
Abstract. In the article presents results of research of primary statistical characteristics of time of processing and data transmission are given in systems of information economic area. On the basis of the received results the analysis of methods of recovery of the passed data is made. In work methods of data processing with admissions are analyzed and the most rational are chosen. Change of the law of distribution of selection depending on a way of elimination of the passed data is investigated. Forecasting of time of availability of the server of information economic area is executed. The goal of this paper is primary statistical characteristics research of data processing and transmission time in IES. Particularly, the server availability time (SAT) was studied within this work. This time consists of Client-to-Server connection time, Server data processing time, and Server answer transfer time. The values of service time were taken for research, that were received while studying of cloud computing performance. It was established that if data loss is not over 6% distribution law of initial model selection saves. If data loss is between 9% and 15% methods primary statistical characteristics of studied selection were determined without consideration of data missing. Missed data was filled with data distributed by the same law as for initial selection with admissions.

Keywords: time of availability of the server, statistical data, the distribution law, indicated of Hurst, method sliding average, admissions of data.

Formulas: 14; fig.: 3, tabl.: 3, bibl.: 6

JEL Classification: G 21, F 29, L 41.
СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ВОССТАНОВЛЕНИЯ ПРОПУЩЕННЫХ ДАННЫХ ВРЕМЕНИ ДОСТУПНОСТИ В ИНФОРМАЦИОННО-ЭКОНОМИЧЕСКИХ ПРОСТРАНСТВАХ

Анотация. В статье приведены результаты исследования первичных статистических характеристик времени обработки и передачи данных в информационно-экономических пространствах. На основании полученных результатов выполнен анализ методов восстановления пропущенных данных. В работе проанализированы приемы обработки данных с пропусками и выбраны наиболее рациональные. Исследовано изменение закона распределения выборки в зависимости от способа устранения пропущенных данных. Выполнено прогнозирование времени доступности сервера в информационно-экономических пространствах. Цель работы состояла в исследование первичных статистических характеристик времени обработки и передачи данных в информационно-экономических пространствах. В частности, в рамках данной работы изучалось время доступности сервера.

Ключевые слова: Информационно-экономические пространства, время доступности сервера, статистические данные, закон распределения, показатель Херста, метод скользящего среднего, пропуски данных.

Формул: 14; рис.: 3, табл.: 3, бібл.: 6
The goal of this paper is primary statistical characteristics research of data processing and transmission time in IES. Particularly, the server availability time (SAT) was studied within this work. This time consists of Client-to-Server connection time, Server data processing time, and Server answer transfer time. The values of service time were taken for research, that were received while studying of cloud computing performance [3]. The research of IES performance and reliability is actual because of user needs to have continuous access to them without data loss.

**Research results.** Statistical data used in paper was received as follows [3]. The Client and Server applications were developed. The Server application was installed on the remote Cloud Computing Azure computer. The Client application was gathering parameters of connection time and remote server data processing time every minute during 24 hours. In common, 1440 values were received, the initial selection based on these values was established, with 0% of lost data. Due to the fact that data transmission in Client-Server system is inevitably followed by data loss for different reasons, the research of data missing influence on primary statistical processing results was made.

The following approach was used for the analysis of missed data influence on the results of statistical characteristics determination. 0, 1, 3, 6, 9, 12, 15% of data was removed from initial selection. Data for removing was selected randomly. The conclusions about lost data influence on the quantity of received statistical results of selections properties were made based on comparisons between received statistical characteristics for each selection.

According to recommendations from paper [5] the following methods of data processing with losing were used:

1. The "Gluing" method - primary statistical characteristics of studied selection were determined without consideration of data missing. With such approach the selection size decreases, and therefore freedom degrees number also decreases when checking of the formulated statistical hypotheses.

2. The "Zero" method - primary statistical characteristics of studied selection were determined with consideration of filling admissions with zeros. In such case freedom degrees number does not decrease, but estimations shift of received characteristics appears comparing to initial selection.

3. The "Average" method - primary statistical characteristics of studied selection were determined with consideration of filling admissions with average values. In such case it is also possible to get estimations shift of received characteristics comparing to initial selection.

4. The "Random" method primary statistical characteristics of studied selection were determined with consideration of filling admissions with quasi-normal numbers distributed by normal law when average value and mean square deviation coincides with with similar characteristics of an initial data set with admissions.

5. In addition to these method, the following "Distributions" approach was used. Missed data was filled with data distributed by the same law as for initial selection with admissions.

The Statgraphics V.15 system was used for determination of distribution law type. The results of this procedure are shown in table 1.

<table>
<thead>
<tr>
<th>Missed data filling approach</th>
<th>0</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>9</th>
<th>12</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gluing</td>
<td>T1</td>
<td>T1</td>
<td>T1</td>
<td>T1</td>
<td>T1</td>
<td>T1</td>
<td>T1</td>
</tr>
<tr>
<td>Average</td>
<td>T1</td>
<td>T1</td>
<td>T1</td>
<td>T1</td>
<td>T2</td>
<td>T1</td>
<td>T3</td>
</tr>
<tr>
<td>Random</td>
<td>T1</td>
<td>T1</td>
<td>T1</td>
<td>T1</td>
<td>T1</td>
<td>T2</td>
<td>T2</td>
</tr>
<tr>
<td>Distributions</td>
<td>T1</td>
<td>T1</td>
<td>T1</td>
<td>T1</td>
<td>T1</td>
<td>T1</td>
<td>T1</td>
</tr>
</tbody>
</table>

* T1 – distribution of maximum value, T2 – loglogistic distribution, T3 – Laplace's distribution.
Received distribution laws look as follows.

T1 - function of the density distribution of maximum value

\[ f(x) = \frac{1}{\beta} \exp\left\{ - \frac{x - \alpha}{\beta} - \exp\left( \frac{x - \alpha}{\beta} \right) \right\} \]  \hspace{1cm} (1)

Distribution parameters \( \alpha \) and \( \beta \) are related to mathematical expectation \( m \) and to dispersion with \( s^2 \) equations:

\[ m = \alpha + \beta \Gamma^{-2}, \quad s^2 = \frac{\beta^2 \pi^2}{6} \]  \hspace{1cm} (2)

T2 – loglogistic distribution with density:

\[ f(x) = \frac{1}{\alpha x} \exp(z), \quad \text{where:} \quad z = \frac{\ln(x) - \mu}{\sigma} \]  \hspace{1cm} (3)

In such case parameters of position \( \mu \) and scale \( \sigma \) are related to mathematical expectation \( m \) and to dispersion with \( s^2 \) conditions:

\[ \mu = \exp(\mu) \Gamma(1 + \sigma) \Gamma(1 - \sigma) \]  \hspace{1cm} (4)

\[ s^2 = \exp(2\mu) \left[ \Gamma(1 + 2\sigma) \Gamma(1 - 2\sigma) - \Gamma^2(1 + \sigma) \Gamma^2(1 - \sigma) \right] \]  \hspace{1cm} (5)

where \( \Gamma(.) \) – Euler's gamma-function.

T3 – Laplace's distribution with density:

\[ f(x) = \frac{\lambda}{2} e^{-\lambda |x-\mu|} \]  \hspace{1cm} (6)

Parameter of this distribution \( \mu \) is equal to mathematical expectation, dispersion \( \sigma^2 \) is related to parameter \( \lambda \) with equation:

\[ \sigma^2 = \frac{2}{\lambda^2} \]  \hspace{1cm} (7)

The conclusion from table 1 is if data loss is not over 6% all studied methods of filling missed data save distribution law of initial model selection. If data loss is between 9% and 15% only "Gluing" and "Distributions" methods save distribution law of initial model selection. When restoring missing data it is important not only to save distribution law of initial selection, but also to achieve the fact that initial model selection (selection without missing) would not be significantly (in statistical sense) different from selections with artificially filled missed data.

Non-parametric criteria of selection coincidence hypothesis checking were used to check the influence of missed data filling approach on received distributions parameters. The checking results and the list of used criteria are presented in table 2. Model selection was used as standard sample in all studied cases. It was established that symbol (*) means absence of difference between compared selections by all criteria. The list of used criteria and their symbols are provided in the note to table 2. All computations were performed using AtteStat program.
Table 2

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero</td>
<td>*</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gluing</td>
<td>* * * * *</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>* * * * Z, A</td>
<td>VW, Z, A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>* * * * * * A, S</td>
<td>*S, K</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distributions</td>
<td>* * * * * * *</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


The presence of symbol in table cell points to criterion that discarded hypothesis about selections coincidence. So, only "Gluing" and "Distributions" methods of restoring missing data are recommended for practical usage.

Statistical row properties were studied for the reasonable choice of forecasting method. At the first stage fractal row properties were defined, specifically Hurst index H.

In paper [4] following procedure of Hurst index computation is proposed. Relation between Hurst index H and data row statistical characteristics is defined with the formula:

\[
R / S = \left( \frac{\pi}{2} N \right)^H ,
\]

where:

- \( S \) – mean square deviation of observations timing row,
- \( N \) – number of observations.

The Hurst index H is defined with the formula:

\[
H = \frac{\lg(R / S)}{\lg(\pi N / 2)} .
\]

Formula (10) is used to compute mean square deviation of observations row S:

\[
S = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{X})^2} .
\]

where: \( \bar{X} \) – arithmetic mean of studied observations timing row for \( N \) timing periods.

The range of accumulated deviation is the most important element of Hurst index computation formula. In general, it is evaluated as follows:

\[
R = \max_{1 \leq i \leq N} Z_u - \min_{1 \leq i \leq N} Z_u
\]

where: \( Z_u \) is accumulated deviation of row elements from mean value:

\[
Z_u = \sum_{i=1}^{N} (x_i - \bar{X}) .
\]

In paper [4] it is recommended to change left part of formula (8) using following correction if the observations number N is less than 250:

\[
R / S_f = R / S \times 0.998752 + 1.051037
\]
Table 3

<table>
<thead>
<tr>
<th></th>
<th>All selection</th>
<th>Last hundred</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arithmetic mean X</td>
<td>2007,678</td>
<td>1958,158</td>
</tr>
<tr>
<td>Standard deviation S</td>
<td>255,464</td>
<td>225,5333</td>
</tr>
<tr>
<td>Range R</td>
<td>57494,667</td>
<td>3698,297</td>
</tr>
<tr>
<td>Normalized Range R/S</td>
<td>225,060</td>
<td>16,39801</td>
</tr>
<tr>
<td>Hurst index Ht</td>
<td>0.709</td>
<td>0.570525</td>
</tr>
</tbody>
</table>

As it is shown in paper [4], in case of Hurst index (Ht) is between 0.326 and 0.674 the model of changing row values is Wiener process. The physical analogue of this process is Brownian motion around average value of observations row.

Evaluation results of index H that is determined on all data array ("all selection") and on last hundred observations ("last hundred") are given in table 3. The obtained data was served as explanation for the choice of sliding average as forecasting method.

The quality of forecasting results was evaluated with retrospective forecast method for six last values using formula:

\[ e = \frac{1}{m} \sum_{i=m-k}^{m} \left| \hat{x}_i - x_i \right| \]

(14)

where: \( e \) – value of average relative forecast error,

\( m \) – capacity of data array for which average relative forecast error was determined

\( x_i \) – actual value,

\( \hat{x}_i \) – calculated value.

Table 4

<table>
<thead>
<tr>
<th>Selection type</th>
<th>Sliding average order</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td>«All selection»</td>
<td>0.072</td>
</tr>
<tr>
<td>«Last hundred»</td>
<td>0.028</td>
</tr>
</tbody>
</table>

The chart of studied process is shown on fig. 1. The x-axis represents number of experiments (server calls), the y-axis represents server availability time in milliseconds.

Fig. 1. Server availability time
Autocorrelation function of SAT value changing process is shown on fig. 2. According to this figure it is obvious that statistical dependence between sequential SAT observations is very weak that indirectly confirms received value of $H$.

Fig. 2. Autocorrelation function of random SAT process

The chart of autocorrelation function of SAT process first differences was received to check its stationarity (fig. 3). According to this chart SAT process can be considered stationary.

Fig. 3. Autocorrelation function of first differences of SAT random process

**Conclusion.** It was established that if data loss is not over 6% distribution law of initial model selection saves. If data loss is between 9% and 15% only "Gluing" and "Distributions" methods save distribution law of initial model selection.

Fractal properties of SAT row were studied using Hurst index in this paper. The method of SAT row values forecasting was chosen on this basis. The quality of forecasting results was evaluated using retrospective forecast method for six last values. Results showed that sliding average method with step equal to three should be used for SAT forecasting.

Література


201

Стаття надійшла до редакції 27.01.2015 © Горбенко А. В., Рубан В. І.

References

Received 27.01.2015 © Gorbenko A. V., Ruban V. I.