УДК 004.052

**А. В. ГОРБЕНКО[1], В. Ю. ДУБНИЦКИЙ[2], В. И. РУБАН[1]**

[1] *Национальный аэрокосмический университет им. Н. Е. Жуковского "Харьковский авиационный институт", Украина*
[2] *Харьковский институт банковского дела Университета банковского дела Национального банка Украины, Украина*

# СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ВОССТАНОВЛЕНИЯ ПРОПУЩЕННЫХ ДАННЫХ ВРЕМЕНИ ИЗМЕРЕНИЯ ДОСТУПНОСТИ СЕРВЕРА В СИСТЕМАХ «ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ»

*В статье приведены результаты исследования первичных статистических характеристик времени обработки и передачи данных в системах «Облачных вычислений». На основании полученных результатов выполнен анализ методов восстановления пропущенных данных. В работе проанализированы приемы обработки данных с пропускам и выбраны наиболее рациональные. Исследовано изменение закона распределения выборки в зависимости от способа устранения пропущенных данных.Показано, что при определении параметров законов распределения выборки с пропущенными данными наилучшим можно считать метод при котором на место пропущенных данных вставляют квазислучайные числа, распределённые по тому же закону, что и исходная выборка.*

***Ключевые слова**: «Облачные вычисления», время доступности сервера, статистические данные, закон распределения, показатель Херста, метод скользящего среднего, пропуски данных*

## Введение

В настоящее время IT-решения, связанные с использованием облачных вычислений (Cloud Computing), приобретают все большую популярность[1]. В данное время Cloud Computing - одно из самых прогрессивных направлений развития информационных технологий. К достоинствам технологии облачных вычислений относят доступность, отказоустойчивость, экономичность, простоту, гибкость и масштабируемость.

**Цель работы** исследование первичных статистических характеристик времени обработки и передачи данных в системах «Облачных вычислений». В частности, в рамках данной работы изучалось время доступности сервера (ВДС). Это время состоит из времени соединения «Клиента» с «Сервером», времени обработки «Сервером» данных и времени передачи ответа «Сервером». Для исследования были взяты значения времени обслуживания, полученные при исследовании производительности Cloud Computing [3].Исследование производительности и надежности Cloud - систем является актуальным потому, что пользователь должен иметь постоянный доступ к ним без потери информации.

**Анализ литературы.** Известны работы [2, 3], посвященные исследованию временных характеристик Cloud computing и Web-сервисов. В этих работах исследованы статистические характеристики времени соединения компьютера с удаленным Cloud-провайдером, времени обработки данных и времени отклика сервера, включающее в себя время обработки данных и время соединения с удаленным сервером.

В процессе передачи данных возможен обрыв соединения и потеря данных. Влияние этих событий на статистические характеристики передаваемых данных в указанных работах не исследованы.

**Полученные результаты**. Статистические данные, использованные в работе были получены следующим образом [3]. Были разработаны программы «Клиент» и «Сервер». Программу «Сервер» устанавливали на удаленном компьютере Cloud Computing Azure. Программа «Клиент» собирала временные характеристики времени соединения и обработки данных удаленным сервером ежеминутно в течение 24 часов. Всего было получено 1440 значений, которые составили исходную выборку, с объемом потерянных данных 0%.

В связи с тем, что в процессе обмена данными в системе «Клиент - Сервер», системе К-С, неизбежны потери данных по самым разным причинам было проведено исследование влияния пропусков данных на результаты первичной статистической обработки.

Для анализа влияния количества пропущенных данных на результаты определения статистических характеристик выборки использовали следующий

прием. Из исходной выборки удаляли 0,1,3,6,9,12,15 процентов данных. Данные, предназначенные для удаления, выбирали случайным образом. Полученные для каждой выборки статистические характеристики сравнивали между собой и на этой основе делали выводы о влиянии количества пропущенных данных на качество полученных статистических выводов о свойствах выборок.

В соответствии с рекомендациями работы [5] были использованы следующие приемы обработки данных с пропусками:

1. Определяли первичные статистические характеристики исследуемой выборки без учета пропусков – метод «Склейка». При таком подходе объем выборки уменьшается, следовательно, уменьшается число степеней свободы, при проверке формулируемых статистических гипотез.

2. Определяли первичные статистические характеристики исследуемой выборки с заполнением пропусков нулями – метод «С нулём». В этом случае число степеней свободы не уменьшается, но возникает смещение оценок получаемых статистических характеристик в сравнении с исходной выборкой.

3. Определяли первичные статистические характеристики исследуемой выборки с заполнением пропусков средними значениями – метод «Среднее». В этом случае также возможно смещение оценок полученных характеристик в сравнении с исходной выборкой.

4. Определяли первичные статистические характеристики исследуемой выборки с заполнением пропусков квазислучайными числами, распределёнными по нормальному закону, среднее значение и среднеквадратическое отклонение которого совпадает с аналогичными характеристиками исходной совокупности данных с пропусками – метод «Случайные».

5. В дополнение к этим методам был использован приём, состоящий в следующем. На место пропущенных данных вставляли данные, распределённые по тому же закону, что и исходная выборка с пропусками – метод «Распределения».

Вид закона распределения определяли используя систему Statgraphics V.15. Результаты применения этой процедуры показаны в табл. 1

Полученные законы распределения имеют следующий вид:

T1 – функция плотности распределения наибольшего значения:

$$f(x)=\frac{1}{\beta}\exp\left\{-\frac{x-\alpha}{\beta}-\exp\left(\frac{x-a}{\beta}\right)\right\}. \qquad (1)$$

Таблица 1

Изменение закона распределения выборки в зависимости от способа устранения пропущенных данных*

| Способ устранения пропусков данных | Количество пропусков (%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0 | 1 | 3 | 6 | 9 | 12 | 15 |
| Склейка | T1 | T1 | T1 | T1 | T1 | T1 | T1 |
| Среднее | T1 | T1 | T1 | T1 | T2 | T3 | T3 |
| Случайные | T1 | T1 | T1 | T1 | T1 | T2 | T2 |
| Распределения | T1 | T1 | T1 | T1 | T1 | T1 | T1 |

* T1 – распределение наибольшего значения,
T2 – лог-логистическое распределение,
T3 – распределение Лапласа.

Параметры распределения $\alpha$ и $\beta$ связаны с математическим ожиданием m и дисперсией $s^2$ равенствами:

$$m=\alpha+\beta\Gamma^{-2}, \qquad s^2=\frac{\beta^2\pi^2}{6} \qquad (2)$$

T2 – логлогистическое распределение с плотностью вида:

$$f(x)=\frac{1}{\sigma x}\frac{\exp(z)}{[1+\exp(z)]^2}, \qquad (3)$$

где:

$$z=\frac{\ln(x)-\mu}{\sigma}.$$

Параметры положения $\mu$ и масштаба $\sigma$ в этом случае связаны с математическим ожиданием m и дисперсией $s^2$ условиями:

$$\mu=\exp(\mu)\Gamma(1+\sigma)\Gamma(1-\sigma), \qquad (4)$$

$$s^2=\exp(2\mu)\left[\Gamma(1+2\sigma)\Gamma(1-2\sigma)-\Gamma^2(1+\sigma)\Gamma^2(1-\sigma)\right], \qquad (5)$$

где $\Gamma(.)$ – гамма функция Эйлера.

T3 – распределение Лапласа, с плотностью :

$$f(x)=\frac{\lambda}{2}e^{-\lambda|x=\mu|}. \qquad (6)$$

Параметр этого распределения $\mu$ совпадает с математическим ожиданием, дисперсиея $\sigma^2$ связана с параметром $\lambda$ равенством:

$$\sigma^2=\frac{2}{\lambda^2}. \qquad (7)$$

Из данных, приведенных в таблице 1 следует, что при потере данных не свыше шести процентов все исследованные способы устранения пропусков сохраняют закон распределения исходной модельной выборки. При потерях данных от девяти до пятнадцати процентов только способы «Склейка» и «Распределения» сохраняют закон распределения исходной модельной выборки. При восстановлении пропущенных данных важно не только сохранить вид закона распределения исходной выборки, но и добиться того, чтобы исходная модельная выборка

(выборка без пропусков) существенно (в статистическом смысле) не отличалась от выборок с искусственно заполненными пропусками данных.

Для проверки влияния способа заполнения пропущенных данных на параметры полученных распределений использовали непараметрические критерии проверки гипотезы о совпадении выборок. Результаты проверок и перечень использованных критериев приведен в таблице 2. В качестве эталона во всех исследованных случаях принимали модельную выборку. При составлении принято, что символ (*) означает отсутствие различия между сравниваемыми выборками по всем использованным критериям. Список применённых критериев и их условные обозначениея приведены в примечании к табл.2. Все расчёты выполнены с использованием программы AtteStat.

Таблица 2

Влияние способа заполнения пропусков данных времени на сохранение вида и параметров закона распределения*

| Способ устранения пропусков | Количество пропусков (%) | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 3 | 6 | 9 | 12 | 15 |
| С нулем | * | К | В,М, ВВ,К | В,М, ВВ,З , К | В,М,В В,С,З, К | В,М,В В,С,З, К |
| Склейка | * | * | * | * | * | * |
| Среднее | * | * | * | З,А | ВВ, З,А | ВВ, З,А * |
| Случайные | * | * | * | * | *А,С | *С,К |
| Распределения | * | * | * | * | * | * |

*В – Критерий Вилкоксона, М – Критерий Манна-Уитни, ВВ – Критерий Ван дер Вардена, С – Критерий Сэвиджа, З – Критерий Зигеля-Тьюки, А – Критерий Ансари-Бредли, К – Критерий Клотца.

Наличие символа в ячейке таблицы указывает критерий, который отверг гипотезу о совпадении выборок. Таким образом, для практического использования можно рекомендовать только способы восстановления пропусков «Склейка» и «Распределения».

Для обоснованного выбора способа прогнозирования были изучены статистические свойства ряда. На первом этапе определяли фрактальные свойства ряда, а именно показатель Херста Н.

В работе [4] предложен следующий порядок расчета показателя Херста. Связь между показателем Херста Н и статистическими характеристиками ряда данных определяют в виде формулы:

$$R / S = \left(\frac{\pi}{2} N\right)^{H}, \qquad (8)$$

где S – среднее квадратическое отклонение временного ряда наблюдений,

N – количество наблюдений.

Тогда величину показателя Херста Н определяют по формуле:

$$H = \frac{\lg(R / S)}{\lg(\pi N / 2)}. \qquad (9)$$

В формуле (10) величину S- среднее квадратичное отклонение ряда наблюдений рассчитывают по формуле:

$$S = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{X})}. \qquad (10)$$

где $\overline{X}$ – среднее арифметическое изучаемого временного ряда наблюдений за N временных периодов.

Размах накопленного отклонения является наиболее важным элементом формулы расчета показателя Херста. В общем виде его рассчитывают так:

$$R = \max_{1 \le u \le N} Z_u - \min_{1 \le u \le N}, \qquad (11)$$

где: $Z_u$ – накопленное отклонение элементов ряда от среднего значения:

$$Z_u = \sum_{i=1}^{u}\left(x_i - \overline{X}\right). \qquad (12)$$

В работе [4] рекомендовано при количестве наблюдений N < 250 корректировать левую часть формулы (8), используя поправку:

$$R / S_T = R / S \times 0,998752 + 1,051037. \qquad (13)$$

Таблица 3

Определение показателя Херста для ВДС

| вся выборка | | последние сто | |
|---|---|---|---|
| Среднеарифметическое X | 2007,678 | Среднеарифметическое X | 1958,158 |
| Стандартное отклонение S | 255,464 | Стандартное отклонение S | 225,5333 |
| Размах R | 57494,667 | Размах R | 3698,297 |
| Нормированный размах R/S | 225,060 | Нормированный размах R/S | 16,39801 |
| Показатель Херста Нт | 0,709 | Показатель Херста Нт | 0,570525 |

Как показано в работе [4] в случае, когда показатель Херста (Ht) находится в интервале от 0,326 до 0,674, то это означает, что моделью изменения значений ряда будет винеровский процесс, физическим аналогом, которого является броуновское движение вокруг среднего значения ряда наблюдений.

В таблице 3 приведены результаты расчёта величины Н определённой по всему массиву данных («вся выборка») и по последним ста наблюдениям («последние сто»).

Полученные данные послужили обоснованием для выбора скользящего среднего в качестве метода прогноза.

Качество результатов прогнозирования определяли методом ретроспективного прогноза по шести последним данным с применением формулы:

$$\varepsilon = \sum_{i=m-k}^{m} \frac{|\hat{x}_i - x_i|}{x_i}, \qquad (14)$$

где: ε – величина средней относительной ошибки прогноза ,

m – объём массива данных, для которого определяли среднюю относительную ошибки прогноза,

$x_i$ – фактическое значение

$\hat{x}_i$ – вычисленное значение.

Таблица 4

Оценка качества прогноза ВДС

| Тип выборки | Порядок скользящего среднего | | |
|---|---|---|---|
| | 3 | 5 | 7 |
| «Вся выборка» | 0,072 | 0,049 | 0,065 |
| «Последние сто значений» | 0,028 | 0,035 | 0,041 |

На рис. 1 показан график изучаемого процесса. По оси x абсцисс отложено количество экспериментов (обращений к серверу), по оси ординат отложено время доступности сервера в миллисекундах.

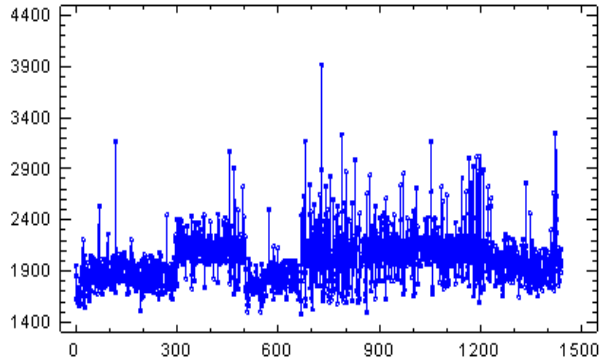


Рис. 1. Время доступности сервера

На рис.2 показана автокорреляционная функция процесс изменения величины ВДС. Из этого рисунка видно, что статистическая зависимость между последовательными отсчётами ВДС очень слабая, что служит косвенным подтверждением полученному значению величины H.

Для проверки стационарности процесса ВДС получен график автокорреляционной функции первых разностей процесса ВДС, показанный на рис.3и из которого следует, что процесс ВДС можно считать стационарным.

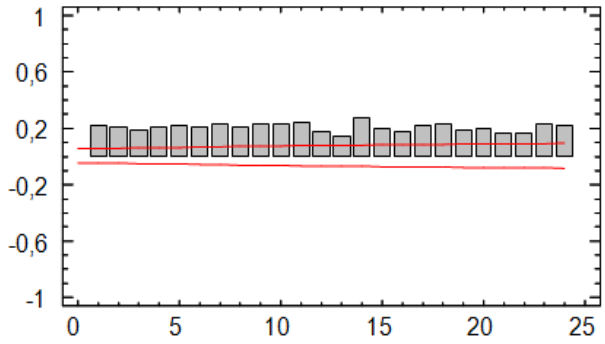


Рис. 2. Автокорреляционная функция случайного процесса ВДС

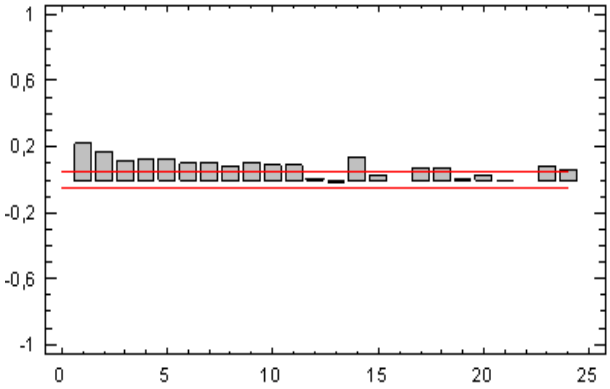


Рис. 3. Автокорреляционная функция первых разностей случайного процесса ВДС

## Выводы

1. Установлено, что при потере данных не свыше шести процентов сохраняется закон распределения исходной модельной выборки. При потере данных от девяти до пятнадцати процентов только способы «Склейка» и «Распределения» сохраняют закон распределения исходной модельной выборки.

2. В работе исследованы фрактальные свойств ряда ВДС с использованием покакзателя Херста. На этой основе выбран метод прогнозирования значений ряда ВДС. Качество результатов прогнозирования определяли методом ретроспективного прогноза по шести последним данным. Результаты показали, что для прогнозирования ВДС следует использовать метод скользящего среднего с шагом, равный трём.

## Литература

*1. Концепция Cloud Computing [Текст] / В. И. Лымарь, М. И. Макарова, Е. В. Зубкова, Т. В. Кортева // Конкурентоспособность территорий : материалы XV Всерос. экон. форума молодых*

*ученых с междунар. участием в рамках III Евразийского экономического форума молодежи «Диалог цивилизаций – „ПУТЬ НАВСТРЕЧУ“» (Екатеринбург, 17–18 мая 2012 г.) : в 9 ч. / [отв. за вып. М. В. Федоров, Э. В. Пешина]. – Екатеринбург : Изд-во Урал. гос. экон. ун-та, 2012. – Ч. 8. Направления : 11. Исследования менеджмента, маркетинга и логистики ; 19. Информационные процессы инновационного бизнеса. – С. 255–256.*

*2. Benchmarking Dependability of a System Web Application. [Text] / Yuhui Chen, Alexander Romanovsky, Anatoliy Gorbenko, Vyacheslav Kharchenko. // 14th IEEE Int. Conf. on Engineering of Complex Computer Systems. – ICECCS'2009: conference proceedings. – Potsdam (Germany), 2009. – P. 146–153.*

*3. Рубан, В. И. Экспериментальное исследование производительности Cloud Computing [Текст] / В. И. Рубан, А. В. Горбенко // Системы управления, навигации и связи. – 2012. – Вип. 4(1). – С. 189–191.*

*4. Найман, Э. Расчёт показателя Херста с целью выявления трендовости (персистентности) финансовых рынков и макроэкономических индикаторов [Текст] / Э. Найман // Економіст. - 2009. - №10. - С. 25-29.*

*5. Литтл, Р. Дж. А. Статистический анализ данных с пропусками [Текст] / Р. Дж. А. Литтл, Д. Б. Рубин ; пер. с англ. А. М. Никифорова. – М. : Финансы и статистика, 1991. – 334 с*

*Поступила в редакцию 3.03.2014, рассмотрена на редколлегии 25.03.2014*

**Рецензент:** д-р техн. наук, профессор В. А. Гороховатский, Харьковский институт банковского дела Университета банковского дела НБУ.

## ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ВІДНОВЛЕННЯ ПРОПУЩЕНИХ ДАНИХ ЗАМІРЮВАННЯ ЧАСУ ДОСТУПНОСТІ СЕРВЕРА В СИСТЕМАХ «ХМАРНИХ ОБЧИСЛЕНЬ»

***А. В. Горбенко, В. Ю. Дубницький, В. І. Рубан***

У статті приведені результати дослідження первинних статистичних характеристик часу обробки і передачі даних в системах «Хмарних обчислень». На підставі отриманих результатів виконаний аналіз методів відновлення пропущених даних. У роботі проаналізовані прийоми обробки даних з пропускам і вибрані найбільш раціональні. Досліджена зміна закону розподілу вибірки залежно від способу усунення пропущених даних. Показано, що при визначенні параметрів законів розподілу вибірки з пропущеними даними найкращим можна вважати метод при якому на місце пропущених даних вставляють квазівипадкові числа, розподілені по тому ж закону, що і початкова вибірка.

**Ключові слова:** «Хмарні обчислення», час доступності сервера, статистичні дані, закон розподілу, показник Херста, метод ковзного середнього.

## THE COMPARATIVE ANALYSIS OF THE METHODS OF RECOVERY OF THE PASSED DATA MESUREMENT OF TIME OF AVAILABILITY OF THE SERVER IN THE SYSTEMS OF "CLOUD COMPUTING"

***A. V. Gorbenko, V. Y. Dubnitskiy, V. I. Ruban***

Results specified of study in primary statistical characteristics of data processinbg and transmission time in cloud calculation systems. On the basis of results obtained methods of missing data recovery were analyzed. Techniques of missing data processing were analyzed and most reasonable of them selected. Sample selection law alteration studied depending on method of missing data removal. The best method for determination of missing data sample distribution law was shown to be one in which missed data were substituted with quasi-random numbers distributed under the same law that the initial sample.

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

**Горбенко Анатолий Викторович** – д-р техн. наук, профессор кафедры компьютерных систем и сетей Национального аэрокосмического университета им. Н. Е. Жуковского «ХАИ», Харьков, Украина, e-mail: A.Gorbenko @ csn.khai.edu.

**Дубницький Валерій Юрійович –** канд. техн. наук, старший научный сотрудник, заведующий научно-исследовательской лаборатории, Институт банковского дела Университета банковского дела НБУ, Харьков, Украина, e-mail: valeriy_dubn@mail.ru.

**Рубан Виталий Иванович** – аспирант кафедры компьютерных систем и сетей Национального аэрокосмического университета им. Н. Е. Жуковского «ХАИ», Харьков, Украина, e-mail: rubanvit@mail.ru.