Asian Journal of Innovation and Policy (2024) 13.1:057-067
DOI: https://doi.org/10.7545/ajip.2024.13.1.057

Factors Influencing Supercomputing Resource Selection with PCA

Hyungwook Shim[*], Myungju Ko[**], Sunyoung Hwang[***], Jaegyoon Hahm[****]

Abstract This paper analyzes the factors influencing the selection of supercomputing resources. Using the results of a survey targeting supercomputing resources in the public sector, a resource selection model was presented through logistic regression and principal component analysis methods. As a result of the analysis, it was confirmed that affiliation, purpose of use, size of research funding, possession of a supercomputer, and whether specialized services are needed have a significant impact on resource selection. In the future, we expect that the results of this study will be used in various ways to manage demand for supercomputing resources.

Keywords Supercomputer, Logistic regression, Principal component analysis, Demand management, Resource selection model

I. Introduction

Supercomputing resources are public goods that are closely related to national scientific and technological capabilities. As various applications using AI are commercialized, large-scale calculations with big data have become essential in most industries. From AI-based government administrative systems to public life services such as transportation, finance, shopping, and media, etc. continue to emerge (Gill, 2022; Shankar, 2022; Arora, 2021). In the near future, similar

Submitted, February 21, 2024; 1st Revised, April 19, 2024; 2nd Revised, May 28, 2024; Accepted, July 4, 2024

* Ph.D, Division of National Supercomputing Center, Korea Institute of Science and Technology Information, Daejeon, Korea; shw@kisti.re.kr
** Ph.D, Division of National Supercomputing Center, Korea Institute of Science and Technology Information, Daejeon, Korea; myju@kisti.re.kr
*** Researcher, Division of National Supercomputing Center, Korea Institute of Science and Technology Information, Daejeon, Korea; sunyoung@kisti.re.kr
**** Corresponding, Center Director, Division of National Supercomputing Center, Korea Institute of Science and Technology Information, Daejeon, Korea; jaehahm@kisti.re.kr

to the electricity that is irreplaceable in our daily lives, supercomputing resources will become a critical resource needed 24 hours a day in all industries. However, supercomputing resources are limited. In particular, public sector research relies on state-run supercomputing resources because it is unavailable to use resources owned by private companies. Therefore, the government's perspective, the distribution of these resources and demand management have become very important to increase national competitiveness.

A supercomputer is a type of computer that quickly and accurately calculates large-scale, complex computational tasks using high-performance computing resources. Recently, 500 computers announced based on performance by "TOP500," a world supercomputer rankin site, are considered supercomputers. Many countries are making efforts to secure supercomputing resources. New resources enter the ranking every year, and all countries in the top 10 appear to have more than 10 supercomputers. In other words, many countries have a large number of supercomputer resources, and as a result, the importance of resource management is increasing along with the expansion of resources (Shim, 2023). When starting to operate multiple public sector supercomputing resources, the distribution of demand between existing resources and newly built resources is very important. From a national perspective, reducing the surplus time of resources and increasing the utilization rate is the basis for increasing technological competitiveness and fostering industry. In order to optimally distribute demand, it is necessary to identify trends in what types of resources users want to select. For this trend, it is needed to find out the factors that influence the selection. Since it is difficult to survey all users who plan to use resources in the future, a statistical estimation process using samples must be conducted. Therefore, this paper surveyed existing users of public sector supercomputing resources in South Korea to estimate the factors affecting resource selection and the size of the influence.

II. Introduction

To date, many studies have been conducted to discover factors affecting people's behavioral choices and to predict the future or suggest implications. In addition, these behaviors and phenomena are modeled, and the coefficients of the model are estimated using various statistical methods to analyze the quantitative size of the influence of each factor. The major studies related to this are as follows. Rozell (1999) used a path analysis of longitudinal data collected from 75 manufacturing employees participating in a computer training course, to test a model of the intrapersonal processes impacting computer-related performance. Gender, computer experience, and attributional style were found

to be predictive of computer attitudes, which were in turn related to computer efficacy, task-specific performance expectations, and post-performance anxiety (Rozell, 1999). Amron (2019) aims to review and identify the relevant factors that influence the acceptance of CC(Cloud Computing) implementation in the organization. It reviewed 55 articles related to CC implementation, and a total of 21 factors have been obtained through several processes. These factors were arranged according to the frequency based on the thematic analysis method (Amron, 2019). Lin (2017) examines the impact of GDP per capita (gross domestic product), energy intensity (EI), carbon intensity (CI), and total population on carbon dioxide emissions in China's transport industry using quantile analysis from 1980 to 2010. In this study, having confirmed stationarity and that there exists a long-term relationship among our variables (carbon emission, gross domestic product, energy intensity, carbon intensity, and urbanization), we checked which variable(s) has a greater impact on carbon emission on different quantiles (Lin, 2017). Wen (2020) establishes a model for the factors affecting the acceptance of online education platforms among college students based on the theory of planned behavior (TPB) and puts forward several hypotheses on the influence of multiple factors on acceptance. Then, a scientific questionnaire was designed and distributed online to college students. The survey data were subject to descriptive analysis and correlation analysis(Wen, 2020). Sisman (2022) aims to determine the potential influencing factors of housing prices through applying global regression models including Ordinary Least Squares (OLS), Spatial Lag Model (SLM), and Spatial Error Model (SEM) and examine their geographic variation by local regression approaches such as Geographically Weighted Regression (GWR) and Multiscale Geographically
Weighted Regression (MGWR). Pendik district of Istanbul (Turkey) was selected as the study are. Souza(2018) intends to identify the main factors that influence the durability of adhesive ceramic external wall claddings, developing models to predict their service life that involve these variables. The service life prediction models proposed in this study are defined based on simple and multiple regression analysis, both linear and nonlinear. The proposed models are defined based on the evaluation of the degradation condition of 96 ceramic claddings of residential buildings in Brasília (Souza, 2018).

Looking at previous research, research to find factors that influence people's decision-making continues to be conducted in many fields. They used various methodologies, including surveys, path analysis, and regression analysis, to discover influential factors that do not appear on the surface, not only in computer science, but also in education, economics, and architecture. We aim to derive factors that influence the selection of supercomputing computational resources that have not been addressed so far in the supercomputing field. PCA was used as a new model improvement tool derived through logistic regression

analysis. It would be used as an important reference to manage demand for supercomputing resources.

This paper consists of a total of 5 chapters. In Chapters 1 and 2, the academic value of this study is derived through the background of the study and analysis of previous research. Chapter 3 describes the research procedures and data. Chapter 4 presents the analysis process and results for deriving influencing factors, and Chapter 5 summarizes the results and presents the limitations and plans of the study.

Ⅲ. Research procedures and data

The procedure for analyzing factors affecting supercomputer resource selection is shown in Fig 1. We survey to obtain data and use it to perform logistic regression analysis. Through this, model suitability is confirmed and a regression model is derived using variables that have a significant impact. Next, the most critical step is to derive the appropriateness of variable selection and improvement factors through principal component analysis. Finally, we re-run the regression analysis on the improved model and examine whether it has improved.

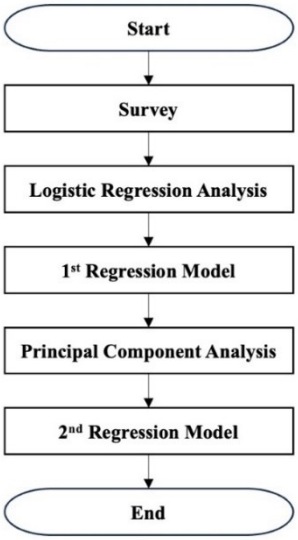

Figure 1. Flowchart

The selection process for these items is explained in detail as follows. First, it was divided into two areas: users and resource providers. User survey items are the basic unique characteristics of resource use and respond to the purpose of use, field of use, etc. Resource provider items included specialized services that users could select according to their preferences. Next, survey items were added using the operation plan submitted by the supercomputing center to the government. Lastly, to supplement the items, a review of the items was conducted targeting existing supercomputer users.

The survey was conducted on users with experience using supercomputers. A total of 200 people were surveyed, and an online survey was conducted.

The survey items were selected as 7 items that were judged to have an impact on users' resource selection and an item of results of resource selection. Affiliation, field of use, purpose of use, size of research funding, possession of a supercomputer, recognition of field-specific resources, and need for specialized services are presented. Respondents must select one of two options for each item. The model for logistic regression analysis is shown in equation (1). x_n and y represent independent and dependent variables, α is a constant, and β is a coefficient.

$$\mathrm{y} = \alpha + \beta_a x_a + \cdots + \beta_f x_f \tag{1}$$

The criteria and basis for variable selection were based on existing research results. Shim (2023) selected variables that could reflect the characteristics of the supercomputer ecosystem consisting of national centers and specialized centers by referring to the government's operating guidelines and annual operating plans. The selection results and detailed description of the variables are as follows. First, a refers to the user's affiliation. They choose between private companies and public institutions, and through this, it will be verified which resources they prefer depending on the type of affiliation. b is a variable that indicates whether the field to be used is included in the core field (autonomous driving, weather, disaster, etc.) or not, and the demand for the core field can be used to make the political priority of the investing field. c indicates the purpose of use and, d is the size of the research fund supported to use the resource. It intended to see the characteristic that the larger the research fund, the less affected by the usage fee. e was selected to see the tendency to use national resources despite having a supercomputer, and f indicates whether the different types of resources (herein referred to as specialized center) are known or not. It gives the user computing service by field (specialized service). Lastly, g refers to the use of customized services specialized in the research field. h indicates resources a user can select. National resources represent established supercomputers; special resources represent the resources that provide specialized service by fields and introduced newly. Seven variables, including

affiliation, field of use, purpose of use, size of research funding, possession of a supercomputer, and recognition of a specialized center, were used as independent variables, and the final eighth user resource was used as a dependent variable. The survey results were coded as shown in Table 1.

Table 1. Data and Code

			Code	
			0	**1**
Independent Variables	a	**Affiliation**	Private	Public
	b	**Field**	Core	Others
	c	**Purpose**	Public R&D	Private R&D
	d	**Fund Size**	Under $10,000	Over $10,000
	e	**Supercomputer Possession**	X	O
	f	**Recognition of Resources**	X	O
	g	**Specialized Service**	X	O
Dependent Variables	h	**Selection**	National Resource	Specialized Resource

Source: The authors

Ⅳ. Analysis Result

The coefficient estimation results are shown in Table 2. First, looking at the p-value, it was determined that the variables a, b, and f were not significant at the 5% significance level. Next, looking at the signs of the coefficients for significant variables, c, d, and g show positive signs, and e shows a negative sign. To interpret this, in the case of c, the closer the purpose of the use is to private R&D, the higher the tendency to select a specialized resource, while in the case of d, the larger the research budget, the more inclined to select a specialized resource. The more specialized services g are needed, the more specialized resources they choose. On the other hand, in the case of e, if you do not have a supercomputer, you tend to choose a specialized resource.

Table 2. Regression analysis results

	B	S.E,	Wald	p-value
a	-.704	.460	2.347	.126
b	-.622	.505	1.513	.219
c	2.561	.702	13.290	.000
d	1.779	.572	9.671	.002
e	-2.370	.614	14.885	.000
f	.639	.455	1.974	.160
g	2.608	.481	29.380	.000
constant	-1.658	.523	10.036	.002

Source: The authors, Summary of statistical analysis results

This can be expressed as a regression model as equation (2).

$$y_g = -1.658 + 0.704x_a - 0.622x_b + 2.561x_c + 1.779x_d - 2.37x_e + 0.639x_f + 2.608x_g \quad (2)$$

The results of re-performing the regression analysis excluding statistically insignificant variables are shown in Tables 3 and 4. The results of the model fit are shown in Table 3. The explanatory power was found to be 24.8% and 34.1%, respectively, through Cox and Snell's R2 and Nagelkerke R2 values.

Table 3. Model fit

Cox and Snell's R^2	Nagelkerke R^2
.248	.341

All variables were significant at the 5% significance level, and the coefficient values were confirmed to have decreased somewhat overall.

Table 4. Regression analysis results

	B	S.E,	Wald	p-value
c	1.870	.560	11.152	.001
d	1.560	.524	8.870	.003
e	-2.209	.523	17.856	.000
g	2.263	.382	35.086	.000
constant	-1.278	.430	8.818	.003

Source: The authors, Summary of statistical analysis results

PCA was additionally performed to confirm the appropriateness of the regression model derived from the regression analysis. PCA uses the statistical properties of the data mean value, data standard deviation value, and covariance of data in each dimension of data set elements. When a data set of related variables is observed, a new method that reflects the original information contained in the variables as much as possible is used. This is a method of creating sets of variables. Therefore, using the component matrix derived through this analysis, we review whether the selection of the excluded variables and the variables included in the current regression model is appropriate, considering significance. To explain in detail, if the correlation with other variables in the component matrix is low and classified as a set with separate components, that variable can be judged to have an independent effect on resource selection and is therefore included in the regression model even if its significance is low. It has to be. Conversely, if multiple variables are grouped into one component, the set of variables with that correlation can be reduced to a single dimension. PCA is conducted in the following steps: KMO (Kaiser-Meyer-Olkin) and Bartlett's tests, confirmation of commonalities, total variance, and component matrix analysis (Liu, 2033; Mansfield, 1977). First, Table 5 shows the test results of KMO and Bartlett. KMO indicates the appropriateness of variable selection, and Bartlett indicates the appropriateness of the model. The KMO was 0.708, which is greater than 0.5, and the Bartlett test also showed that the significance probability was less than 5%, so the data and model can be judged to be at an appropriate level.

Table 5. KMO and Bartlett test

		Value
KMO		.708
Bartlett	**Chi-square(Approximate)**	301.270
	p-value	.000

Source: The authors, Summary of statistical analysis results

According to Table 6, it is divided into four components, and these factors have an explanatory power of approximately 79.3% of the total variance.

Table 6. The Eigenvalue, % of variance and coefficients for each principal component

Component	Eigenvalue	% of variance	% of cumulative variance
1	1.540	22.001	22.001
2	1.481	21.158	43.160
3	1.404	20.061	63.221
4	1.125	16.077	79.298

Source: The authors, Summary of statistical analysis results

The component matrix consisting of a total of 4 components is shown in Table 7. Through factor loading, component 1 can be grouped into g, f, and b, component 2 into d, component 3 into c, e, and component 4 into a.

Table 7. Rotated component matrix (varimax)

	Component			
	1	2	3	4
g	.856	-.011	-.083	.340
f	.693	-.018	.516	-.192
b	.523	.363	.310	.264
d	-.027	-.921	.082	.012
c	.049	.395	.753	.290
e	.110	.068	.887	.223
a	.197	.111	.193	.877

Source: The authors, Summary of statistical analysis results

Through the PCA results, variable a represents an independent component and should be added to the model even if its significance is low. Therefore, the logistic regression analysis was performed again by adding variable a. As a result of the analysis, it can be seen that the model fit is 25.9% and 35.5%, respectively, which is a slight increase compared to the previous model.

Table 8. Model fit

Cox and Snell's R^2	Nagelkerke R^2
.259	.355

Source: The authors, Summary of statistical analysis results

The regression analysis results are shown in Table 9, and the improved regression model is shown in equation (3). The sign of the coefficient did not change, and it was confirmed that the overall value became smaller. The coefficient of a showed a value of -0.716, indicating a tendency to select the specialized resource if it is a private institution.

Table 9. Regression analysis results

	B	S.E,	Wald	p-value
a	-.716	.447	2.569	.109
c	2.095	.598	12.276	.000
d	1.615	.541	8.892	.003
e	-2.057	.534	14.846	.000
g	2.520	.428	34.732	.000
constant	-1.318	.458	8.278	.004

Source: The authors, Summary of statistical analysis results

$$y_g = -1.318x_a + 2.095x_c + 1.615x_d - 2.057x_e + 2.520x_g \quad (3)$$

V. Conclusion

To manage demand for efficient use of public sector supercomputer resources, factors affecting resource selection were analyzed. As a result of the analysis, it was confirmed that affiliation, purpose of use, size of research funding, possession of a supercomputer, and whether specialized services are needed have a significant impact on resource selection. As the number of national supercomputer resources operated increases, it is very important to analyze the factors that influence users' resource selection. As there is not much research related to this, it is believed that the results of this paper can be used as a reference to establish effective resource plans in various countries that are expanding their resources. The limitation is that the survey subjects targeted existing users who had used supercomputers, so it did not reflect the preferences of new users. Therefore, in order to compensate for this limitation in the future, we plan to reanalyze it including prospective users who want to use the resource.

Acknowledgment

This research was supported by the Korea Institute of Science and Technology Information (KISTI).(No. K24L2M1C3).

References

Amron, M.T., Ibrahim, R., Bakar, N.A. A., Chuprat, S. (2019). Determining factors influencing the acceptance of cloud computing implementation. Procedia Computer Science, 161, 1055-1063.

Arora, R. (2021). Toward Efficient Resource Utilization of a GPU-Accelerated AI Supercomputer (Doctoral dissertation, Northeastern University).

Gill, S.S., Xu, M., Ottaviani, C., Patros, P., Bahsoon, R., Shaghaghi, A., Uhlig, S. (2022). AI for next generation computing: Emerging trends and future directions. Internet of Things, 19, 100514.

Lin, B., Benjamin, N.I. (2017). Influencing factors on carbon emissions in China transport industry. A new evidence from quantile regression analysis. Journal of cleaner production, 150, 175-187.

Liu, R.X., Kuang, J., Gong, Q., Hou, X.L. (2003). Principal component regression analysis with SPSS. Computer methods and programs in biomedicine, 71(2), 141-147.

Mansfield, E.R., Webster, J.T., Gunst, R.F. (1977). An analytic variable selection technique for principal component regression. Journal of the Royal Statistical Society: Series C (Applied Statistics), 26(1), 34-40.

Rozell, E.J., Gardner III, W.L. (1999). Computer-related success and failure: a longitudinal field study of the factors influencing computer-related performance. Computers in Human behavior, 15(1), 1-10.

Shankar, S., Reuther, A. (2022, September). Trends in energy estimates for computing in ai/machine learning accelerators, supercomputers, and compute-intensive applications. In 2022 IEEE High Performance Extreme Computing Conference (HPEC), 1-8.

Shim, H., Hahm, J. (2023). Preferences for Supercomputer Resources Using the Logit Model. Journal of information and communication convergence engineering, 21(4), 261-267.

Shim, H., Hahm, J. (2023). A study on demand management plans for National Supercomputer resources. Technology in Society, 75, 102376.

Sisman, S., Aydinoglu, A.C. (2022). A modelling approach with geographically weighted regression methods for determining geographic variation and influencing factors in housing price: A case in Istanbul. Land Use Policy, 119, 106183.

Souza, J., Silva, A., de Brito, J., Bauer, E. (2018). Analysis of the influencing factors of external wall ceramic claddings' service life using regression techniques. Engineering Failure Analysis, 83, 141-155.

Wen, J., Wei, X., He, T., Zhang, S. (2020). Regression Analysis on the Influencing Factors of the Acceptance of Online Education Platform among College Students. Ingénierie des Systèmes d'Information, 25(5).