Research article
DOI: https://doi.org/10.18721/JCSTCS.17304
UDC 519.25

COUNT TIME SERIES ANALYSIS OF JOBS SCHEDULING IN THE HYBRID SUPERCOMPUTER CENTER

S.V. Malov ✉ , *A.A. Lukashin*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ sergey.v.malov@gmail.com

Abstract. Increasing the efficiency of supercomputer centers is an extremely important task, especially in the context of growing demand for high-performance computing and a shortage of supercomputer resources. Statistical analysis of the results of various indicators of supercomputer performance is aimed at creating models of computing resource management and forming a basis for using artificial intelligence methods. The purpose of this research is to study the incoming flow of user requests (jobs), which largely determines the load on supercomputer resources. To analyze the incoming flow of user jobs, generalized linear models and generalized estimating equations, as well as the autoregressive conditional Poisson model, were used. It allowed taking into account the dependence of observations and the effect of overdispersion. Based on the results of supercomputer operation observations, estimates of the time trend were obtained, as well as indicators of changes in the intensity of the job flow within weekly and annual cycles with classification by areas of expertise and computing clusters. Indicators of statistical significance of changes within the weekly and annual cycles were established. As a result of an advanced statistical analysis using multiple comparison methods, statistically significant orders of the main effects of the weekly and annual factors were obtained.

Keywords: count time series, generalized estimating equations, autoregressive conditional Poisson model, multiple comparisons, supercomputer cluster, job scheduling

Acknowledgements: The research was partially financially supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of the state assignment “Development and research of machine learning models for solving fundamental problems of artificial intelligence in the fuel and energy complex” (FSEG-2024-0027). The results of the work were obtained using the computational resources of the shared-use center “Polytechnic Supercomputer Center” of Peter the Great St. Petersburg Polytechnic University (No. 500675, https://ckp-rf.ru/catalog/ckp/500675/).

Citation: Malov S.V., Lukashin A.A. Count time series analysis of jobs scheduling in the hybrid supercomputer center. Computing, Telecommunications and Control, 2024, Vol. 17, No. 3, Pp. 42–53. DOI: 10.18721/JCSTCS.17304

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.17304
УДК 519.25

АНАЛИЗ ВРЕМЕННЫХ РЯДОВ ЧАСТОТ ДЛЯ ПЛАНИРОВАНИЯ ЗАДАЧ ГИБРИДНОГО СУПЕРКОМПЬЮТЕРНОГО ЦЕНТРА

С.В. Малов ✉ , *А.А. Лукашин*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ sergey.v.malov@gmail.com

Аннотация. Повышение эффективности использования суперкомпьютерных центров является крайне важной задачей, особенно в условиях растущего спроса на высокопроизводительные вычисления и дефицит суперкомпьютерных ресурсов. Статистический анализ результатов различных показателей функционирования суперкомпьютера направлен на создание моделей управления вычислительными ресурсами и формирование базы для использования методов искусственного интеллекта. Целью данного исследования является изучение входящего потока заявок пользователей, во многом определяющего загрузку ресурсов суперкомпьютера. Для анализа входящего потока заявок пользователей используются обобщенные линейные модели и обобщенные уравнения оценивания, а также пуассоновская авторегрессионная модель, применение которых позволяет учитывать зависимость наблюдений и эффект избыточной дисперсии. По результатам наблюдений за работой суперкомпьютера получены оценки временного тренда, а также показатели изменений интенсивности потока заявок в рамках недельного и годового циклов с классификацией по областям знаний и вычислительным комплексам. Установлены показатели статистической значимости изменений в рамках недельного и годового цикла с учетом данной классификации. В результате углубленного анализа с использованием методов множественного сравнения получены статистически значимые порядки главных эффектов недельного и годового факторов.

Ключевые слова: дискретные временные ряды, обобщенные уравнения оценки, пуассоновская условно авторегрессионная модель, множественные сравнения, суперкомпьютерный кластер, планирование задач

Финансирование: Исследование выполнено при частичной финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках государственного задания «Разработка и исследование моделей машинного обучения для решения фундаментальных задач искусственного интеллекта в топливно-энергетическом комплексе» (FSEG-2024-0027). Результаты работы получены с использованием вычислительных ресурсов центра коллективного пользования «Суперкомпьютерный центр “Политехнический”» Санкт-Петербургского политехнического университета Петра Великого (№ 500675, https://ckp-rf.ru/catalog/ckp/500675/).

Для цитирования: Malov S.V., Lukashin A.A. Count time series analysis of jobs scheduling in the hybrid supercomputer center // Computing, Telecommunications and Control. 2024. Т. 17, № 3. С. 42–53. DOI: 10.18721/JCSTCS.17304

Introduction

High-performance computing is an important element in computer-aided engineering and fundamental research. Large world-leading research centers use their supercomputers, while the smaller ones use supercomputers operating in shared-use centers. A shared-use center serves a wide variety of users conducting research in various domains including but not limited to mechanical engineering, physics, electronics, life sciences, artificial intelligence etc. This results in very different jobs running on the same supercomputer cluster in terms of number of cores, memory, software, and time [10]. This makes job scheduling more complex and inefficient as it is difficult to set parameters suitable for all types of jobs.

Modern supercomputers, possessing significant computational resources, simultaneously perform many jobs belonging to different fields of knowledge and imposing different requirements to computational resources and software calculations. Users set jobs for execution using a job scheduler, which forms a queue and schedules their using of the supercomputer resources. The statistical analysis of user job flow is significant for understanding the specifics of using supercomputers as shared-use centers. It allows to proceed to the development of intelligent algorithms for increasing the efficiency of supercomputer system resource usage. The load on the supercomputer resources is largely determined by the incoming flow of user jobs, which is studied in this paper.

Statistical data on supercomputer operation provides new opportunities for optimizing resource utilization. Understanding the parameters of user job flow allows to significantly improve the overall performance of supercomputer systems. The work on data collection and analysis is described in [6], and works [1, 10] demonstrate statistical and machine-learning analysis of supercomputer data. We perform statistical analysis of incoming flow of user jobs that determine requirements of the supercomputer resources at any given time. Statistical analysis of the incoming flow of user jobs allows to optimize the tools of queue management for executing computational jobs and distributing them among computational clusters. For this study, a dataset containing two years of supercomputer center jobs information was collected.

The Poisson process model is applicable for an ideal homogeneous flow. The number of jobs in disjoint and identically sized time intervals are independent and identically distributed random variables having a Poisson distribution with some fixed $\lambda > 0$. The homogeneity requirement of the job flow is too restrictive in practical cases, prompting the use of advanced models for analysis. Statistical analysis of heterogeneous job flow is usually based on time series data on the number of jobs obtained in some equal time intervals (e.g. days). Classical methods in time series analysis require observations to be normally distributed, which is not applicable to count data, especially if some atoms have sufficiently high probabilities. In the particular case of counts of jobs with sufficiently high probabilities of small counts, the classical time series analysis is not applicable. The generalized log-linear regression model (see, e.g. [8]) can be used for statistical analysis of homogeneous counting time series, if the observed counts are independent and have Poisson distribution. The property of equidispersion (equality of mean and variance) of the Poisson distribution is often violated in favor of overdispersion. The same estimating equations lead to consistent estimator of the regression parameters under some mild regularity conditions, even if the independence and the Poisson distribution properties are not satisfied and the number of the observed count time series tends to infinity and the length of each time series remains fixed, which is typical for longitudinal data analysis [2]. The consistent robust variance estimator can be obtained using so-called “sandwich” method. The use of so-called “working correlation matrix” and the generalized estimating equations (GEE) [7, 11] gives more efficient estimators of the regression parameters. It should be noted that the consistency of the robust variance estimator is confirmed as the number of the observed time series increases, whereas at a fixed number of the time series of the increasing length, the consistent variance estimation requires some restrictions on the distributions and dependence structure of the observations.

The alternative framework in heterogeneous flow data analysis is the conditional Poisson model. The multivariate 1st order Poisson autoregressive model [4] assumes that the conditional distribution of count Y_{it} at time t has the Poisson distribution with the following parameter:

$$\mu_{it} = a_{it}\lambda_{it} + v_{it}Y_{i,t-1} + \gamma_{it}\sum\nolimits_{j\neq i} v_{ij}Y_{j,t-1}, \tag{1}$$

where $\log\lambda_{it} = \boldsymbol{X}'_{it}\beta$; $\log v_{it} = \boldsymbol{Z}'_{it}\beta$ and $\log\gamma_{it} = \boldsymbol{G}'_{it}\beta$, whereas $\boldsymbol{X}'_{it}$, $\boldsymbol{Z}'_{it}$ and $\boldsymbol{G}'_{it}$ are the regressors. The Poisson autoregressive model as a natural generalization of the Poisson model with independent counts has much wider application area due to the independent counts property violation and the overdispersion effect. In the particular case of the spatial component absence $\gamma_{it} = 0$ (see [3]),

$$\mathbf{D}Y_{it} = \frac{\mathbf{E}Y_{it}}{1-v_{it}^2} \text{ and } \mathbf{Cov}\left(Y_{it}, Y_{it+k}\right) = \frac{v_{it}^k \mathbf{E}Y_{it}}{1-v_{it}^2}.$$

The multivariate Poisson autoregressive spatial model is widely used in epidemiology. A set of statistical tools for multivariate Poisson autoregressive spatial model is implemented in package *surveillance* [5, 9] for the R programming language[1].

For stratified statistical analysis of user job flows the Poisson log-linear generalized model and the independence estimating equations with the robust "sandwich" variance estimator, implemented in the *geepack* R-package, were used, as well as the univariate 1st order Poisson autoregressive model. All observed jobs were divided into 11 groups based on user area of expertise and 5 groups based on the computing clusters, to which the jobs were submitted, and only 4 of the 5 groups were analyzed. The generalized regression models included a smooth time trend as well as weekly/annual periodic factors. The main goal of the statistical analysis was the investigation of the dynamic change of the intensities of job flow over time in the presence of periodic factors classified by user's area of expertise and computing cluster. In addition to the regression fit and the statistical significance analysis of the periodic factors, some significant partial orders of the main effects using advanced contrasts analysis were obtained.

Explanatory analysis of users' job flow

The study examined historical data on job execution in the "Polytechnic Supercomputer Center". In total, the dataset contained 1545793 records of running jobs. Each record contained a user label, the number of requested resources (processors and supercomputer nodes), and job execution parameters, including how many and what resources were issued, when and how the job was completed. Based on the user label, each job was assigned to an area of expertise, such as physics or mechanics. A total of 11 areas of expertise were identified:

- astrophysics;
- bioinformatics;
- biophysics;
- energetics;
- geophysics;
- IT;
- mechanical engineering;
- mechanics;
- physics;
- radiophysics;
- a special group called geovation.

The last group is related to geophysical software, which runs in an automated mode (the jobs are submitted to the supercomputer queue automatically). Also, these jobs are quite small, but there are a lot of them processed in parallel. This explains the significant number of such jobs, but compared to the number of consumed resources (in terms of node-hours) the figures will be different. All jobs were divided into separate queues representing computing clusters, to which they were submitted:

- "Tornado" – a homogeneous cluster based on CPU (612 node cluster with 28-core compute nodes);

[1] The R Project for Statistical Computing, Available: https://www.R-project.org/ (Accessed 25.09.2024)

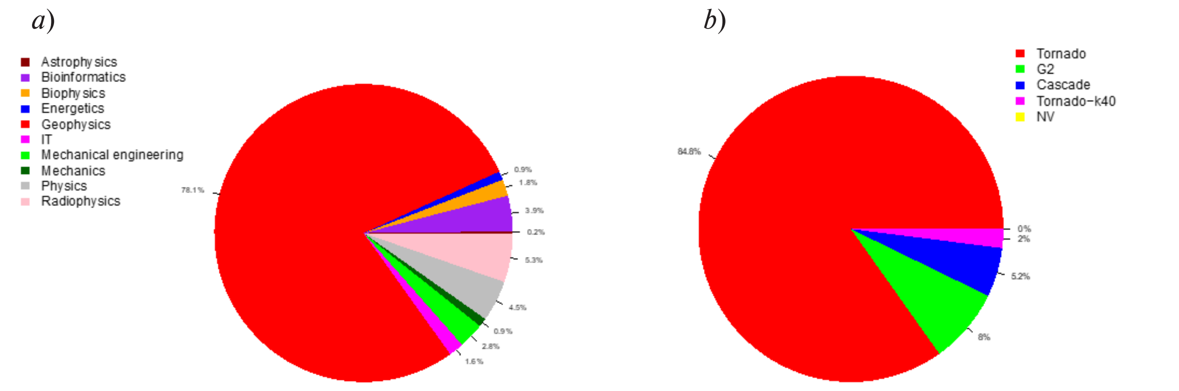

Fig. 1. Percentage of jobs: (*a*) from users in different areas of knowledge; (*b*) different computing clusters

- "G2" – a special cluster for geophysics;
- "Cascade" – a homogeneous cluster with large nodes (81 node cluster with 48-core compute node);
- "Tornado-k40" – a heterogeneous cluster with GPUs (56 node cluster with 28-core nodes with 2 GPUs);
- "NV" – a heterogeneous cluster with GPUs with large nodes (48-core nodes with 8 GPUs).

The percentage of received jobs depending on grouping factors is shown in Fig. 1.

The number of user jobs received from 01.09.2021 to 31.08.2023 is given in Table 1.

Table 1

The number of user jobs divided into groups

Area of expertise	**Computing cluster**					
	Tornado	**G2**	**Cascade**	**Tornado-k40**	**NV**	**Total**
Astrophysics	2812	0	0	0	0	2812
Bioinformatics	59567	0	0	66	0	59633
Biophysics	23830	2	1	3788	0	27621
Energetics	13893	12	238	145	18	14306
Geophysics	4985	8	1632	1199	2	7826
Geovation	984698	122755	77596	10464	0	1195513
IT	17734	3	421	6780	0	24938
Mechanical engineering	35476	0	51	7174	13	42714
Mechanics	14076	31	44	331	0	14482
Physics	67988	0	0	747	0	68735
Radiophysics	82047	0	0	125	0	82172
Total	1307106	122811	79983	30819	33	1540752

It should be noted, that the distribution of numbers in table 1 is highly unbalanced, with the majority of jobs (63.9%) coming from users in the geophysics area of expertise and being processed by the "Tornado" computing cluster. Moreover, the simultaneous use of two grouping factors, area of expertise and computing cluster, is unpractical due to the presence of a large number of empty cells. Since the total number of jobs received on the computing cluster "NV" the corresponding flow was not analyzed.

Fig. 2 shows the additive time trend estimators of the combined flow using the moving average method with a window size of 365 days, the smoothed moving average obtained by kernel smoothing of the

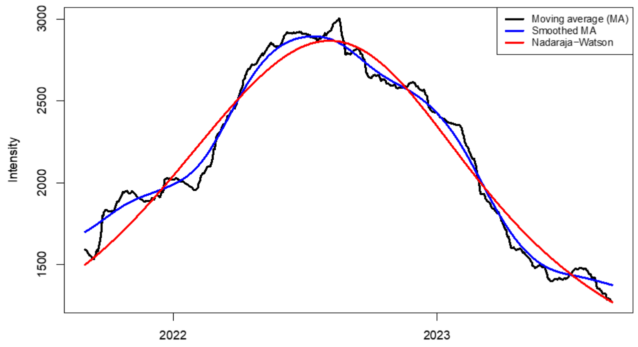

Fig. 2. The additive trend obtained by three different methods

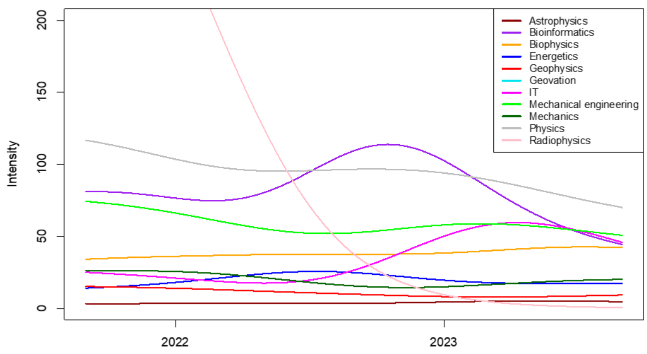

Fig. 3. The additive time trend for user job flow from different areas of expertise
(except for the geovation area)

moving average estimator with the Gaussian kernel and a sufficiently small smoothing parameter of 30, as well as the Nadaraya–Watson estimator with the Gaussian kernel and smoothing parameter of 120. The window size for the moving average method was chosen to exclude the seasonal component effect, and the smoothing parameter for the Nadaraya–Watson estimator was chosen to obtain the estimator sufficiently similar to the moving average. The presence of a time trend in the combined job flow and a sufficient increase in intensity in 2022 are evident, which explains the need to consider the time trend in the statistical analysis models. It should be noted, that the time trend of the combined job flow is determined primarily by jobs in the geovation group, since these jobs are the majority. The Nadaraya–Watson estimators of the additive time trends with the same smoothing parameter depending on the user's area of expertise (except for the geovation group) are presented in Fig. 3. An increase in the intensity of the job flow from the bioinformatics group in the second half of 2022 should be noted, while the other groups are not typical by this effect. Additionally, a significant decrease in the intensity of the job flow for users in the radiophysics area of expertise should be noted and, to a lesser extent, physics area, as well as a slight increase in the intensity of the job flow for users in the IT area of expertise.

A study of changes in the intensities of the job flows over time depending on the computing cluster, to which they were submitted (see Fig. 4.), shows that the increase in the intensity of the job flow observed in 2022 is characteristic only for "Tornado" cluster, and there is also a decrease in the intensity of the job flow for "G2" and an increase for "Cascade" clusters.

The variety of the time trends for different job flows is a strong argument in favor of using nonparametric trend estimates in regression models.

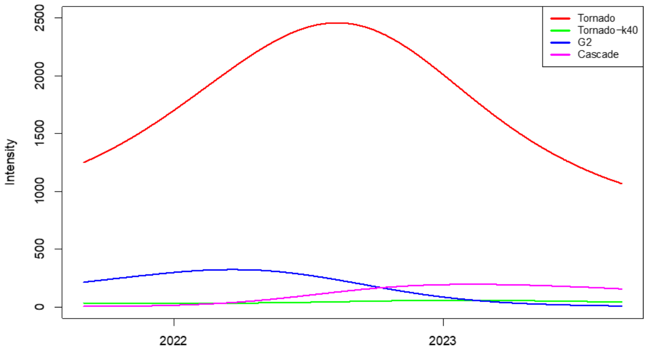

Fig. 4. The additive time trend for user job flow obtained in different computing clusters

Regression analysis of user job flows

The stratified statistical analysis of the job flows groped separately by user's area of expertise and the computing cluster was performed. All statistical inferences were adjusted for 15 flows, including the combined job flow and excluding the "NV" flow, therefore the significance level taking into account the Bonferroni correction is $1/300 \approx 0.003$.

To investigate behavior of the job flows intensities within the annual and weekly cycles, the GEE framework based on the generalized log-linear Poisson model with two additive factors Month and Day of the Week was used:

$$\log(\lambda_t) = \mu_1 + \alpha_j 1_{\{Month=j\}} + \beta_r 1_{\{Day=r\}} + \log(X_t), \tag{2}$$

where λ_t is the intensity of the job flow, X_t is the corresponding estimated time trend and t is a day of observation from the beginning of the study, and independence estimating equations. In order to fit the models the R-function *geeglm*() of package *geepack* was used.

Estimates of the multipliers for weekly and annual cycles are given in Tables 2 and 3, respectively. P-value in the last column characterizes the statistical significance of the effect of the corresponding factor on the intensity of the job flow.

The statistical analysis revealed a statistically significant effect of the annual periodic factor for each of the job flows adjusted to the total number of flows, while the effect of the weekly periodic factor was significant only for the mechanical engineering, mechanics an radiophisics flows, as well as for the combined flow and for "Tornado", "Tornado-k40", "G2" flows. Within the annual cycle, a slight decrease in the intensity of user job flows in the summer was observed, which is typical only for researchers in some areas of expertise, and a large variation in intensity throughout the year for researchers in radiophysics, information technology and bioinformatics areas of expertise. It should also be noted, that the sufficient increase of the intensity of the "Cascade" job flow at the end of the year had occurred.

The advanced statistical analysis of pairwise contrasts for the main effects of periodic factors allowed to find several partial orders with a joint reliability of 95%. Let θ_i and θ_j be the logarithmic main effects of levels i and j, respectively, of the factor under study. The pairwise contrast $\psi_{ij} = \theta_i - \theta_j$ allows to determine, whether the main effect of i-th level is smaller than, equal to or larger than the main effect of j-th level.

In order to obtain statistically significant inferences, two-sided joint confidence intervals for the parameters ψ_{ij} with all pairs of levels i and j were constructed using the Bonferroni method. If the confidence interval for the parameter ψ_{ij} lies entirely to the right of zero, the main effect of level i of the factor is less than the main effect of level j, and if it lies entirely to the left of zero, the main effect of level i is larger than the main effect of level j. All the significant inferences obtained in such a manner have the joint reliability of at least 95%.

Table 2

Multipliers and main effects for weekly cycle

Flow	MULT	MON	TUE	WED	THU	FRY	SAT	SUN	P-value
Astrophysics	0.87	0.98	1.02	1.1	1.05	1.13	0.89	0.88	0.7849
Bioinformatics	0.54	1.93	0.69	1.36	1.22	0.8	0.78	0.73	0.5897
Biophysics	0.91	1.16	1.14	1.11	0.94	1.07	1.09	0.62	0.3994
Energetics	0.84	1.27	1.39	1.09	1.04	0.82	0.73	0.83	$9.9*10^{-3}$
Geophysics	0.83	0.97	1.78	1.15	0.88	1.27	0.59	0.77	0.1600
Geovation	0.83	1.14	1.24	1.25	1.09	1.29	0.69	0.58	$3.9*10^{-3}$
IT	0.58	1.17	1.02	1.85	0.84	0.97	0.59	0.95	0.3356
Mechanical engineering	0.84	1.42	1.45	1.52	1.37	1.34	0.45	0.39	$1.3*10^{-29}$
Mechanics	0.91	1.36	1.32	1.25	1.06	1.17	0.62	0.58	$5.3*10^{-8}$
Physics	0.96	1.07	1.02	1.27	1.2	1.04	0.76	0.76	$1.3*10^{-2}$
Radiophysics	0.22	3.78	1.23	1.12	1.43	1.05	1.92	0.07	$2.9*10^{-18}$
Tornado	0.88	1.25	1.16	1.22	1.08	1.19	0.77	0.58	$2.4*10^{-3}$
Tornado-k40	0.83	1.28	1.4	1.43	0.94	1.17	0.55	0.64	$9.6*10^{-4}$
G2	0.76	1.48	1.41	1.34	1.42	1.4	0.39	0.46	$2.0*10^{-12}$
Cascade	0.74	1.05	1.41	1.36	0.93	1.4	0.65	0.59	$7.4*10^{-3}$
Combined	0.88	1.25	1.19	1.23	1.09	1.21	0.73	0.57	$1.3*10^{-4}$

Table 3

Multipliers and main effects for annual cycle

Flow	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	P-value
Astrophysics	1.53	2.01	1.55	1.66	1.15	1.31	0.65	0.81	1.09	0.56	0.46	0.57	$2.6*10^{-13}$
Bioinformatics	3.04	0.47	0.47	0.38	0.52	2.44	0.61	0.13	2.16	4.95	2.43	1.52	$7.6*10^{-10}$
Biophysics	0.96	0.89	0.86	1.13	1.04	1.97	1.26	0.39	0.84	0.91	1.27	1.22	$2.2*10^{-6}$
Energetics	0.32	0.64	1	1.08	1.67	1.88	1.08	1.91	0.46	1.48	1.11	0.89	$1.5*10^{-21}$
Geophysics	0.61	0.96	0.88	1.75	0.34	0.89	2.55	0.48	0.95	1.43	1.3	1.68	$9.8*10^{-6}$
Geovation	0.5	1.96	1.36	1.35	0.85	1.1	0.41	2.1	1.23	1.05	0.89	0.59	$1.29*10^{-28}$
IT	1.51	4.39	0.91	4.61	1.71	0.08	0.37	0.24	1.8	1.73	1.37	0.68	$4.4*10^{-26}$
Mechanical engineering	1.16	0.93	1.22	1.1	0.69	1.23	0.71	0.66	0.96	1.04	1.06	1.67	$2.8*10^{-13}$
Mechanics	1.28	0.79	2.11	1.17	0.85	0.84	1.11	0.88	0.97	1.03	0.79	0.72	$2.9*10^{-6}$
Physics	0.87	0.98	0.83	0.96	0.76	1.13	1.05	1.11	0.83	0.99	1.78	1	$3.2*10^{-4}$
Radiophysics	5.89	0.24	2.01	2.13	0.14	0.04	0.13	1.21	5.61	2.33	4.5	3.37	$2.3*10^{-26}$
Tornado	0.73	1.72	1.17	1.17	0.73	1.05	0.45	1.87	1.3	1.16	0.91	0.66	$3.5*10^{-23}$
Tornado-k40	1.41	1.23	0.82	0.63	2.21	0.7	0.39	0.56	1.34	1.14	1.65	1.3	$1.2*10^{-6}$
G2	0.66	1.69	1.75	1.74	0.88	1.77	0.83	2.04	0.38	0.3	0.87	1.13	$1.1*10^{-17}$
Cascade	0.28	0.43	1.33	2.48	1.29	0.72	0.41	0.85	1.97	1.76	3.7	0.61	$7.7*10^{-16}$
Combined	0.7	1.62	1.19	1.23	0.78	1.05	0.46	1.75	1.21	1.08	1	0.7	$3.5*10^{-24}$

The P-value is determined as the minimal $\alpha \leq 0.05$, such that all the confidence intervals for the pairwise contrasts of joint significance level $1- \alpha$ that were entirely in the region to the right or to the left of zero still remain in the same region. The obtained significant orders of the main effects can be visualized as a graph. The nodes of the full graph of significant orders are related to the corresponding levels of factor, and the edges are present, if the order (smaller than or larger than) is confirmed statistically at the established level of confidence adjusted to the number of flows and total number of pairwise contrasts. All levels of the factor can be ordered by the value of the estimator, in which case the edge orientation can be omitted. The edges of the reduced graph are arranged in increasing order of the effect level estimators, and the edge between every two nodes (right and left) is present only if every node to the right of the right node and each node to the left of the left node of the pair are connected by an edge at the full graph of significant orders. Nodes that are not informative for the significant orders can be removed. Although the reduced graph is not uniquely defined by the full graph, there is a subjective component in the choice of the reduced graph version, and some significant orders can be missed, the reduced graph seems more practical for interpreting the results of ordering than the full graph.

The results of the advanced analysis for the multiplicative main effects of the weekly and annual periodic factors are presented as the reduced graphs (one for each flow) in Tables 4 and 5, respectively. For example, the intensity of the combined flow in July is significantly smaller than in November, June, October, March, September, April, November and August; the intensity in July, December and January is significantly smaller than in September, April, February and August; the intensity in July, December,

Table 4

Partial orders of the intensities of the job flows within the annual cycle

Flow	Significant partial orders (reduced graph)	P-value
Astrophysics	NOV DEC JUL AUG SEP MAY JUN JAN MAR APR FEB	0.0031
Bioinformatics	APR MAR JAN OCT	0.0017
Biophysics	AUG APR DEC NOV JUN	0.0024
Energetics	JAN SEP FEB DEC MAR APR JUN NOV OCT MAY AUG	0.0026
Geophysics	MAY JUL	$5*10^{-4}$
Geovation	JUL JAN DEC MAY NOV OCT JUN SEP APR MAR FEB AUG	$4.5*10^{-4}$
IT	JUN AUG JUL DEC MAR NOV JAN MAY OCT SEP FEB APR	0.0028
Mechanical engineering	MAY JUL MAR JUN DEC	$6.2*10^{-5}$
Mechanics	DEC NOV FEB MAR	$6.8*10^{-4}$
Physics	NOV MAY	0.0011
Radiophysics	JUN JUL MAY FEB AUG MAR APR DEC NOV SEP JAN	0.0015
Tornado	JUL DEC JAN MAY NOV JUN OCT APR MAR SEP FEB AUG	0.0030
G2	OCT SEP JAN JUL NOV MAY DEC FEB APR MAR AUG	0.0017
Cascade	JAN JUL FEB DEC JUN AUG MAY MAR OCT SEP APR NOV	0.0017
Combined	JUL DEC JAN MAY NOV JUN OCT MAR SEP APR FEB AUG	0.00213

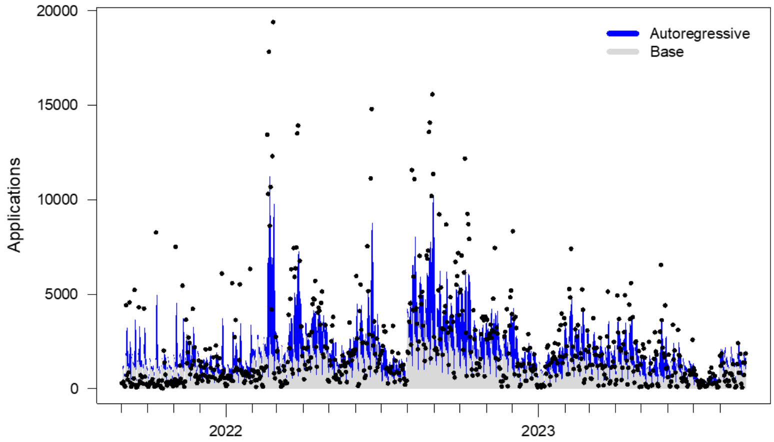

Fig. 5. Observed numbers and Poisson autoregressive predictors

January and May is significantly smaller than in April, February and August; the intensity in July, December, January, May and November is significantly smaller than in August. The intensity of the job flow for users in the bioinformatics area of expertise in April and March is smaller than in January and October (other nodes are omitted). Moreover, for example, the intensity of user job flow in “G2” cluster in December is significantly smaller than in September, but this pairwise order is not marked at the reduced graph in the table, due to the method limitations, since the order of December and September is not significant, whereas the estimator of the intensity in September is smaller than in October.

Table 5

Partial orders of the intensities of the job flows within the annual cycle

Flow	Significant partial orders (reduced graph)	P-value
Mechanical engineering	SUN SAT FRI THU MON TUE WED	$1.9*10^{-8}$
Mechanics	SUN THU FRI WED TUE	0.0032
Radiophisics	SUN TUE THU SAT MON	$1.2*10^{-4}$
G2	SAT SUN WED FRI TUE THU MON	$9.5*10^{-4}$
Combined	SUN TUE WED MON	0.0021

Next, the univariate Poisson autoregressive models (1) was fitted, where the parameterization for λ_t is determined by (2), $\log v_{\text{it}} = \alpha$ for all t, and $\alpha > 0$ is the parameter of autoregression, for each user job flow separately. In order to fit the models, the R-function *hhh4*() of package *surveillance* was used. The obtained estimators of the base and autoregressive components of the combined flow are visualized in Fig. 5.

The stratified statistical analysis of user job flows from different areas of expertise and computing clusters showed the significance of annual and weekly periodic factors for each flow adjusted to the number of flows (the maximal P-value of the likelihood ratio test $8.1*10^{-4}$ was obtained for weekly periodic factor of user job flow in radiophysics) and the regression component is formally significant for all job flows, with the exception of the user job flow with the radiophysics area of expertise. In conclusion, it should be noted, that the estimators of the base component λ_t in the Poisson autoregressive model do not determine the intensity changes due to the presence of the autoregressive component.

Discussion

All the jobs were initially classified by the user's area of expertise and by the computing cluster, to which the job was submitted. Considering the results of the explanatory analysis, a stratified approach to study the user job flow was applied. Based on the number of jobs per day for each group of jobs, a time series was generated.

Two approaches were used for stratified analysis of user job flows: the generalized linear model and generalized estimating equation (GEE) based on pseudo-likelihood function, and the Poisson autoregressive model. The GEE analysis revealed significant difference in the intensities in different month of the year for each of user job flows, but no implicit seasonal changes were found, nor did it reveal a common form of the intensity changes for all the job flows. Advanced statistical analysis allowed to reveal some significant partial orders of month by the intensity values for each of user job flow. The statistically significant difference in the intensities of job flows on different days of the week were found for only a part of the flows: mechanical engineering, mechanics and radiophysics, as well as the "G2" cluster, and the combined flow. For each of these five flows, some partial orders of days of the week in terms of intensity values were obtained.

The Poisson autoregressive analysis showed significantly lower variance of the regression and autoregression parameters estimators, which indicated greater stability of the model compared to GEE. The statistical significance of weekly and annual periodic factors of the base component were detected for each of the user job flows. The statistical significance of the autoregressive component was detected for each of the user job flows, excluding users in radiophysics area of expertise. The statistical significance of the autoregressive component can be explained both by the dependence of observations and overdispersion and indicates the inexpediency of using the Poisson generalized linear model, when the observations are independent.

REFERENCES

1. **Baranov A.V., Nikolaev D.S.** Machine learning to predict the supercomputer jobs execution time. Software & Systems, 2020, Vol. 33, pp. 218–228. DOI: 10.15827/0236-235X.130.218-228

2. **Diggle P.J., Heagerty P., Liang K.-Y., Zeger S.L.** Longitudinal data analysis. 2nd edition. Oxford: Oxford University Press Inc., 2002.

3. **Fokianos K., Rahbek A., Tjøstheim D.** Poisson autoregression. Journal of the American Statistical Association. 2009, Vol. 104, no. 488, pp. 1430–1439. DOI: 10.1198/jasa.2009.tm08270

4. **Held L., Höhle M., Hofmann M.** A statistical framework for the analysis of multivariate infectious disease surveillance counts. Statistical Modelling, 2005, Vol. 5, no 3, pp. 187–199. DOI: 10.1191/1471082X05st098o

5. **Höhle M., Meyer S., Paul M.** surveillance: Temporal and Spatio-Temporal Modeling and Monitoring of Epidemic Phenomena. 2022, Available: https://CRAN.R-project.org/package=surveillance (Accessed 25.09.2024)

6. **Klinkenberg J., Terboven C., Lankes S., Müller M.S.** Data Mining-Based Analysis of HPC Center Operations. 2017 IEEE International Conference on Cluster Computing (CLUSTER), 2017, pp. 766–773. DOI: 10.1109/CLUSTER.2017.23

7. **Liang K.-Y., Zeger S.L.** Longitudinal data analysis using generalized linear models. Biometrika, 1986, Vol. 73, no 1, pp. 13–22. DOI: 10.1093/biomet/73.1.13

8. **McCullagh P., Nelder J.A.** Generalized linear models. 2nd edition. London: Chapman & Hall, 1989.

9. **Meyer S., Held L., Höhle M.** Spatio-temporal analysis of epidemic phenomena using the R package surveillance. Journal of Statistical Software, 2017, Vol. 77, no. 11, pp. 1–55. DOI: 10.18637/jss.v077.i11

10. **Zaborovsky V.S., Utkin L.V., Muliukha V.A., Lukashin A.A.** Improving Efficiency of Hybrid HPC Systems Using a Multi-agent Scheduler and Machine Learning Methods. Supercomputing Frontiers and Innovations, 2023, Vol. 10, no. 2, pp. 104–126. DOI: 10.14529/jsfi230207

11. **Zeger S.L., Liang K.-Y., Albert P.S.** Models for longitudinal data: A generalized estimating equation approach. Biometrics, 1988, Vol. 44, no. 4, pp. 1049–1060.

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Malov Sergey V.
Малов Сергей Васильевич
E-mail: sergey.v.malov@gmail.com
ORCID: https://orcid.org/0000-0003-0093-6506

Lukashin Alexey A.
Лукашин Алексей Андреевич
E-mail: lukash.spb.ru@gmail.com

Submitted: 25.06.2024; Approved: 17.09.2024; Accepted: 27.09.2024.

Поступила: 25.06.2024; Одобрена: 17.09.2024; Принята: 27.09.2024.