


Article

Big Data Analytics in Government: Improving Decision Making for R&D Investment in Korean SMEs

Eun Sun Kim ¹, Yunjeong Choi ² and Jeongeun Byun ^{2,*} 

¹ Data Analysis Division, Korea Institute of Science and Technology Information, 66 Hoegi-ro, Dongdaemun-gu, Seoul 02456, Korea; kimes@kisti.re.kr

² Technology Commercialization Center, Korea Institute of Science and Technology Information, 66 Hoegi-ro, Dongdaemun-gu, Seoul 02456, Korea; yjchoi@kisti.re.kr

* Correspondence: jebyun@kisti.re.kr; Tel.: +82-2-3299-6295

Received: 31 October 2019; Accepted: 20 December 2019; Published: 25 December 2019



Abstract: To expand the field of governmental applications of Big Data analytics, this study presents a case of data-driven decision-making using information on research and development (R&D) projects in Korea. The Korean government has continuously expanded the proportion of its R&D investment in small and medium-size enterprises to improve the commercialization performance of national R&D projects. However, the government has struggled with the so-called “Korea R&D Paradox”, which refers to how performance has lagged despite the high level of investment in R&D. Using data from 48,309 national R&D projects carried out by enterprises from 2013 to 2017, we perform a cluster analysis and decision tree analysis to derive the determinants of their commercialization performance. This study provides government entities with insights into how they might adjust their approach to Big Data analytics to improve the efficiency of R&D investment in small- and medium-sized enterprises.

Keywords: big data; decision tree; government; national R&D project; small and medium-sized enterprises; commercialization performance

1. Introduction

The concept of Big Data analytics (BDA) pertains to accumulating, combining, analyzing, and using large-scale data for various purposes and of various types. BDA enables organizations in both the private sector and, increasingly, the public sector to make better decisions (i.e., more quickly and efficiently) based on evidence and insights [1–3]. Indeed, Big Data applications in government are no longer unusual. Many countries have come to regard Big Data as a growth engine for the future as well as a solution to existing economic and social problems. Over the past decade, governments globally have announced comprehensive strategies for using Big Data at the national level. They first focused on the construction of infrastructure to open access to data and promote its utilization. Thereafter, they supported legal and institutional improvements to empower the private sector to use public data and create added value (indirect role) as well as used Big Data for policymaking (direct role) [4].

Indeed, building on the constructed infrastructure, most governments endeavor to expand their own use of Big Data to formulate policies based on concrete data rather than depending on mere experience or intuition. The use of Big Data has thus far been limited because of the lack of actual data available to the government to implement such data-driven policies. In particular, the use of Big Data has been scarce because of the limitations of the infrastructure required to (i) accumulate and generate reliable data, which is essential for utilization; and (ii) convert the accumulated and generated data into a form that can actually be used in practice. However, an infrastructure that can

generate, accumulate, and analyze data has now been established and the discussion on data-driven policies is re-emerging. The establishment of data-driven policies using Big Data and BDA can help public administrators at all levels of government and in different areas reach their goals. It can also prevent the inefficient operation of the government, bad policymaking, and the selection and execution of misguided alternatives. In summary, complex policy issues affected by various variables can be handled efficiently and effectively using Big Data and BDA, and the new data-driven insights gained can aid the decision-making of the government.

Governments have been using Big Data for policymaking in several ways. Basic statistical or quantitative analyses have been performed based on census data on numbers of people, their living conditions, and other socioeconomic characteristics collected from sampling or public administration records on taxes, employment, and so on. At a more advanced level, some governments have built public health and medical systems through the integration and analysis of data such as medical records and insurance information, performed disaster forecasting using traffic data, and established public safety policies based on crime-related data analysis. Specifically, in the context of the present study, many countries have expanded governmental investment in research and development (R&D), especially as science and technology have emerged as the driving forces behind economic development and national competitiveness. Together with this trend, there is an increase in demand to improve the efficacy of governmental R&D investment to ascertain whether such investment actually provides a return, determine the issues if the return on investment is low, and understand the strategic investment methods needed to improve results.

This study expands the concept of BDA to the governmental sector and derives the optimal solutions for governmental R&D investment in Korea. Hence, it departs from those fields in which Big Data has already been used for policymaking and focuses on policymaking using Big Data and BDA in new fields to understand how to improve the efficiency of government-sponsored R&D projects. We perform a cluster analysis and a decision tree analysis, a predictive modeling method widely used for machine learning, based on data on around 43,800 national R&D projects in Korean small and medium-sized enterprises (SMEs: SMEs in this study are described as establishments with fewer than 250 employees), which play a key role in the national economy. Since the 2000s, the Korean government has expanded its R&D investment in SMEs to enable such firms to commercialize the results of R&D projects and thereby raise added value. This study thus analyzes the government-sponsored R&D projects conducted by SMEs to identify the factors that determine whether the goals of such governmental support are achieved. Methodologically, it uses data taken from the world's first national R&D information knowledge portal supplied by the National Science and Technology Information Service (NTIS). This portal provides information on about 540,000 national R&D projects, such as governmental R&D investment, number of projects, and innovative performance. Using these NTIS data on a variety of national R&D projects, we derive the determinants of commercialization performance and suggest a systematic method of increasing the efficiency of national R&D investment, with a particular focus on increasing the commercialization performance of SMEs.

The results of this study contribute to the body of knowledge on this topic by establishing strategies for using Big Data to achieve data-driven policymaking. As governmental investment in R&D has reached 20 trillion KRW in Korea (the average exchange rate in 2018 was KRW 1,101.48=USD), we expect the results of this study to be particularly useful for planning governmental investment in R&D more efficiently and expanding commercialization successes.

The remainder of the paper is organized as follows. In Section 2, we examine the theoretical background by reviewing previous studies of governments' use of Big Data. We also examine the status of governmental R&D investment in Korea and identify problems in the government's decision-making process. Section 3 describes the data and analytical method used to solve the problems faced by the Korean government. Section 4 presents the results of the analysis, and Section 5 summarizes the results and concludes.

2. Theoretical Background

2.1. Government and Big Data

Evidence-based policy, which refers to establishing policies grounded on objective and scientific research and ensuring they are designed and implemented based on concrete data, has existed since ancient times. In Ancient Greece, Aristotle argued that diverse sources of knowledge should be included to set rules or develop regulations; Aristotle's concept of diverse knowledge has been interpreted to include scientific knowledge [5]. In the medical field in the early 1990s, the phrase "evidence-based" was formulated to refer to medical practices based on evidentiary data [6] and the phrase has since entered into generalized use. It is only in recent years, however, that the emphasis on evidence-based practices has entered the field of government [7,8].

Governmental institutions have traditionally selectively generated and managed the information the government needs, using data for institutional maintenance and reinforcing organizational capabilities rather than making them publicly available. Gradually, however, governments have been encouraged to change this monopolistic approach to information management. Owing to rapid changes in sociocultural environments and behaviors caused by globalization and the increasing complexity and diversity of society as well as the development of ICT, there have been significant changes in the environment in which governmental policies are implemented [9]. It is increasingly argued that the government should eschew opinion-based policies and the selective application of evidence, driven by ideological perspectives, prejudices, and conjecture, and instead make policy decisions based on citable evidence, with ample access to data [10,11]. Research has found that through evidence-based policymaking, governments can gain trust in a changing environment [12], justify policy decisions, make policy decisions more quickly, resolve conflicts in the process of formulating and implementing policies, and improve the quality of policies [13–15].

The key factors to evidence-based policymaking are securing the objectivity of the materials or data used [16] and conducting scientific analysis [17]. Therefore, to reach the stage where evidence-based policies can be established, it is first necessary to collect high-quality data that enable the suitable analysis of the issue in question, select scientific methods to analyze this accumulated data, and apply the analytical results to the process of designing policies. However, in only a few limited fields, such as healthcare, security, and public safety, and environmental monitoring and response measures, have governments been able to secure sufficient data with proven objectivity, conduct scientific analysis, and apply these findings to formulate policies. One of the main reasons for this slow progress has been the lack of accumulated data. However, data-based practices are now expected to become applicable to a wider range of fields thanks to improvements in data collection, integration, and analysis techniques [18].

Studies have analyzed the application of BDA to governmental practices in the healthcare, security, and public safety sectors. In the healthcare sector, governments use Big Data to find the strongest scientific basis for suppressing increases in medical expenses. One of the top priorities of governments has been building an infrastructure that links the Big Data from various organizations through the construction of databases that connect the individual patients of public administrative and medical organizations by developing a network of data on existing medical services [1,19]. Using such databases, studies have analyzed the optimal treatments and cost reductions based on predictions of high-cost patients, readmitted patients, and occurrences of complications and medical incidents; other studies have focused on applying these data and achieving service optimization through personalized medical services, clinical decision support systems, and mobile devices [20,21].

In the public safety sector, governments identify crime trends by analyzing the times, areas, and types of crime incidents within criminal records. These data are used to establish public safety policies such as dispatching more police officers to certain crime-prone areas. Based on the analyzed information, an application has been developed to improve citizens' safety; this application notifies citizens in areas in which crime is expected to occur to reduce crime rates. Studies of these issues are

referred to as security informatics, an area of expertise continuously advancing through the integration of technical, organizational, and policy-based approaches [22–24].

2.2. R&D Policy of the Korean Government: the Korea R&D Paradox

With the advent of the knowledge-based society in the 21st century, science and technology have emerged as new growth engines for strengthening national competitiveness, outstripping the importance of other factors of production such as capital and labor. As a result, countries globally are continuously expanding investment in R&D to secure these growth engines. The Korean government has also increased its R&D investment in pursuit of economic development through science and technology. As of 2017, R&D investment in South Korea amounted to 78.8 trillion KRW, the fifth largest in the world and the largest globally in proportion to GDP. Of this total, the government's R&D expenditure was KRW 19.4 trillion, nearly 5.5 times greater than the 3.5 trillion KRW spent in 2000 [25]. In 2017, government-funded research institutes received 7.9 trillion KRW, academia 4.4 trillion KRW, SMEs 4.1 trillion KRW, large firms 0.4 trillion KRW, and other actors, including public research institutes, 2.6 trillion KRW. In particular, R&D investment in SMEs has been steadily increasing, rising from 2,854 billion KRW in 2013 to 4,119 billion KRW in 2017 (The proportion of R&D support for SMEs was calculated based on the purchasing power parity index in 2010; the equivalent index of South Korea was 56.8%, significantly higher than the percentages in the United States (11.4%), France (24.8%), and the United Kingdom (25.2%) [26]); however, investment in large firms has been decreasing, falling from 861 billion KRW to just 419 billion KRW in 2017 [25]. The government expects to improve commercialization performance and economic growth by implementing R&D support for SMEs, which account for 99% of all enterprises in Korea.

However, despite this proactive support, the level of commercialization performance achieved as an outcome of government-sponsored R&D projects has continued to be low. According to the government's plan announced in 2014 to promote innovation among SMEs, the success rate of commercialization attributed to government R&D projects for SMEs has been only around 50% [27]. Recently, the Presidential Advisory Council on Science and Technology formulated and approved a national R&D innovation plan including a provision to double R&D investment for SMEs. The plan sets quantitative targets to support SMEs, requiring government agencies and public institutions, which have annual R&D budgets of 30 billion KRW, to invest a certain percentage of their R&D funding in SMEs [28]. The problem is that performance has been analyzed in a fragmented manner based only on R&D investment and the number of commercialized projects, and there is no systematic analysis of which projects involving SMEs have achieved successful commercialization outcomes thanks to R&D investment and whether such commercialization has generated actual sales.

The Korean government established the NTIS in 2006 to share and jointly utilize information on national R&D projects, which had previously been managed by individual departments. However, the government's utilization of NTIS data has been limited to merely presenting R&D expenditure by actor, research phase, and region using basic statistical analysis or releasing the number of achievements such as papers, patents, and technology transfers, including commercialization performance. Although the government has collected sufficient data on national R&D projects, it has been unable to effectively apply data analytics to formulate data-driven R&D policies.

3. Analytical Method

3.1. Analysis Procedure

This study aims to derive the optimal solution for enhancing the efficiency of governmental R&D investment in SMEs. As most of the data on R&D projects carried out in 2018 have not yet been entered into the NTIS, we extracted data on 48,309 national R&D projects conducted by SMEs from 2013 to 2017. Python was used for data preprocessing and analysis.

We next employed cluster analysis to group the data and thus examine the determinants of commercialization performance. Cluster analysis is suitable for the exploration of the large amounts of R&D project data made available by the Korean government. In addition, it can classify these R&D project data to show their characteristics. Using the results of the cluster analysis, we can then understand the structure of project data in high-performing projects when commercialization performance outcomes are not revealed by using indicators such as average investment.

We then clustered the data into groups using the self-organizing map (SOM) algorithm [29]. Cluster analysis methods such as principal component analysis can efficiently form clusters using a small quantity of data to interpret large-scale multidimensional data. However, some data are lost because of the linear data reduction issue; another problem is that these methods are unsuitable for analyzing non-linear targets [30,31]. To avoid these problems, we thus used the SOM algorithm, which can process large-scale data quickly and performs the strongest of all available hierarchical cluster analysis methods [32].

Table 1 shows the 13 input variables for the cluster analysis. These variables were based on the project information available from the NTIS in 2017, which the Korean government uses for the investigation, analysis, and evaluation of national R&D programs [25]. Based on the clustering results, we used four indicators of commercialization performance in the NTIS, namely the average number of commercialized projects, commercialization period, sales from commercialized projects, and the number of jobs created by commercialized projects, to compare and analyze each cluster (Table 2). Finally, we conducted a decision tree analysis using the classification and regression tree (CART) algorithm to identify the determinants of commercialization performance for the projects in the clusters [33]. The input variables for the decision tree analysis included not only the variables in Table 1, but also the categorical variables that could not be used in the cluster analysis. Table 3 shows the added variables and their descriptions.

Table 1. Input variables for the cluster analysis.

Variable	Description
Project period	Total project period
Source of funding	Source of funding for R&D projects classified into a general account, a special account, and funding
Government investment	R&D expenditure invested by the central government
Private cash ratio	Cash ratio accounts for among the private-sector contributions by the relevant performing organization and/or the local government in addition to the proportion of the budget provided by the central government
Private non-cash ratio	Non-cash ratio accounts for among the private-sector contributions by the relevant performing organization and/or the local government in addition to the proportion of the budget provided by the central government
Research commissioned	Number of research projects commissioned and managed by enterprises when some of the R&D projects are performed under a contract or jointly performed
Funding for research commissioned	Total support funding for research projects commissioned and managed by enterprises when some of the R&D projects are performed under a contract or jointly performed
Research commissioned by an actor	Actor performing under a contract or jointly performing some of the R&D project managed by enterprises; classified into enterprise, university, government-funded research institute, foreign research institute, and other
Continuation	R&D projects classified into new or continued projects. The latter refers to projects whose project period has expired, but that have been confirmed to continue

Table 1. *Cont.*

Variable	Description
Phase	R&D phase classified into basic research, applied research, development research, and other
Characteristics	Characteristics of the final outcomes of R&D projects in terms of: project period, project budget, capabilities of researchers, current level of R&D; classified into idea development, prototype development, product or process development, and other
Practical use	Target of R&D projects; classified into practical and non-practical use
Technology life cycle (TLC)	Technology life cycle of R&D projects; classified into introduction, growth, maturity, decline, and other

Table 2. Indicators of commercialization performance.

Indicator	Description
Number of commercialized projects	Number of commercialized projects
Commercialization period	Difference between the year of commercialization and year of the project start
Sales	Sales from commercialized projects
Job creation	Number of jobs created by commercialized projects

Table 3. Input variables for the decision tree analysis.

Variable	Description
Name of department	Name of the administrative department that manages all aspects of the planning, evaluation, and management of R&D projects
Research field	Research field of R&D projects; into nature, life, and artificial following the national standard classifications of science and technology
Application field	Application field of R&D projects; classified into industry and the public sector

3.2. Analysis Model

3.2.1. Cluster Analysis: The SOM Algorithm

The SOM algorithm, proposed and developed by Kohonen [29,34], is an unsupervised neural network used to visualize and analyze high-dimensional data in the form of maps arranged in easy-to-understand low dimensional neurons. It consists of two layers of artificial neural networks; one is the input layer that receives input vectors and the other is the competitive layer comprising a two-dimensional grid. In this layer, vectors are clustered at one point according to the characteristics of the input vector. The input layer has the same number of neurons as the number of input variables, and the competitive layer has the same number of neurons as the number of clusters predetermined by the user. The data in the input layer are arranged in the competitive layer through learning, which is called a map. The sorted data is displayed as a grid on the map. Data with similar patterns are located close together on the map, while data with different patterns are located far away from each other. This allows us to easily visually assess not only similarities in the clusters but also similarities between the clusters. To determine the optimal number of clusters, we compared the silhouette coefficient with the number of clusters and conducted the analysis based on the number of clusters with the highest coefficient.

3.2.2. Decision Tree Analysis

Decision tree analysis classifies decision rules into a tree structure to perform the classification and prediction. It is a data mining-based distribution technique that searches for large amounts of unexpected or valuable structures. After the major input variables in a large amount of data are found, decision tree analysis is useful for effectively analyzing the interactions between the individual factors

to determine how the various interactions affect the target. In addition, since the analysis process is expressed through a tree structure, it is easy to interpret.

The CART algorithm can be applied regardless of the scale of the objective or input variable. Moreover, the decision tree can be easily interpreted by dividing it through binary splits rather than multiple splits. Another advantage of this approach is that the process of analysis is expressed in the form of trees, which simplifies the interpretation and requires no assumptions of linearity or normality in the variables. This enables the use of both continuous and categorical variables.

Depending on the type of objective variable, the CART algorithm classifies continuous and categorical variables under the classification tree and regression tree, respectively. In cases where the objective variable is categorical, such as in this study, the Gini index and entropy are used to measure impurity. Optimal splitting is conducted by selecting the input variables that minimize the Gini index and entropy [35]. Furthermore, it is robust in response to outliers, and is a non-parametric method that does not require assumptions about the distribution. Since the first separation occurs for the variable with the strongest explanatory power, it is an effective method for identifying important variables. Hence, this study used the CART algorithm to derive a predictive model for the creation of qualitative commercialization performance, which is then verified using 10-fold cross-validation. To evaluate the performance of the predicted outcomes, we use the receiver operating characteristic curve to calculate the area under the curve (AUC) [36].

4. Results

4.1. Clustering Results

We used the SOM algorithm to cluster the 48,309 national R&D projects conducted by SMEs from 2013 to 2017. First, we compared the silhouette coefficients for each number of clusters to select the optimal number of clusters. Upon comparing the values from 2×2 up to 10×10 , we found that the 3×3 clusters had the highest silhouette coefficient value (0.4523) and therefore we conducted clustering using 3×3 clusters (Figure 1).

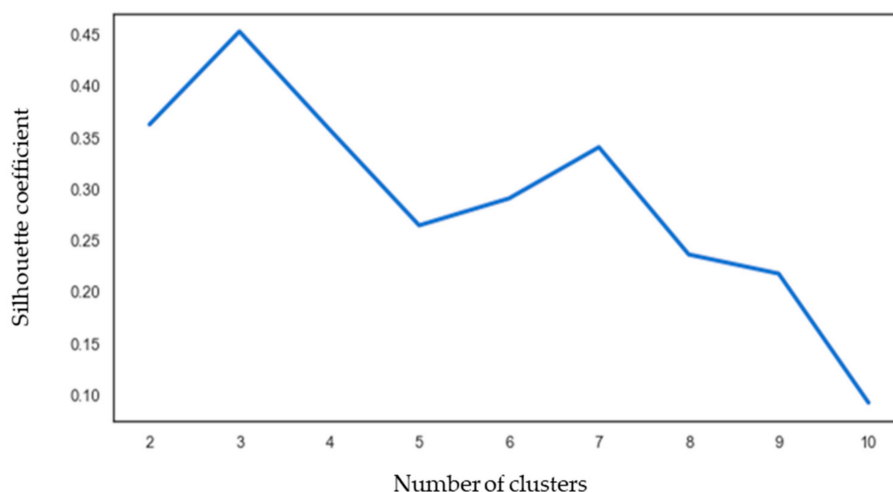


Figure 1. Silhouette coefficients by the number of clusters.

Figure 2 presents the results of the cluster analysis using 3×3 clusters, showing the distribution of observations across each cluster. Cluster 21 (C21) was the largest, with 15,006 projects, followed by C20 and C02, while C10 was found to be the smallest cluster. As the clustering results included no outlier clusters, we analyzed all the clusters to calculate the average governmental investment per R&D project, as shown in Figure 3, and the average number of projects in which R&D successfully led to commercialization. In the case of successfully commercialized projects, we examined the

average number of commercialized projects, average commercialization period, average sales from commercialized projects, and average number of jobs created by commercialized projects (Figure 4).

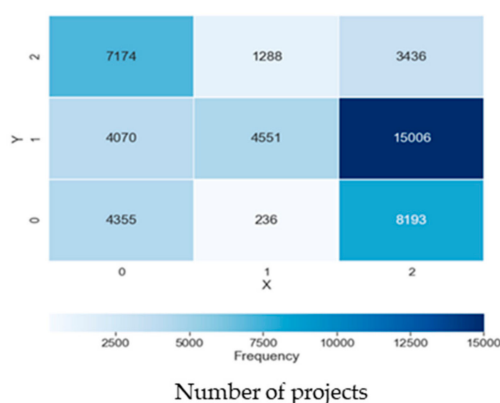


Figure 2. Clustering results using the SOM algorithm.

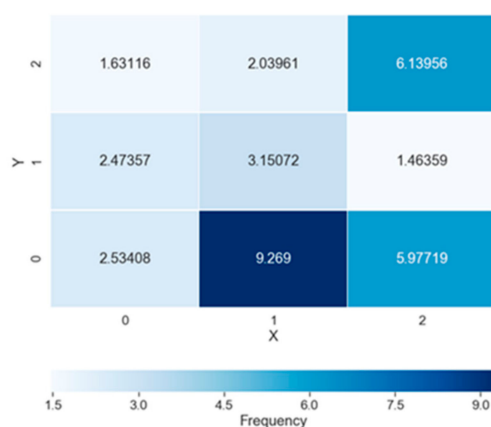


Figure 3. Average government investment per R&D project (Unit: 100 million KRW).

First, the clusters in which R&D projects led to the highest commercialization performance were C10 and C00. The average time required in C10 and C00 to yield commercialization performance was relatively short, at 0.37 and 0.16 years, respectively. However, while the projects in these clusters reached commercialization in the short term, they were found to have performed more poorly than those in other clusters in terms of qualitative performance, such as sales and job creation. This finding indicates that rapid commercialization in more R&D projects does not necessarily lead to qualitative performance. In particular, although C10 received the largest amount of governmental investment, it was observed to have poor commercialization performance.

Conversely, the cluster with the longest commercialization period, C01, was found to be among the three worst performing clusters in terms of generated sales and job creation. This finding shows that a longer average commercialization period also does not necessarily lead to strong qualitative performance. For C21, another cluster that had a longer commercialization period, it took more than one year to achieve commercialization. However, C21 was among the three best performing clusters in terms of sales as well as exhibiting relatively strong job creation performance. Most of the projects in C20, which performed well in terms of both measures of qualitative performance (sales and job creation) were found to have reached commercialization within six months of the completion of the R&D projects. While the cluster with the highest revenue, C22, appears to have generated high revenue due to the large number of commercialized projects, it also performed relatively well in terms of job creation while also requiring only a short time to reach commercialization (under six months). Considering these findings, we conclude that the commercialization period does not appear to be a determining factor for the qualitative performance of commercialization.

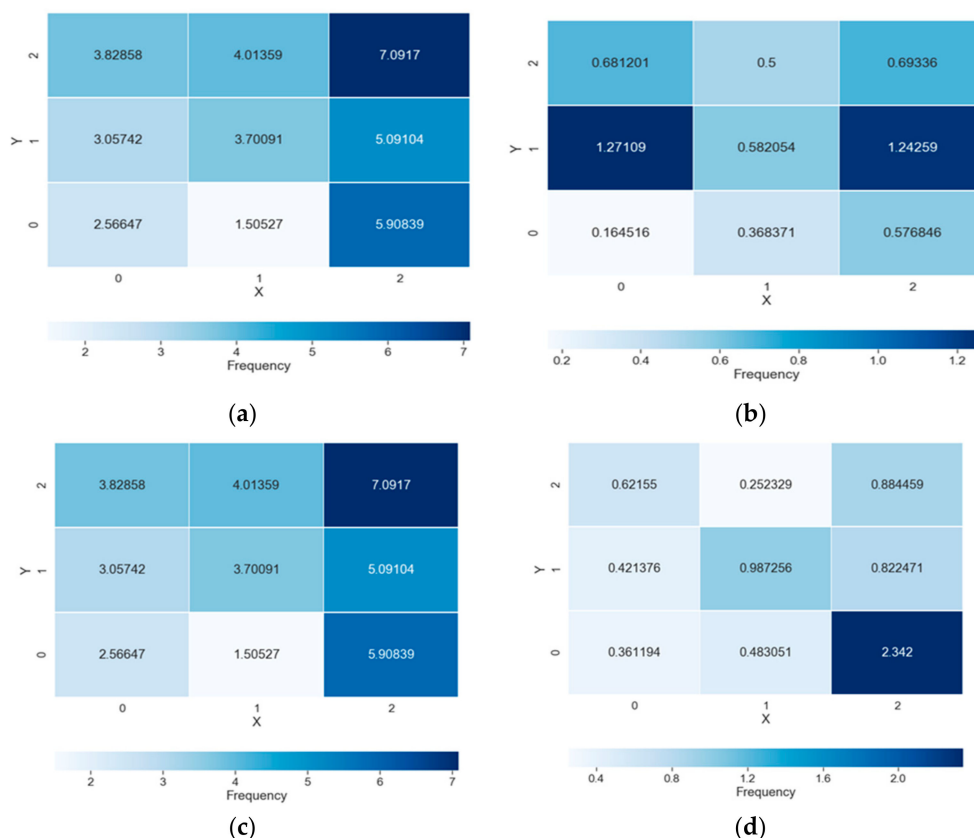


Figure 4. Commercialization performance of each cluster. (a) Average number of commercialized projects (Unit: number of projects). (b) Average commercialized period (Unit: year). (c) Average sales from commercialized projects (Unit: 100 million KRW). (d) Average number of jobs created by commercialized projects (Unit: number of jobs).

4.2. Determinants of Commercialization Performance in Each Cluster

Based on the results of the cluster analysis, we conducted the decision tree analysis to identify the specific factors that led to the qualitative performance in C20 and C22. We also examined which factors led to the qualitative performance in C21 as opposed to another cluster that had similar times to commercialization, C01, which required a longer period to reach commercialization than C20 and C22. Table 4 reports the measured AUC values. Values closer to 1 indicate the higher accuracy of the predictive model; an AUC value of 1 indicates perfect accuracy, while values lower than 1 but greater than or equal to 0.9 may be interpreted as indicating high accuracy. Since all the AUC values measured for each cluster exceed 0.9, the predictive models for each cluster derived in the decision tree analysis can be regarded as being reliable.

Table 4. AUC values of each cluster.

Cluster	AUC Value
C00	0.9996
C01	0.9994
C02	0.9998
C10	0.9881
C11	0.9982
C12	0.9996
C20	0.9982
C21	0.9998
C22	0.9999

The results of the decision tree analysis for each cluster are as follows. In the case of C20, which yielded the strongest qualitative performance in terms of job creation, projects designated for “practical use”, characteristics equal to “other development”, and a technology life cycle equal to “other” had a 0.9814 probability of being in C20. Next, projects designated for “practical use”, characteristics equal to “other development”, a technology life cycle equal to “emerging”, and a phase equal to “applied research” had a 0.9600 probability of being in C20. Projects designated for “practical use”, characteristics equal to “other development”, a technology life cycle equal to “growth”, “maturity”, or “decline”, and a phase equal to “applied research” had a 0.8312 probability of being in C20 (Figure 5).

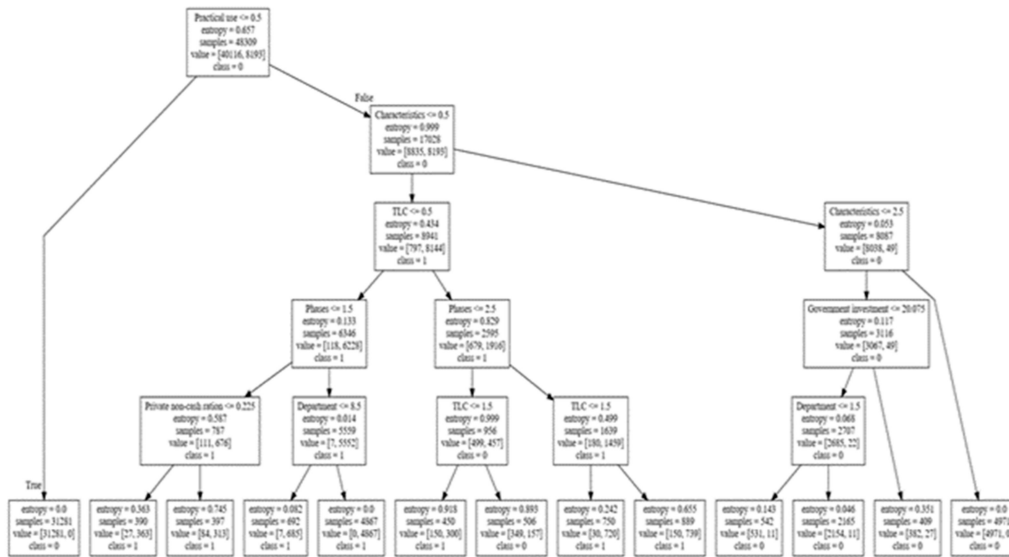


Figure 5. Decision tree analysis results for C20.

In the case of cluster C22, which demonstrated the strongest qualitative performance in terms of sales, new projects “not designated for practical use” with characteristics equal to “idea development” or “other development” had a 0.9994 probability of being in C22 (Figure 6). Among the projects in C21, which had a longer period to commercialization than C20 and C22 but yielded strong qualitative performance in terms of sales and job creation, new projects “not designated for practical use” with characteristics equal to “product or process development” had a 0.9993 probability of being in C21 (Figure 7). Among the projects in C01, which had a longer period to commercialization, as in the case of C21, but performed poorly in terms of sales and job creation, new projects designated for practical use with characteristics not equal to “other” had a 0.9801 probability of being in C01 (Figure 8).

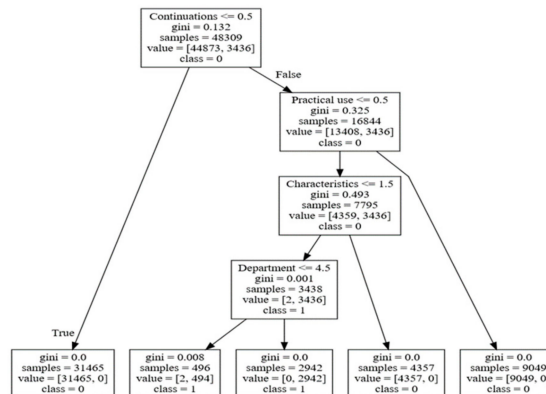


Figure 6. Decision tree analysis results for C22.

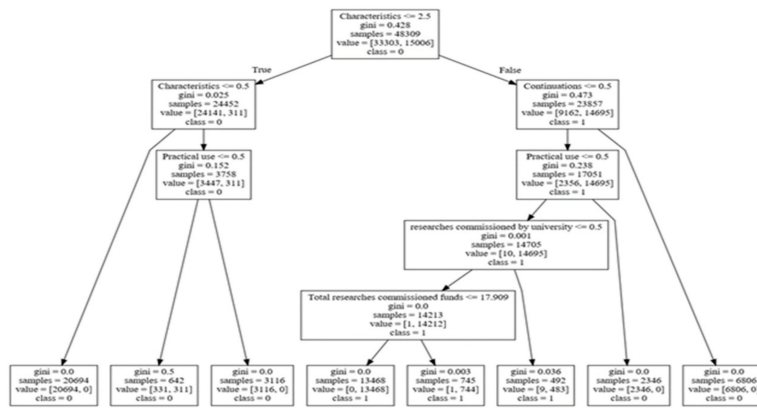


Figure 7. Decision tree analysis results for C21.

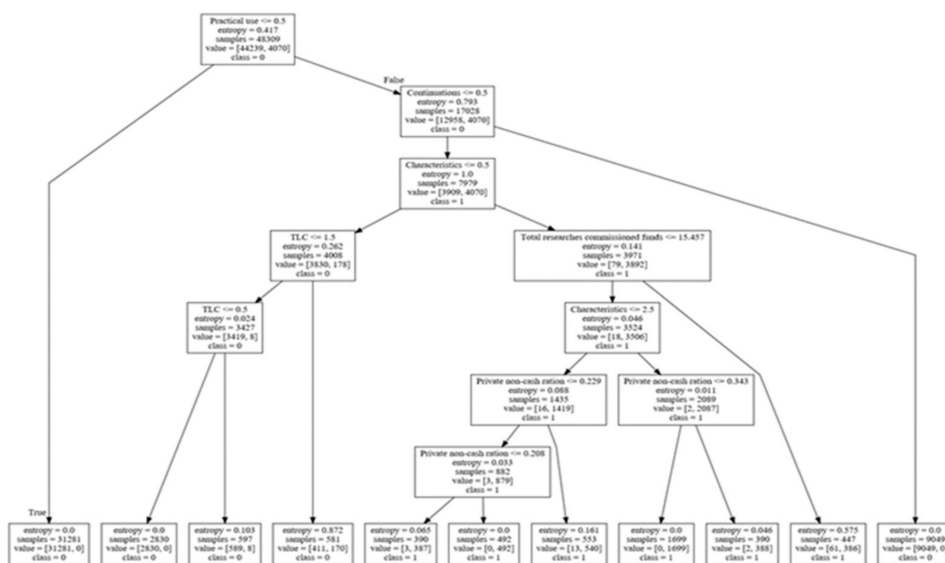


Figure 8. Decision tree analysis results for C01.

Of the projects that required a longer-than-average commercialization period, those not designated for practical use were found to perform better. Projects not designated for practical use with characteristics equal to “product or process development” were found to take longer until commercialization but performed better in terms of sales and job creation when they were commercialized successfully. Therefore, projects not designated for practical use with characteristics equal to “product or process development” appeared to require sufficient time rather than rapid commercialization.

C22 showed a large number of projects for non-practical use. In addition, C21, which generated strong qualitative performance, had many projects for non-practical use. However, C01 had a low number of commercialized projects and did not create high qualitative performance, and projects for practical use belonged to C01. Projects for practical use are those in which firms participate to commercialize technology to generate economic and social value from sales and job creation. However, such projects for practical use failed to achieve the expected levels of commercialization, indicating a mismatch in government R&D policies.

5. Conclusions and Implications

This paper presented a new case of a government’s application of BDA. Based on data on national R&D projects in Korea, we conducted cluster and decision tree analyses to identify the determinants of commercialization performance. These analyses showed a low success rate of commercialization for

national R&D projects. Among successfully commercialized projects, many were not for practical use, indicating a mismatch in government R&D policies. In addition, many projects were commercialized but failed to create sales or jobs; this shows a lack of social and economic value creation, which is the primary goal of governmental R&D investment in SMEs, and thus a failure to realize a return on investment.

The findings of this study suggest the following policy implications. First, considering the finding that governmental investment did not lead to the determinants of commercialization performance, policymakers must be selective and focused when they design R&D policies for SMEs, as the expansion of inputs does not necessarily lead to an increase in outputs. It seems that the linear-based viewpoint, in which increasing R&D investment simply leads to national economic growth, has prevailed in the policymaking arena. In other words, the results of the cluster analysis show that under the current structure of R&D support, large investment projects do not lead to qualitative commercialization performance. Moreover, although the proportion of projects with less investment is large, such projects do not lead to qualitative performance, either. As such, to enhance the effects of R&D support, it is necessary to first elaborate on how to select supported targets and determine the optimal investment for them.

Second, policymakers must conduct integrated reviews of projects designated for practical use. Whether being designated for practical use was analyzed as the determinants of commercialization performance. Interestingly, however, projects not designated for practical use, but aimed at the development of products or processes, have higher commercialization performance than those projects designated for practical use and with the characteristics of R&D that may directly lead to commercialization performance such as prototypes or product/process development. Policymakers focus on the practical application and commercialization of technologies when providing R&D support for SMEs as well as expanding the proportion of projects designated for practical use; however, the findings of this study show that the effectiveness of such efforts is low. As such, it is necessary to fully review the achievability of objectives and possibility of the realization of performance when selecting projects for practical use rather than first expanding the proportion of projects for practical use.

Finally, policymakers must review the R&D information collected by the NTIS when establishing data-driven R&D policies. It is difficult to interpret those R&D characteristics analyzed as determinants of commercialization performance when they are categorized as “others”, making it hard to apply them when formulating policy. Indeed, it is difficult to identify their exact intent because of a lack of standardization. It is thus necessary to ensure collected items can be converted into analyzable data to help policymakers apply the data derived from national R&D projects in practice.

Hence, this study makes a significant contribution to the literature by expanding the field of governments' application of BDA and presenting a case of policymaking based on data. In addition, it shows that the government should be concerned about what data can be made available in the future to make policy decisions. Future research is, however, necessary to more closely examine the factors identified in this analysis as determinants of commercialization performance.

Author Contributions: E.S.K. conceived the methodology and wrote the first draft. Y.C. investigated previous research and analyzed the data. J.B. developed the overall idea and reviewed the final draft. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This research was supported by the Korea Institute of Science and Technology Information.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Davenport, T.H.; Harris, J.G. *Competing on Analytics: The New Science of Winning*; Harvard Business Press: Boston, MA, USA, 2007; pp. 3–22.

2. LaValle, S.; Lesser, E.; Shockley, R.; Hopkins, M.S.; Kruschwitz, N. Big data, analytics and the path from insights to value. *MIT Sloan Manag. Rev.* **2011**, *52*, 21.
3. European Commission. *Big Data Analytics for Policy Making*; European Commission: Brussels, Belgium, 2016; Available online: https://joinup.ec.europa.eu/sites/default/files/document/2016-07/dg_digit_study_big_data_analytics_for_policy_making.pdf (accessed on 10 October 2019).
4. Jiang, H.; Shao, Q.; Liou, J.J.; Shao, T.; Shi, X. Improving the sustainability of open government data. *Sustainability* **2019**, *11*, 2388. [[CrossRef](#)]
5. Flyvbjerg, B. *Making Social Science Matter: Why Social Inquiry Fails and How It Can Succeed Again*; Cambridge University Press: Cambridge, UK, 2001; pp. 55–60.
6. Solesbury, W. *Evidence Based Policy: Whence It Came and Where It's Going*; ESRC UK Centre for Evidence Based Policy and Practice, Queen Mary, University of London: London, UK, 2001; pp. 1–11.
7. La Caze, A.; Colyvan, M. *Evidence-Based Policy: Promises and Challenges*. 2006. Available online: <http://colyvan.com/papers/ebp.pdf> (accessed on 11 October 2019).
8. Watts, R. Truth and politics: Thinking about evidence-based policy in the age of spin. *Aust. J. Publ. Admin.* **2014**, *73*, 34–46. [[CrossRef](#)]
9. Mulgan, G. Government, knowledge and the business of policy making. In Proceedings of the National Institute of Governance Conference, Canberra, Australia, 23–24 April 2003.
10. Gray, J.A.M. *Evidence-Based Healthcare*; Churchill Livingstone: New York, NY, USA, 1997.
11. Sutcliff, S.; Court, J. *Evidence-Based Policymaking: What Is It? How Does It Work? What Relevance for Developing Countries?* Overseas Development Institute: London, UK, 2005.
12. Jennings, E.T., Jr.; Hall, J.L. Evidence-based practice and the use of information in state agency decision making. *J. Public Adm. Policy Res.* **2011**, *22*, 245–266. [[CrossRef](#)]
13. OECD. *Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information*; OECD: Paris, France, 2008; Available online: <https://legalinstruments.oecd.org/public/doc/122/122.en.pdf> (accessed on 10 October 2019).
14. Davis, P. Is evidence government possible? Paper presented at the 4th Annual Campbell Collaboration Colloquium, Washington, DC, USA, 19 February 2004.
15. Triantafyllou, P. The political implications of performance management and evidence-based policymaking. *Am. Rev. Public Adm.* **2015**, *45*, 167–181. [[CrossRef](#)]
16. Ferrandino, J. The enemy of teaching evidence-based policy: The Powell-Bush doctrine of public affairs. *J. Public Aff. Educ.* **2014**, *20*, 73–89. [[CrossRef](#)]
17. Jost, J.T.; Federico, C.M.; Napier, J.L. Political ideology: Its structure, functions, and elective affinities. *Annu. Rev. Psychol.* **2009**, *60*, 307–337. [[CrossRef](#)]
18. Esty, D.; Rushing, R. The promise of data-driven policymaking. *Issues Sci. Technol.* **2007**, *23*, 67–72.
19. Bradley, C.J.; Penberthy, L.; Devers, K.J.; Holden, D.J. Health services research and data linkages: Issues, methods, and directions for the future. *Health Serv. Res.* **2010**, *45*, 1468–1488. [[CrossRef](#)]
20. Curtis, L.H.; Brown, J.; Platt, R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. *Health Aff.* **2014**, *33*, 1178–1186. [[CrossRef](#)]
21. Zhang, Y.; Xu, Y.; Shang, L.; Rao, K. An investigation into health informatics and related standards in China. *Int. J. Med. Inform.* **2007**, *76*, 614–620. [[CrossRef](#)] [[PubMed](#)]
22. Chen, H.; Wang, F.Y.; Zeng, D. Intelligence and security informatics for homeland security: Information, communication, and transportation. *IEEE Trans. Intell. Transp.* **2004**, *5*, 329–341. [[CrossRef](#)]
23. Chen, H.; Wang, F.Y. Guest editors' introduction: Artificial intelligence for homeland security. *IEEE Intell. Syst.* **2005**, *20*, 12–16. [[CrossRef](#)]
24. Ku, C.H.; Leroy, G. A decision support system: Automated crime report analysis and classification for e-government. *Gov. Inf. Q.* **2014**, *31*, 534–544. [[CrossRef](#)]
25. Ministry of Science and ICT (MSIT); Korea Institute of Science and Technology Evaluation and Planning (KISTEP). *Implementation Plan for Survey of National R&D Program and Manual in 2017*; MSIT, KISTEP: Seoul, Korea, 2017. Available online: www.korea.kr/common/download.do?tblKey=EDN&fileId=212304 (accessed on 15 July 2019).
26. Oh, S.; Kim, S. R&D support for SMEs: Current status and performance analysis. *STEPI Insight.* **2018**, *224*, 1–30.
27. Yang, H.B. Status of technology commercialization of SMEs and policy tasks. *J.S.F.* **2017**, *37*, 23–45.

28. Presidential Advisory Council on Science and Technology (PACST). *National Innovation Plan*; PACST: Seoul, Korea, 2018.
29. Kohonen, T. Essentials of the self-organizing map. *Neural Netw.* **2013**, *37*, 52–65. [[CrossRef](#)]
30. Liu, Y.; Weisberg, R.H.; Mooers, C.N. Performance evaluation of the self-organizing map for feature extraction. *J. Geophys. Res.* **2006**, *111*, C05018. [[CrossRef](#)]
31. Reusch, D.B.; Alley, R.B.; Hewitson, B.C. North Atlantic climate variability from a self-organizing map perspective. *J. Geophys. Res.* **2007**, *112*, D02104. [[CrossRef](#)]
32. Mangiameli, P.; Chen, S.K.; West, D. A comparison of SOM neural network and hierarchical clustering methods. *Eur. J. Oper. Res.* **1996**, *93*, 402–417. [[CrossRef](#)]
33. Rutkowski, L.; Jaworski, M.; Pietruczuk, L.; Duda, P. The CART decision tree for mining data streams. *Inform. Sci.* **2014**, *266*, 1–15. [[CrossRef](#)]
34. Kohonen, T. Physiological interpretation of the self-organizing map algorithm. *Neural Netw.* **1993**, *6*, 895–905. [[CrossRef](#)]
35. Burrows, W.R.; Benjamin, M.; Beauchamp, S.; Lord, E.R.; McCollor, D.; Thomson, B. CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada. *J. Appl. Meteorol.* **1995**, *34*, 1848–1862. [[CrossRef](#)]
36. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).