

First, we select the subspace and the projection of vector attachments without spaces. Choose a set of numbers

$$I_r = \{i_1,...,i_r\} \subset \{1,...,d\},$$

with the help of which the subspace is formed

$$M_r = Sp(U_{i_1},...,U_{i_r}),$$

which corresponds to the detachable component.

Choose your own vectors corresponding to $F_N^{(1)}$, is similar to how it is done at the stage of grouping in the basic SSA algorithm. One of the signs of the desired own vector is that its shape is similar to the form of a component of the series $F_N^{(1)}$. We design vectors of attachments without spaces on the selected subspace M_r:

$$\hat{X}_i = \sum_{k \in I_r} (X_i, U_k) U_k, i \in C$$

Next, we construct a projection of vectors of an embedding with spaces. For each such vector in the field with (the set is proper for each vector), the procedure consists of two parts:

- computing $\hat{X}_i$ for $I \setminus P$;
- computing $\hat{X}_i$ for P.

Since the neighboring attachment vectors contain general information, its use causes a large number of possible ways of solving the problem, including for vectors $I = P$.

To fill the vectors with spaces, you need to enter the breakdown of the entire set of spaces into groups so that at least the well-known values of a row, which are placed in succession, differentiate between groups of missed values. The computational procedure can be applied to each group of passes (for each set of vectors).

There are various ways to restore the positions of missing components in the set of vectors embedding group missed values, they can be divided into two groups:

- the simultaneous method of recovery (the use of the formula, which expresses the vector X_{P} by $X_{I \setminus P}$);
- a group of ways of successive filling (to the right, to the left, from two sides to the middle, on two sides with averaging). These methods are based on the fact that trajectory matrices have values on diagonals with indexes $(i, j), i + j = const$ are the same. So, it is possible to restore the gaps in one of the vectors of the embedding, and the gaps in the neighboring vectors fill the values obtained for the already restored vector.

The advantage of the successive recovery method, compared with the simultaneous, is that the weak conditions are applied, and the disadvantage is that the recovery error can accumulate.

The result of step C is the matrix

$$\hat{X} = [\hat{X}_1 \ldots \hat{X}_K],$$

which serves as an approximation of the trajectory matrix of a series $F_N^{(1)}$ for the correct choice of the set I_r.

Diagonal averaging. At the last step of the basic algorithm matrix $\hat{X}$ transformed into a new series $\tilde{F}_N^{(1)}$ (restored row) with diagonal averaging operation.

III. Description of the Research Process

Let's consider the simulation model of the process of retrieving missed values of the time series of knowledge testing results by the SSA method.

As a baseline, we use real test results [11-17] for 105 students who completed 35 attempts to pass tests.

We introduce the following notation:

N_s – number of students,

N_a – number of attempts for each student.

Value of the time series of tests of knowledge testing for a student in-one attempt:

$$\{ f_{ij} \}, i = \overline{0, N_s - 1}; j = \overline{0, N_a - 1}$$

Value of the restored time series of knowledge testing tests for the i-th student in the j-th attempt (the basic method of SSA is used):

$$\{ g_{ij} \}, i = \overline{0, N_s - 1}; j = \overline{0, N_a - 1}$$

Time series with missed values will be obtained by removing from the source $\{ f_{ij} \}$ n number of values, $n = \overline{1, \tau}$, where τ – the threshold of the number of missing components. In our case $\tau = 15$.

We get a family of time series

$$\{ f_{ij}^{(n,k)} \}, i = \overline{0, N_s - 1}; j = \overline{0, N_a - n - 1}; n = \overline{1, \tau}; k = \overline{1, K_n}$$

where K_n – the number of options for removing values from the original time series.

It is clear that, $K_n = C_{N_a}^n$, and the total number of time series (for all students) equals $N_s K_n$. In our case, when $n > 4$ the number of variants exceeds 10^6, because a million variants are quite enough for practical purposes.

A modern computer can calculate this amount in a reasonable time, so we will not generate more 10^4 time series for each student. In this case, specific time series are generated randomly using the Monte-Carlo method.

So,

$$K_n = \begin{cases} C_{N_a}^n, n \le 4, \\ 10^4, n > 4. \end{cases}$$

Apply modification of the SSA method to restore the time series with the missing values for each row in the family $\{ f_{ij}^{(n,k)} \}$.

We receive a family of restored time series

$$\{ g_{ij}^{(n,k)} \}, i = \overline{0, N_s - 1}; j = \overline{0, N_a - 1}; n = \overline{1, \tau}; k = \overline{1, K_n}.$$

We study $\{ g_{ij} \}$ and $\{ g_{ij}^{(n,k)} \}$ series.

We will assume that for each n average (for attempts) the relative value of the module for the difference between the values of the restored time series with spaces and without spaces is a random variable with the number of implementations $N_s K_n$:

$$y_i^{(n,k)} = \frac{1}{N_a} \sum_{j=0}^{N_a - 1} \frac{\left| g_{ij}^{(n,k)} - g_{ij} \right|}{g_{ij}}, \quad i = \overline{0, N_s - 1}; \quad n = \overline{1, \tau}; \quad k = \overline{1, K_n}$$

We call this random variable the "error" of the algorithm for restoring the values of the time series with n missing values.

In [18, 19] it is shown that such a value can be used as a measure of proximity of time series.

The statistical analysis of this random variable allows us to accept the hypothesis of a normal distribution law for all n with a significance level of at least 0.95.

The mathematical expectation of the "error" of the recovery algorithm is determined by the formula:

$$\bar{y}^{(n)} = \frac{1}{N_s K_n} \sum_{i=1}^{N_s} \sum_{k=1}^{K_n} y_i^{(n,k)},$$

and the standard deviation equals

$$\delta^{(n)} = \sqrt{\frac{1}{N_s K_n - 1} \sum_{i=1}^{N_s} \sum_{k=1}^{K_n} (y_i^{(n,k)} - \bar{y}^{(n)})^2}$$

Table I presents the statistical results of simulation of the time series recovery algorithm with spaces. The confidence intervals of the "error" algorithm with a confidence level of 90% were determined for specific values of the number of missed values (from 1 to 15).

Statistical analysis proves that if the number of missed values does not exceed seven, then the "error" of the recovery algorithm not more than 20%. This means that the difference between the estimates does not exceed one point on a five-point scale. Many years of testing knowledge at the Kharkiv National University of Urban Economy O. M. Beketov shows that tests of big errors do not adequately assess student knowledge of the students adequately. Therefore, if the number of missed values is greater than seven, the SSA algorithm cannot be used to restore the time series of student test results.

TABLE I. STATISTICAL ANALYSIS RESULTS OF THE ALGORITHM "ERRORS" FOR THE TIME SERIES VALUES RECOVERY

Number of missed values	Comparative values ($\bar{y}^{(n)}$)	Standard deviation ($\delta^{(n)}$)	Lower limit of confidence interval	Upper limit of the confidence interval
1	0,012	0,006	0,002	0,022
2	0,018	0,007	0,006	0,030
3	0,032	0,010	0,016	0,048
4	0,039	0,011	0,021	0,057
5	0,068	0,013	0,047	0,089
6	0,087	0,015	0,062	0,112
7	0,171	0,018	0,141	0,201
8	0,272	0,021	0,237	0,307
9	0,319	0,022	0,283	0,355
10	0,346	0,025	0,305	0,387
11	0,382	0,028	0,336	0,428
12	0,409	0,034	0,353	0,465
13	0,453	0,041	0,386	0,520
14	0,481	0,052	0,395	0,567
15	0,516	0,059	0,419	0,613

IV. RESULTS AND CONCLUSIONS

The use of simulation models made it possible to obtain statistical estimations of the efficiency of restoring values of the time series of knowledge test results. The algorithms restore the time series with sufficient accuracy (error not more than 20%) if the number of missed values is not more than seven.

Applying the method of singular spectral analysis for time series with spaces will allow determining the effectiveness of educational influences in situations where the basic method cannot be applied.

Prospective are further statistical studies of the student testing results, the construction of statistical and simulation models of the learning process, the development of expert systems, knowledge bases and decision support systems that allow you to choose rational learning strategies for each student.

REFERENCES

[1] Голяндина Н. Э. Метод Гусеница – SSA: Анализ временных рядов / Н. Э. Голяндина. – СПб: 2004. – 76 с.

[2] Голяндина Н.Э. Варианты метода «Гусеница»-SSA для анализа многомерных временных рядов / Н.Э. Голяндина, В.В. Некруткин, Д.В. Степанов // Труды II Международной конференции «Идентификация систем и задачи управления» SICPRO'03, Москва, 29-31 января 2003. М.: Институт проблем управления им. В.А.Трапезникова РАН. 2003. – С. 2139-2168.

[3] Golyandina N. Analysis of Time Series Structure: SSA and Related Techniques / N. Golyandina, V. Nekrutkin, A. Zhigljavsky. - London: Chapman & Hall/CRC, 2001. - 305 p.

[4] Golyandina N. Singular Spectrum Analysis with R. / N. Golyandina, A. Korobeynikov, A. Zhigljavsky. - Springer Verlag, 2018. - 246 p.

[5] de Carvalho M. Real-time nowcasting the US output gap: Singular spectrum analysis at work / Miguel de Carvalho, Antonio Rua // International Journal of Forecasting. - #33.- P185-198.

[6] Воеводина М. Ю. К вопросу об определении эффективности обучающих воздействий / М.Ю. Воеводина // Системи обробки інформації.– Харків: 2007.– № 2(60).– С.153-158.

[7] Воеводина М. Ю. Исследование эффективности обучающих воздействий / М.Ю. Воеводина // Тезисы доклада на XXIV научно-технической конференции ХНАГХ, 2008. – С. 140-141.

[8] Бочаров Б.П. Інформаційні технології в освіті : монографія / Б.П. Бочаров, М.Ю. Воєводіна; Харків. нац. ун-т міськ. госп-ва ім. О. М. Бекетова. – Харків: ХНУМГ ім. О. М. Бекетова, 2015. – 197 с.

[9] Bocharov B. Decision support system for the management of distributed automated teaching system / B. Bocharov, M. Voevodina, Y. Levikov // Материалы научно-технической конференции «Информационные системы и технологии». – Х.: 2012. – С.14.

[10] Голяндина Н.Э. Метод "Гусеница"-SSA для анализа временных рядов с пропусками / Н.Э. Голяндина, Е.А. Осипов // Математические модели. Теория и приложения.– СПб: изд-во НИИХ, 2005. – С.24-28.

[11] Бондаренко М.Ф. Теория интеллекта / М.Ф. Бондаренко, Ю.П. Шабанов-Кушнаренко. – Харьков: Изд-во СМИТ, – 2007. – 576 с.

[12] Бабаєв В. М. Організаційна культура керівника : навч. посіб. / В. М. Бабаєв, Н. В. Шаронова. – Х. : НТУ "ХПІ", 2005. – 259 с.

[13] ISO 29990:2010. Learning services for non-formal education and training. Basic requirements for service providers [http://www.iso.org/iso/ru/catalogue_detail?csnumber=53392].

[14] ISO/IWA 2 «Quality management systems. Guidelines for the application of ISO 9001:2000 in education» [http://www.iso.org/iso/ru/catalogue_detail? csnumber=42356]

[15] Шаронова Н.В. Модель извлечения глубинных знаний для систем организационного управления / Н.В. Шаронова, В.А. Тарловский, Н.Ф. Хайрова // Информационные технологии: вестник ХНТУ. – №2(38). – 2010. – С.97-102.

[16] Методы и средства принятия решений в социально-экономических и технических системах: учебное пособие / Э.Г. Петров, М.В. Новожилова, И.В Гребенник, Н.А. Соколова.; Под общ. ред. Э.Г. Петрова. - Херсон: ОЛДІ-плюс, 2003. – 380 с.

[17] Бочаров Б. П. Опыт использования дистанционных технологий при подготовке специалистов городского хозяйства / Б.П. Бочаров, М.Ю. Воеводина // Коммунальное хозяйство городов.– К: «Техніка», 2008, № 81.– С.409-413

[18] 84. Бурнаев Е.В. Меры близости на основе вейвлет коэффициентов для сравнения статистических и расчетных временных рядов / Е.В. Бурнаев, Н.Н. Оленев // Труды XLVIII научной конференции МФТИ. Часть VII. – М: МФТИ, 2005.– С. 108-110.

[19] 85. Бурнаев Е. В. Модель функционального состояния участников лабораторных рынков / Е. В. Бурнаев, И. С. Меньшиков // Известия РАН. Теория и системы управления. – 2009. – №. 6. – С. 187-204.