대한경영정보학회
「경영과 정보연구」 제39권 제1호 2020년 3월
http://dx.doi.org/10.29214/damis.2020.39.1.003

Verifying the Classification Accuracy for Korea's Standardized Classification System of Research F&E by using LDA(Linear Discriminant Analysis)

Seokin Joung* · Yeongwha Sawng** · Euhduck Jeong***

〈요 약〉

Recently, research F&E(Facilities and Equipment) have become very important as tools and means to lead the development of science and technology. The government has been continuously expanding investment budgets for R&D and research F&E, and the need for efficient operation and systematic management of research F&E built up nationwide has increased. In December 2010, The government developed and completed a standardized classification system for national research F&E. However, accuracy and trust of information classification are suspected because information is collected by a method in which a user(researcher) directly selects and registers a classification code in NTIS.

Therefore, in the study, we analyzed linearly using linear discriminant analysis(LDA) and analysis of variance(ANOVA), to measure the classification accuracy for the standardized classification system(8 major-classes, 54 sub-classes, 410 small-classes) of the national research facilities and equipment established in 2010, and revised in 2015. For the analysis, we collected and used the information data(50,271 cases) cumulatively registered in NTIS(National Science and Technology Service) for the past 10 years. This is the first case of scientifically verifying the standardized classification system of the national research facilities and equipment, which is based on information of similar classification systems and a few expert reviews in the in-outside of the country.

As a result of this study, the discriminant accuracy of major-classes organized hierarchically by sub-classes and small-classes was 92.2 %, which was very high. However, in post hoc verification through analysis of variance, the discrimination power of two classes out of eight major-classes was rather low. It is expected that the standardized classification system of the national research facilities and equipment will be improved through this study.

Key Words: Research Facilities and Equipment, Standardized Classification, Discriminant Analysis, LDA, Classification Verification

논문접수일: 2020년 02월 20일 수정일: 2020년 03월 03일 게재확정일: 2020년 03월 04일

* First Author, Team Leader, National Research Facilities & Equipment Center, KBSI, jsi494@kbsi.re.kr
** Corresponding Author, Professor, Department of Management of Technology, Konkuk University, sawng@konkuk.ac.kr
*** Co Author, Director of the Center, National Research Facilities & Equipment Center, KBSI, edjeong@kbsi.re.kr

Ⅰ. Introduction

Throughout history, science and technology have contributed significantly to national economic growth and industrial development. During the past six decades after the Korean War, Korea's science and technology, and the national economy have continued to develop together. Recently, research facilities and equipment (F&E) have become especially important as tools and as a means to lead the development of science and technology.

As a result, the government has been continuously expanding investment budgets for research and development (R&D) and research F&E, and the need for efficient operation and systematic management of research F&E to be built nationwide has increased. The government began to establish a full-time management system for national research F&E in September 2009. As a part of this, in December 2010, the government developed and completed a standardized classification system for national research F&E, including scientific equipment and research facilities in the existing industrial technology equipment category. Additionally, in February 2015, the existing system (sub-/small classes) was reorganized with the increase of convergence equipment and new high-tech equipment, which is now useful as a standard of management for F&E.

However, the accuracy of and trust in information classification are suspect because information is collected by a method in which a user (researcher) directly selects and registers a classification code in NTIS. The cause of the error that occurs when a user (researcher) registers the research F&E information in NTIS is considered to be a ambiguity of the classification criteria or a violation of MECE principles (mutually exclusive and collectively exhaustive) as a categorized thinking tool proposed by McKinsey consultants (Lee and Chen, 2018). This hinders the judgment when the user selects the classification code. On the contrary, despite accurate user registrations, the problems may arise due to standard error or a quantitative imbalance of the standardized classification system itself.

Therefore, the study analyzed linearly using linear discriminant analysis (LDA) and analysis of variance (ANOVA) from the information data (total 50,271 cases), cumulatively registered in NTIS for the past 10 years (2006-2015), to verify the classification accuracy for the standardized classification system of the national research F&E.

Ⅱ. Literature Review

1. Literature Review on Technology Classification

Technology classification is used in numerous countries around the world, including the US and European nations, to systematically manage the country's science and technology activities and budget allocation. Technology classification can be divided into classification

systems according to technology content, based on the principles of the technology and academic disciplines, and classification systems based on the purpose of the technology in terms of need (Shin et al., 1994; Jeong et al., 1994). Technology classification is currently classified into two types: flat classification systems and binary classification systems (Seol and Song, 2000; Lee et al., 2016). Unlike flat classification, a traditional classification method, binary classification is a matrix-type classification system that separates research and application sectors. Flat classification systems have the advantages of being simple to create and use, while binary classification systems can express technology convergence and emphasize new areas of technology. Furthermore, the vertical and horizontal structures should be considered when developing the classification system. In a vertical structure, the classification accuracy increases as the number of classes increase; in terms of actual utilization, however, the number of classes is not proportional to utilization efficiency. The horizontal structure refers to the relationship between classification items of the same class. This is closely related to the principles of the classification system, in which the system also considers importance, scale, and the like, in addition to the source of the sector when determining which class to classify a particular sector, i.e., whether to place it in a major class or a sub-class.

The following three technology classification systems are the most representative cases in Korea. Shin et al. (1994) used the expert panel method to form seven main sectors: information/electronics/communication, machinery/equipment, materials/processing, life science, energy/resources/nuclear energy, environment/earth science, and construction technology; and eight major classes, including "other," with 42 sub-classes, 204 small classes, and 1,045 detailed classes, thus forming a seven-tiered classification system. This system is a simple list of technical items and has the advantage of being suitable for research and development. As this technology classification was developed in consideration of the characteristics of each field by encompassing element technology as much as possible, the principles and standards of technology classification differ among fields. Moreover, though the classification system is specific and detailed, it is limited to technology of the pure science sectors. However, it is significant in that it is the first classification system created for planning and managing national research and development projects.

The classification table of Cho et al. (2001) was a modified and supplemented version of the technology classification table of Jeong et al. (1994); it was created in a preliminary study to establish the national science and technology standardized classification. As in 1994, the system was created as a flat technology classification system using the expert panel method. It is a seven-tiered classification system consisting of seven major classes, 35 sub-classes, and 183 small classes. Considering the development of technology, the scale of national R&D projects, and the efficiency of utilization, a three-tiered classification

system was created comprising 12 major classes, 94 sub-classes, and 573 small classes. In particular, this is the first classification system to include technology innovation and science and technology policy in the sector of science and technology. In addition, Hong et al. (2016) researched an appropriate technology classification methodology, emphasizing the information utilization aspects of science and technology as a valuable basis for various science and technology policy activities and policy decision-making. Yoo et al. (2018) utilized the 2016 national R&D project survey analysis data to improve the National Science and Technology (S&T) Standard Classification System, revised in 2013. When the temporary classification was newly established as the regular classification, the range and degree of influence were statistically predicted to predict changes in the existing classification. As such, the direction of technology classification research has recently changed toward methodological discussions of technology classification considering the effectiveness of actual information, or statistical analysis of the characteristics of accumulated technology information by classification, rather than discussion of principles on classification, standards, and the like.

2. Case studies of domestic and international standardized classification systems

Domestic standardized classifications are based on international standard classifications, and are essential to enhance the reliability of statistics and expand their usability through quality assurance. Statistics Korea creates and publishes standardized classifications to enable statistics institutions to prepare statistics in the economic and social sectors based on the same standards. Among six standardized classifications, the industrial, expenditure by purpose, and education classifications can fully reflect the economic and social functions and objectives of national R&D projects. Korea Standard Industrial Classification systematically classifies the types of industrial activities mainly performed by production units, such as business units or enterprise units(Sung, 2010). The purpose of the standardized industrial classification is to provide accuracy and comparability in the collection, reporting, and analysis of statistical data on industrial activities. In the standard industrial classification, R&D activities are classified as specialized scientific and technology service industries, as they do not provide services consumed in the normal production process. The activities of national R&D projects themselves are thus classified into one of the major classes of specialized scientific and technology service industries; however, national R&D projects can be divided by classification standard based on the inputs and outputs of economic activities related to the project. KCEAP is a classification system for transactions performed by households, the government, nonprofit organizations, and producers. Its purpose is to be used for the

creation of a system of national accounts (SNA) and as classification items for household expenditure surveys and consumer price surveys. As national R&D projects are promoted by the Government Budget Appropriations for R&D (GBARD), they can be classified by government function classification. The Frascati Manual (OECD, 2015) stipulates that government R&D budgets are classified by socioeconomic objectives for investigation, though they can also be investigated by government function classification. The Korea Standard Classification of Education is a classification system for collecting, aggregating, and analyzing educational data based on the International Standard Classification of Education (ISCED-1997). The classification system consists of level classifications that classify the educational programs, as well as the individual's education level, status of completion, and further education, in addition to field classifications that classify the characteristics of the study according to the curriculum and subject matter. The education-level classification is a classification system divided into stages according to early childhood curricula and curricula from elementary to graduate school.

In addition to the standardized classification system of the economic and social sectors announced by Statistics Korea under the Statistics Act, with the purpose of supporting and managing technology development and research projects, the classification systems used to classify technology and research fields include the Academic Research Area Classification, the Industrial Technology Classification, and the National S&T Standard Classification. The Academic Research Area Classification is a classification system used by the National Research Foundation of Korea to efficiently implement and manage academic research support projects. The current Academic Research Area Classification table is used for the management and statistics of academic research support, investigation of the current status of research activities of universities, the receipt and examination of research projects, and the selection of evaluators. This classification table is designed based on the analysis of major fields of study of domestic and international universities and graduate schools, the names of majors, and curricula, as well as the comparison of items in the various academic classification tables. The Industrial Technology Classification is a classification system stipulated in the Common Operational Guidelines for Industrial Technology Innovation Projects to efficiently plan, assess, and manage industrial technology innovation projects. It was established by the Ministry of Trade, Industry, and Energy, and consists of the major classes of machinery/materials, electricity/electronics, information and communication, chemical, bio/medical, energy/resources, and knowledge services. The National S&T Standard Classification was first established in 2002 based on Article 27 of the Framework Act on Science and Technology and Article 41 of the Enforcement Decree, and has since been periodically revised and supplemented. The

National S&T Standard Classification Systemseeks to efficiently manage information, human resources, R&D projects, and the like related to science and technology, and is utilized as the basic system for the planning, evaluation, and management of national science and technology, the classification standards of science-related surveys, and the management and distribution of knowledge and information. The S&T standard classification is divided into three science and technology sectors and three humanities and society sectors (six sectors, 33 major classes, 369 sub-classes, and 2,899 small classes), in addition to public and industrial sectors to which research results can be applied (two sectors, 33 major classes, and no sub-classes or small classes).

The Frascati Manual (OECD, 2015) developed a field of research and development (FORD) that focuses on the content of R&D with the objective of measuring R&D. This classification system was derived based on the knowledge sources of R&D activities, R&D stakeholders, methods, and techniques, and the experience, knowledge level, and application sectors of scientists and R&D workers. This classification system is consistent with UNESCO's recommendation concerning the International Standardization of Statistics on Science and Technology (1978), which is an important basis for the OECD S&T classification system used in previous Frascati revisions. In R&D areas, the methods for conducting R&D are constantly changing, with continuous changes as new areas are being created. The OECD Frascati Manual recommends that GBARD be classified by socioeconomic objectives (SEO). While it may be difficult to identify the content and objectives of R&D for classification, the SEO classification is recommended because classification by objective is more fundamental from the perspective of government policy objectives for R&D. In addition, as information on the main objective of R&D is difficult to obtain from the party performing the project, classification information is collected through budget data. Moreover, even if government-funded R&D programs have multiple objectives, the classification is based on the identification and classification of the main objectives of the national R&D budget. Eurostat participated in creating the OECD classification as part of the NESTI Task Force, which spearheads the OECD's Frascati Manual. Subsequently, the classification system was revised based on OECD guidelines. For the classification of research fields, the field of study (FOS) of the OECD Frascati Manual was applied. The Nomenclature for the Analysis and Comparison of Scientific Programmes and Budgets (NABS), an independent system used by GBOARD and the Statistical Classification of Economic Activities in the European Community, was used. NABS, originally devised and published in 1969, revised in 1975, 1983, and 1992, and to improve connectivity with the OECD, it was recently revised in 2007 to comprise 13 chapters and 12 sub-chapters.

With the signing of a closer economic relationship between Australia and New

Zealand in 1983, a common classification to facilitate the comparison of R&D statistics between the two nations and countries around the world became necessary. Consequently, the Australian and New Zealand Standard Industry Code (ANZSIC) was enacted in 1993, revised in 2006, and is currently being used. In 2008, the two countries enacted the Australia and New Zealand Standard Research Classification (ANZSRC) to be used jointly. Leading institutions and universities in both countries, centered around the Australian Bureau of Statistics (ABS) and Statistics New Zealand (StatisticsNZ), participated in the establishment of a joint classification system. Through the extensive consultation of technology subcommittees, ANZSRC strengthened its effectiveness so the system could be smoothly adopted and used as a standard classification for both countries even in research activities, in addition to the accumulation of R&D statistical information. ANZSRC replaced the Australian Standard Research Classification (ASRC), which was commonly used in Australia. In New Zealand as well, institutions switched to ANZSRC from their own classifications. Particular features of ANZSRC include its enhancement primarily of the energy, environment, and environmental science-related fields, reflecting the emergence of new R&D fields, increased policy interest in energy production, climate change, and environmental protection, and gradually increasing research interests and achievements. The ANZSRC was created with reference to the OECD Frascati Manual (2002) and NABS (2007) based on ASRC. Like the classification of ASRC, it joins three classification systems (research activities/fields/objectives) to form a three-dimensional matrix structure. Type of activity (TOA) is a classification of R&D activities according to the research results, with four types: pure basic research, strategic basic research, applied research, and experimental development. Basic research is experimental/theoretical research performed to acquire new knowledge for the advancement of knowledge, while strategic basic research is experimental/theoretical research for solving real-life problems. Applied research seeks to obtain new knowledge that can utilize the results of previous or basic research, while experimental development seeks to produce new materials, products, policies, and the like using knowledge derived from research and practical experience, or to essentially improve what has already been created. The Field of Research (FOR) classification is a classification of R&D activities based on the research field, primarily dividing the R&D methodology. The categories of each classification include new research areas, as well as major research areas performed by national research institutes and organizations. In total, it consists of 22 major classes, 157 sub-classes, and 1,238 small classes. The SEO system classifies according to the researcher's perception of the R&D objective or output, and is divided into economic, social, technical, or scientific areas according to the main objective of R&D. The attributes applied in the SEO classification system consist of a combination of processes,

products, health, education, and other specific areas of social and environmental aspects, with five sectors, 17 divisions, 119 groups, and 847 objectives.

3. The Standardized Classification System for National Research Facilities and Equipment

3.1 The Standardized Classification System's Establishment and Revision

Thanks to the economic growth spurred by science and technology, and national development, the government has invested extensive amounts of the R&D budget not only for science and technology R&D but also for the construction and operation of national research F&E. As a result, with the increasing importance of efficient investment and systematic management of nationwide research F&E, in September 2009, the government began to establish a national research F&E life cycle management system. Particularly, as a preliminary task to investigate and analyze the state of construction and operation of national research F&E, the classification of information on national research F&E and the standardization of laws and systems were urgently required. Thus, in December 2010, under the four classification principles (inclusiveness, connectivity, usability, and reliability), Korea's first national research F&E standardized classification system (eight major classes, 48 sub-classes, and 209 small classes) applying the decimal system to sub-classes and small classes was developed (NFEC, 2015).

Since then, the increase of convergence equipment and new high-tech equipment through the development of science and technology, and the spread of digital convergence has exceeded the scope of existing classification systems, thus accelerating the proportion of unclassified research F&E. This was a cause for decreased utilization as a national statistics and research F&E management tool. To address these problems, in March 2014, the government began to revise the first standardized classification. Thus, the government developed a new classification by analyzing the characteristics of 18,990 pieces of equipment classified as "other" among research F&E registered in NTIS and promoted integration between classifications and deletion. The government launched a standardization committee with eight major classes and 23 experts, verified the first standardized classification amendment, and in February 2015, established the national research F&E classification system (eight major classes, 54 sub-classes, and 410 small classes) (NFEC, 2015).

There were three major amendments to the first standardized classification system: the abolition of the decimal system, the increase of sub-classes and small classes, and the introduction of a two-dimensional classification system. First, the existing rules that limited the number of sub-classes and small classes to 10 or less were removed, and the classification system was reorganized to consider only the characteristics of the research F&E without

limiting the number of classifications. Second, terminologies were revised, unnecessary classifications were eliminated, classifications were integrated, and research F&E classified as "other" were added to new classifications (six sub-classes and 201 small classes). Third, the system was reconstructed into a two-dimensional classification system, and considered the uses and technology fields used in R&D of research F&E, as well as perfect matching with key investment areas.

Thus, the government systematically registered the information of national research F&E, and standardized the classification system for managing and using it. Since its establishment in December 2010, the government has strived to reflect the opinions of researchers through continuous system improvement (first revision in 2015).

<Table 1> Major/Sub Class in The Standardized Classification System(2006~2015)

Major class(8)	Sub-class(54)	Case	Major class(8)	Sub-class(54)	Case
1. Optical Equipment	1. The rest of optical equipment	262		28. Power generation equipment	596
	2. Telescope	23		29. Magnetic force generating/ measuring equipment	127
	3. Microscope	2,135		30. Modification equipment	96
				Sub-total	6,500
	4. Camera and visual equipment	1,410	5. Data Process Equipment	31. The rest of data process equipment	72
	5. Equipment creating and measuring light wave	1,874		32. Hardware	3,276
	6. Radiation generating/ measuring equipment	594		33. Software	1,696
				Sub-total	5,044
	7. Image analysis equipment	430	6. Physical State Measurement Equipment	34. The rest of physical state measurement equipment	251
	Sub-total	6,728			
2. Compound Analysis Equipment	8. The rest of compound analysis equipment	311		35. Temperature/heat/humidity/ moisture measurement equipment	1,011
	9. Reaction/compound/Mill equipment	1,719		36. Length/position measuring equipment	874
	10. Bio manufacturing/analysis Equipment	2,119		37. Time/frequency/speed/revolution measurement equipment	119
	11. Separation and purification equipment	1,252		38. Mass, weight/volume/density measurement equipment	116
	12. Separation analysis equipment	2,416		39. Force/torque/pressure/vacuum measurement equipment	607
	13. Spectroscopic equipment	2,353		40. Sound/noise/vibration/shock measurement equipment	500
	14. Mass analysis equipment	1,090		41. Fluid flux dynamics measurement equipment	576

	15. particle analysis equipment	607		42. Surface characteristic measurement equipment	276
	Sub-total	11,867		Sub-total	4,330
3. Process and Test for Machines Equipment	16. The rest of process and test for machines equipment	540	7. Clinical Medical Equipment	43. The rest of clinical medical equipment	48
	17. Cutting equipment	818		44. Clinical diagnosis video equipment	166
	18. Molding equipment	2,497		45. Bio measurement diagnosis equipment	329
	19. Automation and transfer equipment	952		46. Clinical diagnosis analysis equipment	66
	20. Textile machine equipment	403		47. Special equipment for professional medicine	346
				Sub-total	955
	21. Semiconductor equipment	2,896	8. Environment Creation/ Production/ Breeding Facility	48. The rest of environment creation/ production/ breeding facility	135
	22. Heat fluid equipment	3,090		49. Environment creation facility	436
	23. Material characteristics test equipment	2,712		50. Movable facility	115
	Sub-total	13,908			
4. Electricity and Electron Measurement Equipment	24. The rest of electricity and electron measurement equipment	303		51. Biological breeding/experimental facility	167
	25. Measurement equipment	3,434		52. Production facility	10
	26. Analysis equipment	1,452		53. Radiation disposal/shielding facility	26
	27. Sign creation equipment	492		54. Waste disposal facility	50
				Sub-total	939

3.2 The Standardized Classification System's Composition and Utilization

The national research F&E standardized classification system is the most basic classification standard for categorizing national research F&E. According to the standard, the research F&E information is registered and stored in NTIS. It is additionally used as a standard for information linkage with project management systems owned by other departments and institutions. Even when providing functions to identify the redundancy and adequacy of research F&E and to search research F&E information in NTIS, the standardized classification system is used when designing the algorithms in the system. Thus, over the past decade, the standardized classification system has provided functions of the human backbone to manage and utilize national research F&E. The research F&E information established by the government R&D budget must be registered in NTIS within 30 days after acquisition, according to

the Regulations on the Management of National R&D Projects. The registered information is then classified and managed according to the national research F&E standardized classification system (MSIT, 2019). The standardized classification system for national research F&E is hierarchically structured and mutually exclusive of the major classes (8), the sub-classes (54), and the small classes (410). Among the eight major classes, the first, optical equipment, consists of seven sub-classes and 54 small classes, with 6,728 classified research F&E. The second major class, compound analysis equipment, consists of eight sub-classes and 89 small classes, with a total of 11,867 classified research F&E. The third major class, processing and testing for machines and equipment, consists of eight sub-classes and 125 small classes, with a total of 13,908 classified research F&E. The fourth major class, electricity and electron measurement equipment, consists of seven sub-classes and 38 small classes, with a total of 6,500 classified research F&E. The fifth major class, data processing equipment, consists of three sub-classes and 10 small classes, with a total of 5,044 classified research F&E. The sixth major class, physical state measurement equipment, consists of nine sub-classes and 38 small classes, with a total of 4,330 classified research F&E. The seventh major class, Clinical Medical Equipment, consists of 5 sub-classes and 27 small classes, with a total of 955 classified research F&E. The eighth and final major class, environment creation /production/breeding facility, consists of seven sub-classes and 29 small classes, with a total of 939 classified research F&E.

<Table 2> Condition of Statistics in The Standardized Classification System

Classes (num/%)	1. Optical Equipment	2. Compound Analysis Equipment	3. Process and Test for Machines Equipment	4. Electricity and Electron Measurement Equipment	5. Data Process Equipment	6. Physical State Measurement Equipment	7. Clinical Medical Equipment	8. Environment Creation/ Production/ Breeding Facility	Total
Sub Classes	7 (13)	8 (15)	8 (15)	7 (13)	3 (6)	9 (17)	5 (9)	7 (13)	54 (100)
Small Classes	54 (13)	89 (22)	125 (31)	38 (9)	10 (2)	38 (9)	27 (7)	29 (7)	410 (100)
F&E	6,728 (13)	11,867 (24)	13,908 (28)	6,500 (13)	5,044 (10)	4,330 (9)	955 (2)	939 (2)	50,271 (100)

Based on the data registered in NTIS over the past decade (2006-2015), Table 2 shows the statistics of the major class-sub-class-small class structure of the national research F&E classification system and the number of research F&E belonging to the lowest

classification. Based on these results, the composition ratio of sub-class and small class by major class, and the distribution of research F&E classified by small class, were found to be somewhat unbalanced.

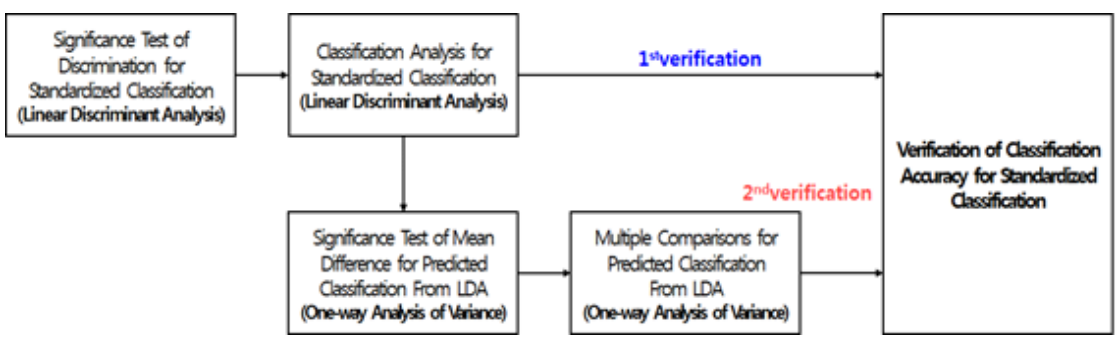
III. Research Model and Methods

1. Development of the Research Model

Korea established the standardized classification system for national science & technology (33 major classes, 371 sub-classes, 2,902 small classes) for efficient operation and management of national R&D projects, and it was first enacted in 2003. The standardized classification for national science and technology is being used in various public fields such as allocation and coordination of R&D budget, management of R&D and performance, project planning and evaluation, and research and analysis of R&D activities (Yoo et al., 2018).

It reflected the specificity of domestic S&T development based on the international S&T classification system, and was developed with a consideration of generality, exclusiveness, similarity, scale, and universality as basic principles. The national standard system for the research F&E was developed by referring to both the standard classification system for national S&T and the advanced cases (e.g., US, Germany), and reorganizing Korea's industrial technology classification (NFEC, 2015). In this study, verification of the accuracy and completeness of the standardized classification system in all areas is a measure of how mutually exclusive and encompassing the whole is. Furthermore, this study attempted to statistically analyze and verify if the items of the sub-classes and small classes are evenly arranged to a suitable degree, and if the distribution of the research F&E belonging to the small classes are appropriately distributed. For this purpose, this study conducted two stages of verification by applying two statistical analysis methods (LDA and ANOVA) as shown in Figure 1. In the first step, LDA, the discrimination significance was examined for eight major classes of the national research F&E classification system. This was done to confirm in advance the normality of the multivariate and the homogeneity of the covariance matrix, prerequisites for discriminant analysis, and to identify statistically significant differences between the eight major classes. After satisfying the discrimination significance, the classification level of each major class could be analyzed by quantifying the distance between the eight major classes, thus quantitatively verifying the accuracy of the classification. In the second step, ANOVA, using the classification values of the predicted groups generated through the linear discriminant function, the significance of the mean difference between the groups was verified. This was performed to confirm the consistency of the optimal group classification

predicted by the discriminant model with the eight major classes of the existing national research F&E standardized classification. Thus, whether or not it is consistent indicates the accuracy of the classification. In addition, a one-to-one mean comparison between the groups was performed through multiple comparison of variance analyses and a post-hoc analysis, thus confirming the distances between the groups through the mean values. Accordingly, this study applied a two-step statistical verification via these two statistical methods. Through this, the classification Based on Fisher's linear discriminant analysis, Wilks's Lambda was .004, and the significance level (Sig.) was .000. As a result, the level of discrimination between sub-class (M_class54) and small class (S_class410) for the eight major classes (difference between the eight groups) was statistically significant. The smaller the Wilks's Lambda is than 1, the higher the discrimination ability. The Wilks's Lambda of this discrimination function is .004, which is very small, indicating very high discrimination ability. accuracy of the national research F&E standardized classification system was precisely analyzed and verified.



<Figure 1> Research Framework

2. Research Methodology

This study applied discriminant analysis for statistical verification and variance analysis for the post-hoc verification, an attempt to derive accurate verification results and various implications. LDA is a classification method originally developed in 1936 by R. A. Fisher. LDA is based on the concept of searching for a linear combination of variables that best separates two classes(Daniela M. W. and Robert T., 2011). Discriminant analysis analyzes the discrimination of the groups based on predefined group information (the classification variable). The appropriate criteria (the discrimination function) necessary to classify the individuals into specific groups are searched to predict which object belongs to which group. Discriminant analysis is a data mining technique that uses the vector

dimension reduction technique without dividing the groups to increase the variance between classes and decrease the variance within classes to separate data with similar properties into different groups. If there are two-dimensional data forming three classes, the dimension can be reduced through projection to one-dimensional subspace using various transformation matrices. When reduced to the best one-dimensional subspace, each class can be easily distinguished even though the data exist in one dimension. Thus, LDA is a technique for reducing the dimensions by minimizing the separation between classes.

Discriminant analysis, together with clustering analysis, are individual directed techniques that derive an equation to discriminate the individuals using the measured characteristic (variable) values for the individuals(Sung and Cho, 2016), and then identify new groups of individuals or calculate their similarities to form clusters of similar individuals. Clustering analysis classifies individuals into clusters by calculating the distance (similarity) between them using the original measured variable. On the other hand, discriminant analysis obtains the discriminant of the linear combination of the original measured variables and uses it to identify the new group of individuals. Thus, discriminant analysis is a method for predicting the group to which an object of observation belongs (Guo et al., 2007, Duintjer T. and Schlesinger, 2007, Dudoit et al., 2002, Dipillo, 1976), similar to clustering analysis in that it assigns the group to which each individual belongs. However, discriminant analysis is clearly different from clustering analysis in that after performing the analysis, the group itself is not presented as an analysis result; rather, it determines to which group the new object belongs, with the number of subgroups present in the population determined in advance, depending on the theoretical background or purpose of the study.

Thus, this study attempted to verify the validity and accuracy of the classification system based on the information accumulated over a decade in the national research F&E standardized classification system. This study adopted SPSS 14.0 as statistical tool to perform discriminant analysis and variance analysis step by step.

Ⅳ. Research Analysis and Results

As considering that it has already been verified through a large number of documents and expert reviews, we focused on analyzing the structure of elements between the classes. This study was conducted on the research F&E's construction information (total 50,271) collected by NTIS during the last 10 years (2006-2015), and it was not measured as the sample, so that both reliability and feasibility analysis of the data were excluded.

In this study, the analytical data were already normalized as they correspond to the population. Next, based on the verification of

the homogeneity of the covariance matrix, both variables of sub-class (M_class54) and small class (S_class410) showed statistical significance (Sig.=.000, $p<.05$) for the 95% confidence interval, thus satisfying the homogeneity of variance, a prerequisite for discriminant analysis.

Based on Fisher's linear discriminant analysis, Wilks's Lambda was .004, and the significance level (Sig.) was .000. As a result, the level of discrimination between sub-class (M_class54) and small class (S_class410) for the eight major classes (difference between the eight groups) was statistically significant. The smaller the Wilks's Lambda is than 1, the higher the discrimination ability. The Wilks's Lambda of this discrimination function is .004, which is very small, indicating very high discrimination ability.

In the discriminant function, the square of the canonical correlation coefficient indicates the explanatory power of the discriminant. Through this analysis, the canonical correlation coefficient of the discriminant function is .990; thus, its explanatory power is $(.990)^2 = .980$. This means that 98% of the variance of the discriminant score, the dependent variable, is explained by two variables, sub-class (M_class54) and small class (S_class410).

<Table 3> Verification of homogeneity of population mean

Class	Wilks's Lambda	F	df1	df2	Sig.
M_class54	.025	284871.835	7	50263	.000
S_class410	.049	139899.303	7	50263	.000

<Table 4> Wilks's Lambda

Function	Wilks's Lambda	Chi-square	df	Sig.
1	.004	277728.814	14	.000

<Table 5> Eigenvalue

Function	Eigenvalue	Variance's %	Accumulation %	Canonical Correlation
1	51.567	93.2	93.2	.990

Lee and Lim (2009) explained that the discrimination ability is significant when the discriminant loading value is ±.30 (or .40) or above. Through this analysis, the discriminant loading value of sub-class is 1.932 and the discriminant loading value of small class is -1.176, thus confirming that both sub-class (M_class54) and small class (S_class410), the independent variables, have discrimination ability.

The classification function in Table 6 is used to decide which group to classify the

new classification object. Independent variable values of the new classification object are each assigned to the following classification function. If the resulting value is large, then it is classified into the large value group, and if the resulting value is small, then it is classified into the small value group. However, this study omitted this step because the purpose of verifying the existing classification supersedes the prediction of discrimination. However, through the classification chart of the canonical discrimination function expressed by the classification function, the degree of difference between the eight major classes or the distance between the groups can be measured. The classification in Figure 2 shows that the interval between groups three and four, and between groups seven and eight is relatively close.

<Table 6> Classification Function Coefficient

Independent	L_class8							
	1	2	3	4	5	6	7	8
M_class54	4.385	9.284	11.976	12.745	20.423	27.487	36.893	41.853
S_class410	-.244	-.450	-.439	-.367	-.814	-1.243	-1.796	-2.085
(Constant)	-8.109	-34.125	-78.901	-114.360	-203.398	-310.502	-505.159	-629.826

Fisher's Linear Discriminant Function

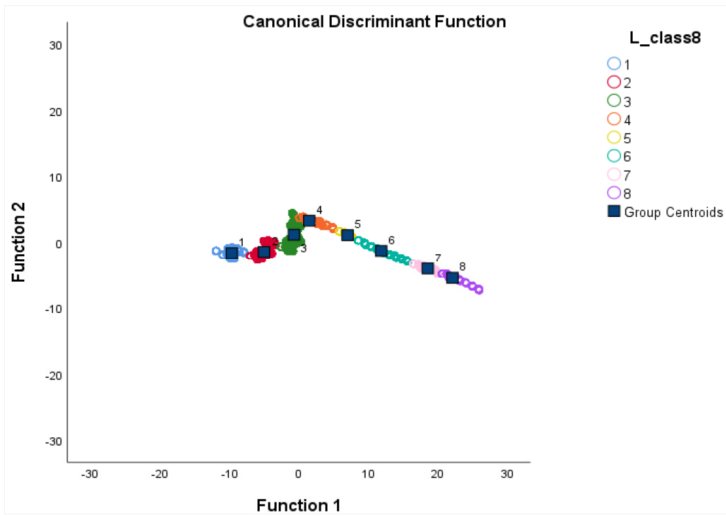
This study visually confirmed the difference in distance between the groups using the classification chart of the canonical discrimination function. Next, this study uses the classification results in Table 7, reflecting accurate figures to examine the degree of data distribution (ratio) between the detailed existing classification and the classification predicted by the canonical discrimination function. The comparison of the predicted groups determined by the canonical discrimination function estimated by sub-class (M_class54) and small class (S_class410) with the existing major class classification showed that the classification accuracy of the existing major class classification was 92.2%, a very high level. However, among the eight major class groups, group three showed the lowest accuracy at 76.6%, followed by group six at 86.7%, which was somewhat low.

This study verified the classification accuracy in the first step through LDA and then conducted ANOVA to re-confirm the accuracy. A variance analysis between the groups predicted through the discriminant analysis and the existing groups was performed to re-verify the classification accuracy using the mean differences between the groups.

In this study, the variance homogeneity of the analytical data, a prerequisite for variance analysis, was statistically significant as shown in Table 8 (Sig. =. 000, $p < .05$).

The ANOVA test of the 8 major class groups predicted through discriminant analysis

showed that the statistic F was 324875.441 and Sig was .000, demonstrating statistical significance and that the mean difference between the eight groups was clearly distinguished.



<Figure 2> Classification chart of canonical discrimination function

<Table 7> Classification Result[a]

		L_class8	Predicted Group								Total
			1	2	3	4	5	6	7	8	
Original Value	Frequency	1	6728	0	0	0	0	0	0	0	6728
		2	0	11867	0	0	0	0	0	0	11867
		3	0	540	10656	2712	0	0	0	0	13908
		4	0	0	0	6404	96	0	0	0	6500
		5	0	0	0	0	5044	0	0	0	5044
		6	0	0	0	0	302	3752	276	0	4330
		7	0	0	0	0	0	0	955	0	955
		8	0	0	0	0	0	0	0	939	939
	%	1	100.0	.0	.0	.0	.0	.0	.0	.0	100.0
		2	.0	100.0	.0	.0	.0	.0	.0	.0	100.0
		3	.0	3.9	76.6	19.5	.0	.0	.0	.0	100.0
		4	.0	.0	.0	98.5	1.5	.0	.0	.0	100.0
		5	.0	.0	.0	.0	100.0	.0	.0	.0	100.0
		6	.0	.0	.0	.0	7.0	86.7	6.4	.0	100.0
		7	.0	.0	.0	.0	.0	.0	100.0	.0	100.0
		8	.0	.0	.0	.0	.0	.0	.0	100.0	100.0

a. 92.2% of the original collective cases were properly classified.

<Table 8> Verification of variance homogeneity

		Levene's F	df1	df2	Sig.
L_class8	Based on average.	7635.683	7	50263	.000
	Based on median.	1468.976	7	50263	.000
	Based on median with modified degrees of freedom.	1468.976	7	18730.577	.000
	Based on cutting average.	6396.908	7	50263	.000

<Table 9> ANOVA

L_class8	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	136910.123	7	19558.589	324875.441	.000
Within Groups	3026.001	50263	.060	-	-
Total	139936.124	50270	-	-	-

This study compared the means of the eight groups through a post-hoc analysis, the results of which are shown in Table 11. In the one-to-one mean comparison of all the groups, all groups showed significance (Sig. > .05) at the 95% confidence interval, confirming that the classification of the groups predicted through the discriminant analysis was accurate. However, based on the analysis of the mean distribution of the predicted groups, as shown in Table 10, the mean difference between group three and group four and the mean difference between group six and group seven were somewhat weak. This verification result is consistent with the result of the previously performed discriminant analysis, re-confirming the result of the discriminant analysis performed in step one.

<Table 10> Mean Distribution of Predicted Groups

Scheffe [a,b]		L_class8							
1 Predicted Group	N	Significance Level = Subset for 0.05							
		1	2	3	4	5	6	7	8
1	6728	1.00							
2	12407		2.04						
3	10656			3.00					
4	9116				3.70				
5	5442					5.04			
6	3752						6.00		
7	1231							6.78	
8	939								8.00
Sig.	-	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

The mean for the groups in the homogeneous subset is displayed.
a. Use harmonic mean sample size 2898.168.
b. The collective size is not the same. The harmonic mean of the group sizes is used. I type error levels are not guaranteed.

<Table 11> Multiple Comparisons

Scheffe, Dependent Variable: L_class8

(I) Predicted Group from Discriminant Analysis	(J) Predicted Group from Discriminant Analysis	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-1.044*	.004	.000	-1.06	-1.03
	3	-2.000*	.004	.000	-2.01	-1.99
	4	-2.703*	.004	.000	-2.72	-2.69
	5	-4.038*	.004	.000	-4.05	-4.02
	6	-5.000*	.005	.000	-5.02	-4.98
	7	-5.776*	.008	.000	-5.80	-5.75
	8	-7.000*	.009	.000	-7.03	-6.97
2	1	1.044*	.004	.000	1.03	1.06
	3	-.956*	.003	.000	-.97	-.94
	4	-1.659*	.003	.000	-1.67	-1.65
	5	-2.994*	.004	.000	-3.01	-2.98
	6	-3.956*	.005	.000	-3.97	-3.94
	7	-4.732*	.007	.000	-4.76	-4.70
	8	-5.956*	.008	.000	-5.99	-5.93
3	1	2.000*	.004	.000	1.99	2.01
	2	.956*	.003	.000	.94	.97
	4	-.703*	.004	.000	-.72	-.69
	5	-2.038*	.004	.000	-2.05	-2.02
	6	-3.000*	.005	.000	-3.02	-2.98
	7	-3.776*	.007	.000	-3.80	-3.75
	8	-5.000*	.008	.000	-5.03	-4.97
4	1	2.703*	.004	.000	2.69	2.72
	2	1.659*	.003	.000	1.65	1.67
	3	.703*	.004	.000	.69	.72
	5	-1.335*	.004	.000	-1.35	-1.32
	6	-2.297*	.005	.000	-2.32	-2.28
	7	-3.073*	.007	.000	-3.10	-3.05
	8	-4.297*	.008	.000	-4.33	-4.27
5	1	4.038*	.004	.000	4.02	4.05
	2	2.994*	.004	.000	2.98	3.01
	3	2.038*	.004	.000	2.02	2.05
	4	1.335*	.004	.000	1.32	1.35
	6	-.962*	.005	.000	-.98	-.94
	7	-1.738*	.008	.000	-1.77	-1.71
	8	-2.962*	.009	.000	-2.99	-2.93
6	1	5.000*	.005	.000	4.98	5.02
	2	3.956*	.005	.000	3.94	3.97

	3	3.000*	.005	.000	2.98	3.02
	4	2.297*	.005	.000	2.28	2.32
	5	.962*	.005	.000	.94	.98
	7	-.776*	.008	.000	-.81	-.75
	8	-2.000*	.009	.000	-2.03	-1.97
7	1	5.776*	.008	.000	5.75	5.80
	2	4.732*	.007	.000	4.70	4.76
	3	3.776*	.007	.000	3.75	3.80
	4	3.073*	.007	.000	3.05	3.10
	5	1.738*	.008	.000	1.71	1.77
	6	.776*	.008	.000	.75	.81
	8	-1.224*	.011	.000	-1.26	-1.18
8	1	7.000*	.009	.000	6.97	7.03
	2	5.956*	.008	.000	5.93	5.99
	3	5.000*	.008	.000	4.97	5.03
	4	4.297*	.008	.000	4.27	4.33
	5	2.962*	.009	.000	2.93	2.99
	6	2.000*	.009	.000	1.97	2.03
	7	1.224*	.011	.000	1.18	1.26

*The mean difference is significant at the .05 level.

Ⅴ. Conclusion

This study statistically verifies the MECE fitness of the current classification system and the information classification using the information on the construction of the research F&E (total 50,271 cases) collected by NTIS over the past 10 years (2006-2015). In addition, this study identifies errors and problems of existing classification system and information registration, and suggests implications and ideas for the improvement of related policies.

In this study, it is found that the combination of linear discriminant analysis and variance analysis is more appropriate than the cluster analysis as a suitable research methodology in order to measure the classification standard of the previously established standardized classification system and the accuracy of the classified information. The verification of the classification accuracy of the standardized classification system of national research F&E by discriminant analysis demonstrated that the classification accuracy was statistically significant. As for the standardized classification systems of national research facilities and equipment, fitness of sub-classes classified in major classes is 98%, fitness of small classes classified in major classes is 88%, and fitness of small classes classified in sub-classes is 85%. The accuracy of the classification decreased as the level went down to the lower level, but the accuracy of the whole classification system

was very high as 92%.

However, for the two groups of processing and testing for machine equipment and physical state measurement equipment, the classification accuracy was somewhat reduced while maintaining a close distance with the surrounding groups. The reason is that the equipment characteristics are similar, such as the thermos-fluid/material property tests included in machining/test equipment or the surface characteristic measurements included in physical measurement equipment. Another reason is that there is a close connection between two equipments, and it is judged to be due to error registration of information or similarity of roles/functions. In future, it is necessary to complement the mutual exclusiveness of the components between machine equipment and the two groups (compound pretreatment analysis equipment, and electrical/electronic equipment) in the major classes stage. Moreover, it is necessary to complement the mutual exclusiveness of the components between physical state measurement equipment and the two groups (data process equipment, clinical medical equipment) in the major classes stage. We are able to adjust the classification of actual data with reference to data of sub-classes classified in surrounding groups for the sub-classes of machines equipment group and physical measurement equipment group in order to make the classification accuracy for Korea's Standardized Classification System of Research F&E.

No previous case has verified the accuracy and completeness of the standardized classification system established in Korea and overseas, so it is meaningful in that it is the first attempted study. Also, by analyzing the information (50,271 cases) managed according to the standardized classification for 10 years, it was confirmed whether or not the actual information is registered and managed with an accurate classification code. Therefore, we believe this study is directly beneficial to the improvement of policy management and classification system for national F&E.

In future, we will carry out the study on the measurement of similar distances in the groups of standardized classification to develop the new methodology for reviewing redundancy.

References

1. Cho, H. H., Sin, T. Y., Song, W. J., An, D. H., Song, S. S., Kim, S. K., Han, Y. B. (2001), *Study on Writing National Science Technology Standardized Classification Table*, Seoul: Science & Technology Policy Institute.
2. Daniela M. W. and Robert T.(2011), "Penalized classification using Fisher's linear discriminant", *Journal of the Royal Statistical Society*: Series B(Statistical Methodology), 73(5), 53-772.
3. Dipillo P.(1976), "The application of bias to discriminant analysis", *Communication in Statistics Theory and Methodology*, A5, 843-854.
4. Dudoit S., Fridlyand J., Speed T. P.(2002), "Comparison of discrimination methods for

the classification of tumors using gene expression data", *Journal of the American Statistical Association*, 97(457), 77-87.

5. Duintjer T. J. and Schlesinger P.(2007), "Improving implementation of linear discriminant analysis for the high dimension/ small sample size problem", *Computational Statistics & Data Analysis*, 52(1), 423-437.
6. Guo Y., Hastie T., Tibshirani R.(2007), "Regularized linear discriminant analysis and its application in microarray", *Biostatistics*, 8(1), 86-100.
7. Hong, S. K., Kim, M., Lee, H. E., Choi, H. R., Kim, B. S., Kwon, S. A.(2016), *Study on classification criteria of national R&D projects for systematic information provision*, Seoul: Korea Institute of Science & Technology Evaluation and Planning.
8. Lee C. Y. and Chen, B. S.(2018), "Mutually Exclusive and Collectively Exhaustive Feature Selection Scheme", *Applied Soft Computing*, 7(68), 961-971.
9. Lee, D., Lee, H., Yoon, I.(2016), "Development of a Classification Scheme for Management of Technology Research: Approach on Research Designs and Methodologies", *Management & Information Systems Review*, 35(4), 269-287.
10. Lee H. S. and Lim J. H.(2009), *Manual for SPSS 14.0*, Seoul: Beob Moon Sa.
11. Ministry of Science and ICT(2019), *Manual on the Management of National R&D Facilities and Equipment*, Gwacheon: Ministry of Science and ICT.
12. National Research Facilities and Equipment Center(2015), *the Standardized Classification System for Research Facilities and Equipment*, PRISM Research Report(22), Daejeon: Korea Basic Science Institute.
13. OECD(2015), *Guidelines for collecting and reporting data on research and experimental development*, Frascati Manual.
14. Seol, S. S. and Song, C. H.(2000), *Theory and Practice of Knowledge Activities Classification*, Hannam University Press.
15. Sin, T. Y, Park, J. H., Jeong, G. H.(1994), *R&D for Korean Technology Classification System*, Seoul: Science & Technology Policy Institute.
16. Sung, B. S. and Cho, W. G.(2016), "Discriminant Factors Influencing Utilization of Genetic Resource", *Management & Information Systems Review*, 35(3), 95-113.
17. Sung, T. K.(2010), "An Essay on the Relationship between Standards and Technological Innovation", *Management & Information Systems Review*, 29(4), 225-244.
18. Yoo, J. Y., Choi, M. J., Kang, S. Y., Lee, B. R.(2018), *the National Science and Technology Standardized Classification System*, Seoul: Korea Institute of Science & Technology Evaluation and Planning.

Abstract

선형판별분석(LDA)기법을 적용한 국가연구시설장비 표준분류체계의 분류 정확도 검증

정석인* · 송영화** · 정의덕***

정부는 연구시설장비가 과학기술의 발전을 견인하는 매우 중요한 도구이자, 수단으로 여겨지면서 국가적으로 R&D와 연구시설장비에 대한 예산 투자를 지속적으로 확대하였다. 또한, 기 구축된 국가연구시설장비의 효율적 운영 및 체계적 관리의 필요성이 점차 대두되면서 2010년 12월, 국가연구시설장비 표준분류체계를 개발하였다. 그러나 연구현장에서는 국가연구시설장비의 NTIS(National Science and Technology Service) 정보수집 초기단계로 누적정보 부족에 따른 표준분류체계의 과학적 검증절차 부재와 동일계층 간 분류기준의 비일관성 문제가 여전히 한계로 제기되고 있다.

따라서 본 연구는 지난 2010년, 2015년 각 제/개정된 국가연구시설장비 표준분류체계(대분류 8개, 중분류 25개, 소분류 410개)의 분류 정확도를 측정하고자 선형판별분석(LDA)과 분산분석(ANOVA) 기법을 적용하여 2단계로 분석하였다. 또한, 본 연구 분석을 위해 지난 10년 동안 NTIS에 누적 등록된 정보데이터(Big-Data) 50,271건을 수집하여 이를 활용하였다. 이는 단순히 국내외 유사 분류체계와 전문가 의견을 토대로 만들어진 現 국가연구시설 표준분류체계를 과학적으로 실증 검증한 첫 연구 사례에 해당된다.

본 연구 결과, 대분류 이하 중분류와 소분류로 분류된 개체 수의 집단별 판별정확도는 92.2% 로 매우 높은 수준이었고, 분산분석을 통한 사후검증에서는 대분류 8개 중 2개 집단의 변별력이 다소 낮게 나타나, 現 표준분류체계 중 일부 개선이 필요한 것으로 조사되었다. 본 연구를 통해 現 국가연구시설장비 표준분류체계가 향후 지속적으로 개선되길 바란다.

핵심주제어: 연구시설장비, 표준분류체계, 판별분석, LDA, 분류검증

* 한국기초과학지원연구원 국가연구시설장비진흥센터 팀장(제1저자), jsi494@kbsi.re.kr
** 건국대학교 경영대학 기술경영학과 교수(교신저자), sawng@konkuk.ac.kr
*** 한국기초과학지원연구원 국가연구시설장비진흥센터 선터장(공동저자), edjeong@kbsi.re.kr