

**PREDICTION OF DRMS WORKLOAD BY IDENTIFICATION OF PATTERNS
IN JOB SUBMISSION PROCESSES**

Andrey V. Gritsenko

Department of Applied Mathematics and Computer Science

Stavropol State University, Stavropol, 1, Pushkina str., 355009, Russia



Abstract

Resource management and scheduling algorithms used in modern distributed resource management systems allow to utilize up to 80% of resources. In the course of studying the issue to augment this percentage a new approach, which consists in the application of prediction methods together with traditional for DRMS scheduling methods, was proposed. It is generally known that a fairly large part of all submitted jobs is formed by recurring job and using information about these jobs one can alter a job schedule for the purpose of increase of resource utilization.

The process of simulation of a real sequence of submitted jobs with hidden cycles that inherently resemble white noise is described. On the basis of information about these cycles a prediction model is generated, which allow to simulate the process with enough high accuracy, and it in its turn confirms the assumption about potential appliance of forecasting methods in DRMS algorithms in order to raise their effectiveness.

Key words: *distributed resource management systems, DRMS, job scheduling, forecasting methods, stochastic process simulation.*

1. INTRODUCTION

Widespread using of distributed computing systems entailed development of diverse resource management and scheduling algorithms at first, and then appearance of a great amount of software applying those algorithms. At present the distributed resource management systems' market is represented by developments of both such companies as Adaptive Computing (their products are Maui and Moab) and Oracle (Oracle Grid Engine), as well as of those that are not so well-known [2]. In spite of all superficial differences of these software products, there is some similarity among them – scheduling algorithms of these products take into account only those jobs that were already submitted and added to a query when constructing a schedule. In the article [1] there is assigned reasoning for a point of view, which claims that appliance of current jobs only at the stage of scheduling may be the weakness of an algorithm and may lead to the effectiveness loss. In particular it is stated that in some cases it is not unlikely that the same task may be submitted with a certain periodicity, and the assumption is put forward that the opportunity to utilize forecasting methods in scheduling algorithms may be very useful and could significantly increase the overall effectiveness of DRMS.

2. XLSTAT STATISTICAL ANALYSIS

The prevalent belief that job submission is a stochastic process and its composition resembles white noise because of great amount of users who take part in it, is one of the main reasons why forecasting methods are still not used in job schedulers' algorithms.

To verify this assumption XLSTAT – an add-on for Microsoft Office Excel that possesses an extensive set of functions for statistical analysis of time series – was applied [3, 4]. As an input data we used a time series with three variables: date of task initiation, amount of processes, required for a task to execute and task duration – that was obtained as a result of the examination of the workload of the distributed computing system at Stavropol State University. Several prolonged periodical patterns were revealed among other data of this process.

First of all we divided initial time series into 2 of smaller dimensions: first contained information only about required number of processes ('Procs') and second – about job duration time ('Time') – and then analyzed these time series using following XLSTAT features: Descriptive Analysis [3] and Normality Tests [4].

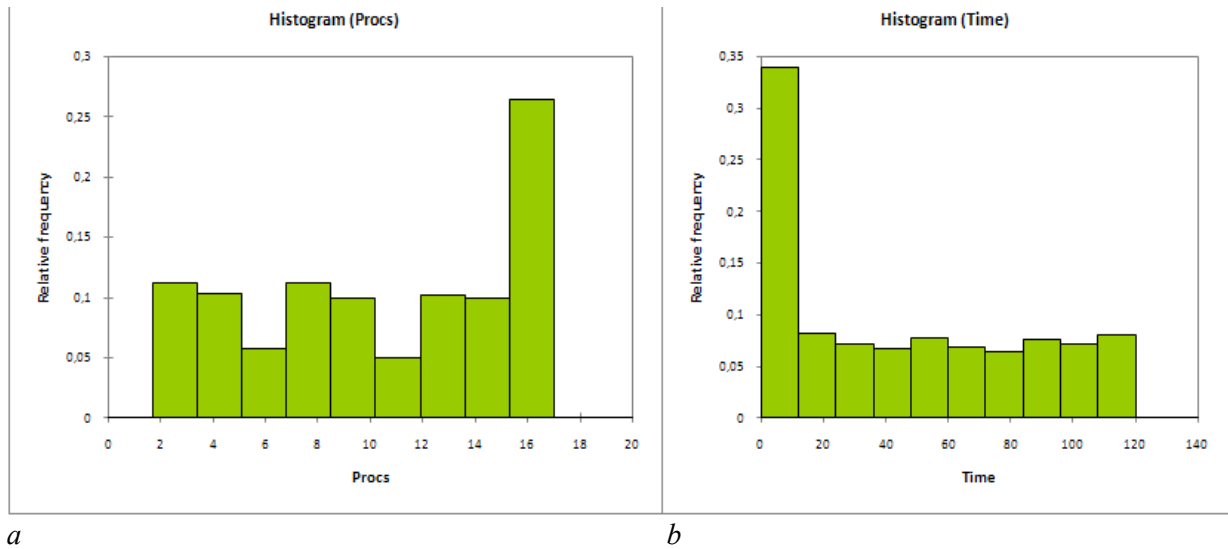
Table 1. XLSTAT Descriptive analysis result table

Statistic	DF	Value		p-value	
		'Procs'	'Time'	'Procs'	'Time'
Jarque-Bera	2	345.320	377.898	< 0,0001	< 0.0001
Box-Pierce	6	17.443	22.235	0.008	0.001
Ljung-Box	6	17.469	22.256	0.008	0.001
McLeod-Li	6	30.015	19.065	< 0,0001	0.004
Box-Pierce	12	44.827	44.891	< 0,0001	< 0.0001
Ljung-Box	12	44.931	44.978	< 0,0001	< 0.0001
McLeod-Li	12	71.131	34.034	< 0,0001	0.001

Descriptive Analysis feature involves Jarque-Bera, Box-Pierce, Ljung-Box and McLeod-Li tests, frequency histograms and autocorrelogram plotting. The Jarque-Bera test is a normality test, based on the skewness and kurtosis coefficients: the bigger the value of the Chi-square statistic (third column), the more unlikely the null hypothesis that the data are normally distributed. The other three tests (Box-Pierce, Ljung-Box, McLeod-Li) are computed at different time lags (second column) and allow to test if the data could be assumed to be a white noise or not. These tests are also based on the Chi-square distribution. Here the p-value, which corresponds to the probability of being wrong when rejecting the null hypothesis (or accepting if true), is much less than 0.01 for both 'Procs' and 'Time' time series eo ipso suggesting one to conjecture that these time series are white noise [3].

Other normal tests combined in the context of Normality Tests, such as Shapiro-Wilk, Anderson-Darling and Lilliefors tests, also say that the null hypothesis should be rejected and the alternative hypothesis should be accepted. The falsity of the null hypothesis about normal distribution of 'Procs' and 'Time' time series can be graphically illustrated by means of frequency histograms (pic. 1)

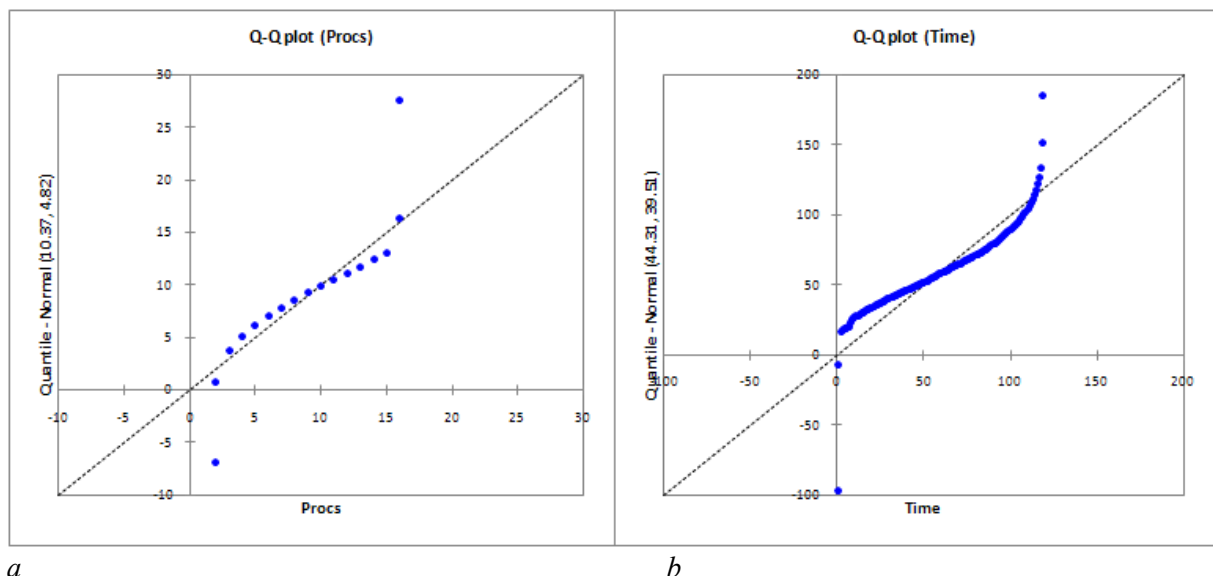
The frequency histogram gives us another argument to discard the null hypothesis:



Picture 1. Frequency histogram of 'Procs' (a) and 'Time' (b) time series

As we can see values in all intervals are distributed relatively uniformly except for two significantly extended rectangles, and as we would see further these rectangles are correspond to the cycle values.

In addition to normality tests the feature of the same name also constructs the Q-Q plot comparing the cumulative distribution function of the input data to the cumulative distribution function of normal distribution with the same mean and standard deviation [4].

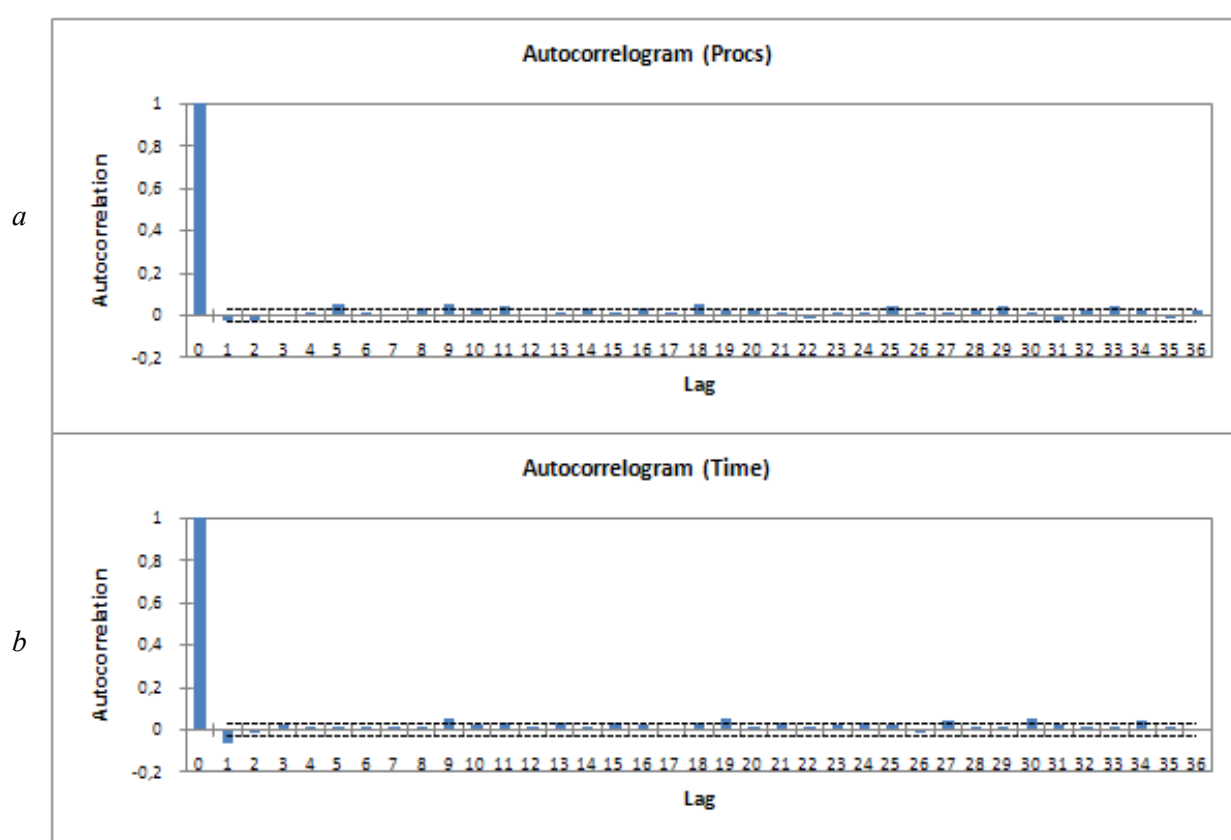


Picture 2. Q-Q plots for 'Procs' (a) and 'Time' (b) time series

In the case of values of the input time series follow a normal distribution, we should observe an alignment with the first bisecting line, in other cases some deviations from the bisecting line should be

observed. On picture 2 we can observe that there are two strong deviations on both of Q-Q plots, indicating that the distributions of the time series are most probably not normal. We can also notice that the most of these deviations corresponds to the abnormal value on the histogram charts, i.e. to the cycle values.

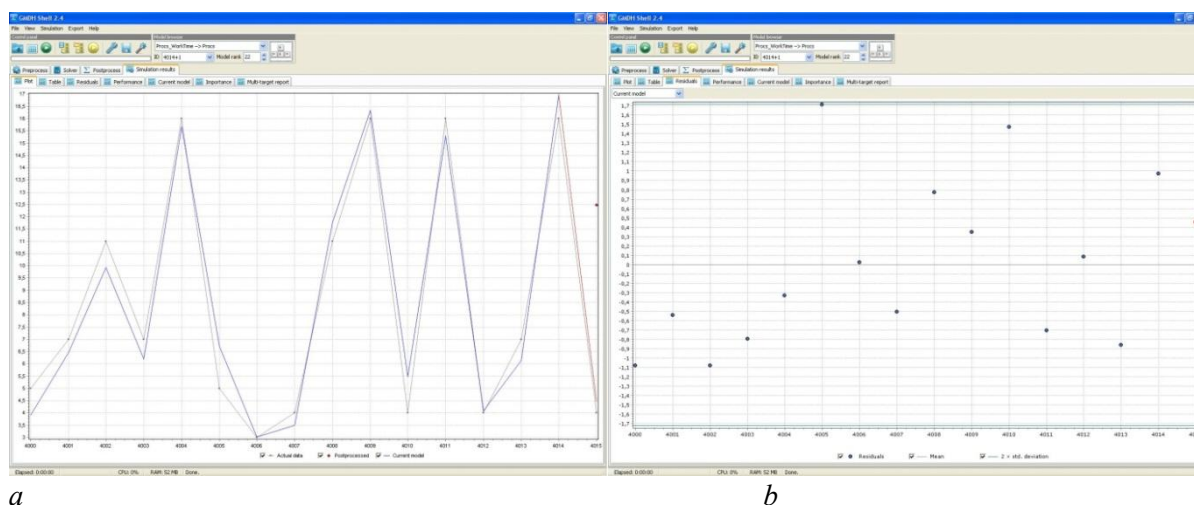
As a result of all of the above, it is most unlikely that one should accept the null hypothesis that the data are normally distributed. At the same time one should also conclude that data is not generated by a white noise process taking into consideration that it is known for certain that there are some recurring jobs. However, despite this fact, from the current autocorrelograms (pic.3) it follows that no cycles or other periodic patterns are present in the initial time series. This phenomenon can be explained by the fact that repetitive patterns are buried under the noise of other jobs, which is the result of the variable number of jobs submitted per day.



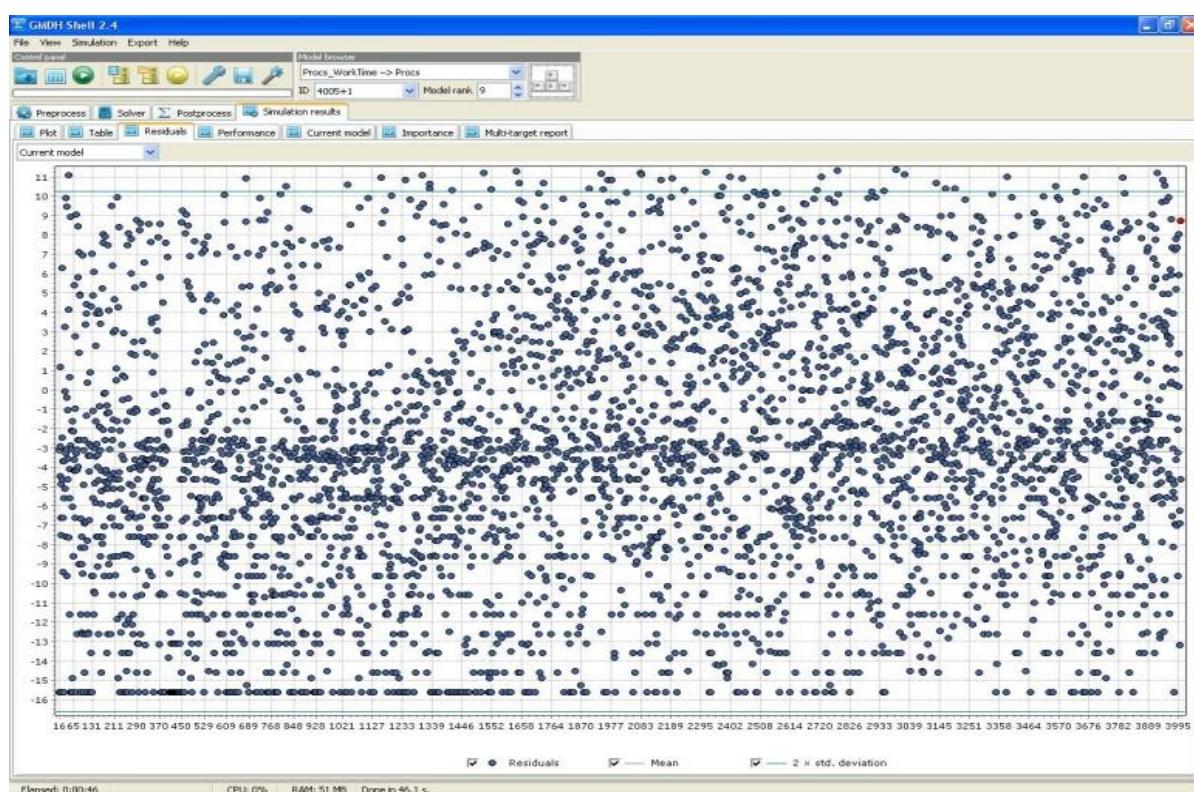
Picture 3. Autocorrelograms for 'Procs' (a) and 'Time' (b) time series

3. STOCHASTIC PROCESS MODELING USING GMDH SHELL

Though the preceding statistic analysis revealed the resemblance of the observable job submission process with the white noise an attempt to simulate the behavior of this process using specific software was made. As the appliance of traditional forecasting methods, e.g. ARIMA, was not successful since error of these methods exceeded several times the maximal value of the sample we used GMDH Shell instead – a software package developed by Geos Research Group and utilizing Group Method of Data Handling to solve different problems such as time series forecasting, regression and classification [5].



Picture 4. Simulation model (a) and its error (b) for 100 element sample. Last 15 elements are shown

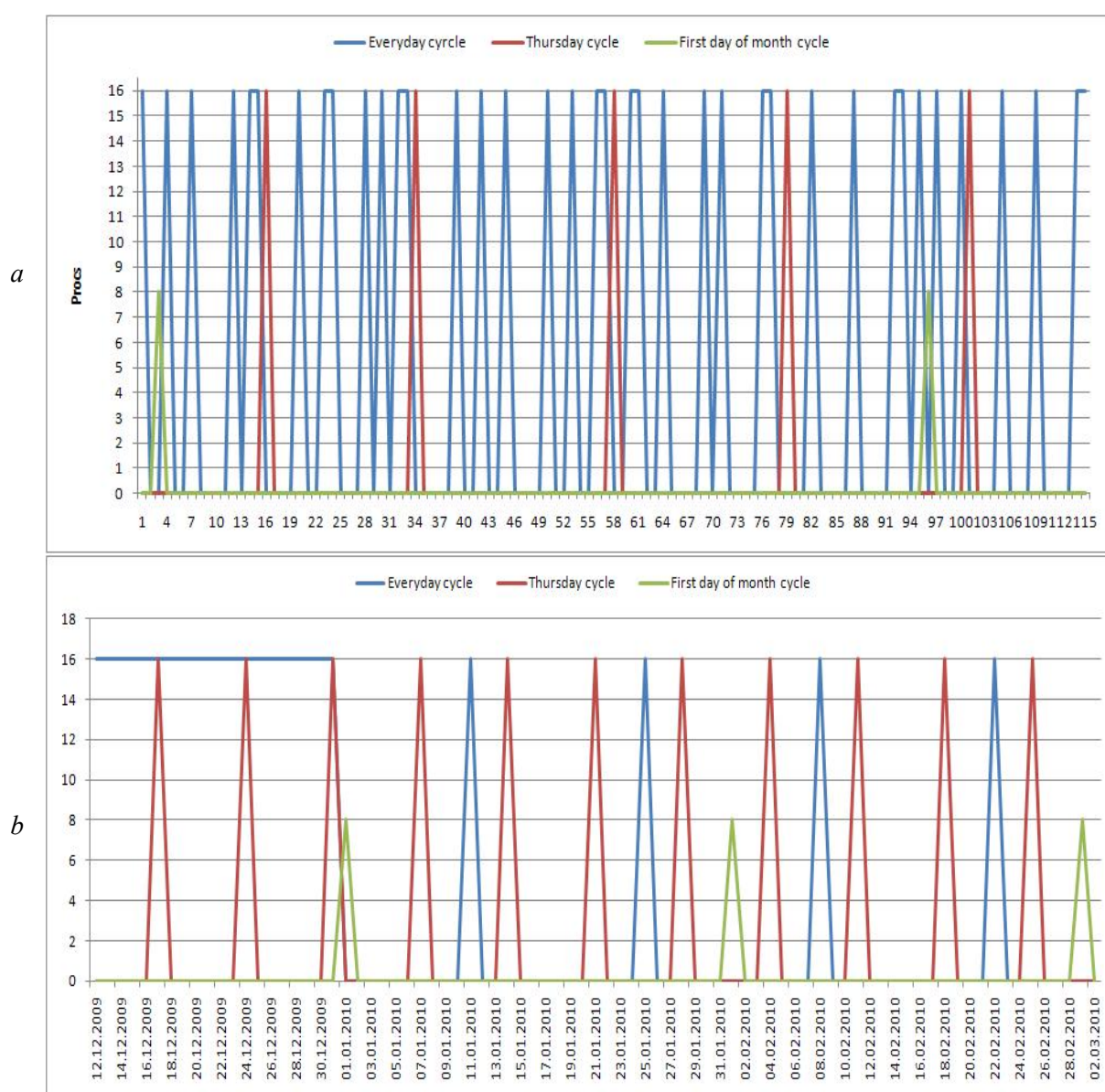


Picture 5. Residuals for the model of 4000 elements size

In spite of strong mathematical background GMDH Shell did not always give appropriate results: for a small learning window size (the dimension of the input dataset) this package generated rather accurate simulation models of the given stochastic processes with the errors that do not exceed 10% of the maximum value (pic. 4), though such models do not take into consideration recurring patterns because

of the periods of these patterns that could be comparable to the size of sample. Consequently we had to enlarge the size of the training data, but simultaneously with the growth of the input dataset the residuals grew as well and for learning window size of the order of 4000 they exceeded any admissible level (pic. 5), eo ipso making it embarrassing to use this software product for the current problem.

Taking into account the fact that we know for sure about recurring jobs but none of the described approaches allowed us to gain desired results it was proposed to use a trick consisting in the explicit display of the existing cycles. As a result we discovered 3 prolonged cycles: jobs that were submitted every first day of a month, jobs that were submitted every Thursday for a parochial period of time and those that firstly were submitted everyday but then their submission period changed to two weeks (pic. 6).



Picture 6. Patterns in the initial time series (a) and per diem view (b)

Having knowledge about existing cycles we can use it to construct more reliable prediction model using as a basis one of the simulation models produced by GMDH Shell. In this new model we use a special reliability coefficient, which indicates the probability that for the next time step certain cycle will still exist, dependent on the average duration of other cycles. Selecting an appropriate model and reliability coefficient for the current data we could gain satisfactory results (pic. 7)

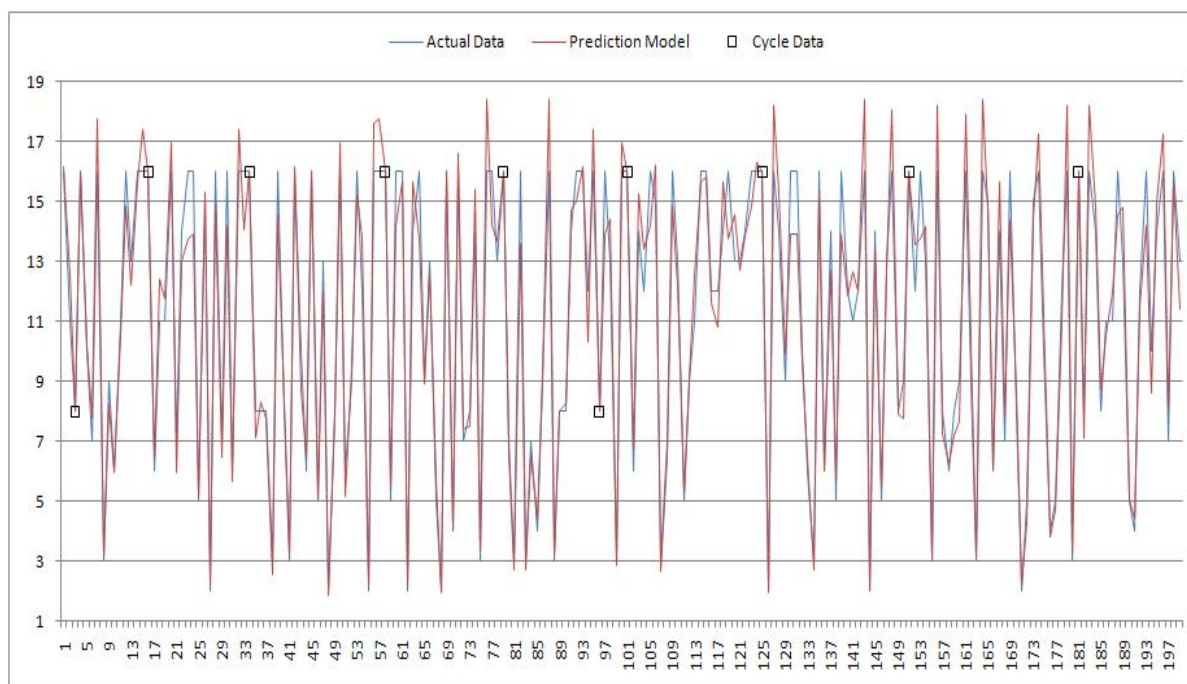


Рис.7. Prediction model that uses information about existing cycles in input data

5. CONCLUSION

It is obvious that the described above approach of obtaining a prediction model that also considers information about periodical patterns should be used only in the aggregate with traditional job-scheduling and resource-managing methods to gain us benefit, but not as a separate DRMS algorithm

The further investigations in this direction would consist in finding the most applicable way of adjusting the reliability coefficient and using self-training methods to make it possible for the algorithm to alter its parameters according to the input data.

REFERENCES

1. Gritsenko A., Shulgin A. Applying of forecasting methods in job management and resource allocation algorithms in high performance systems issue investigation. In the World of Scientific Discoveries, 2011, vol. 8(20), p 43-55

Journal of International Scientific Publications:
Materials, Methods & Technologies, Volume 6, Part 1

2. GRIDCLUB.RU. GRIDCLUB.RU: Grid technologies Internet portal [online]. Moscow: GRIDCLUB.RU, 2010. Date of revision 10 March 2012 [cited 28 January 2012]. Cluster systems. Available from World Wide Web: <<http://www.gridclub.ru/software/cluster.html>>
3. Addinsoft. Statistics package for Excel – XLSTAT [online]. New York: Addinsoft, 2012. Date of revision 20 February 2012 [cited 20 February 2012]. Description and transformation of a time series with XLSTAT-Time. Available from World Wide Web: <<http://www.xlstat.com/en/learning-center/tutorials/using-differencing-to-obtain-a-stationary-time-series.html>>
4. Addinsoft. Statistics package for Excel – XLSTAT [online]. New York: Addinsoft, 2012. Date of revision 20 February 2012 [cited 20 February 2012]. Sampling and Normality test. Available from World Wide Web: <<http://www.xlstat.com/en/learning-center/tutorials/sampling-data-in-a-distribution-and-running-a-normality-test-with-xlstat.html>>
5. Geos Research Group. GMDH Shell [online]. USA: Geos Research Group, 2012. Date of revision 10 March 2012 [cited 10 March 2012]. Product Overview. Concepts. Available from World Wide Web: < <http://www.gmdhshell.com/docs/concepts> >