CONFERENCE PROCEEDINGS/FULL PAPERS
ISBN: 978-625-99063-3-1/July 2023

31st RSEP International Conference on Economics, Finance and Business
22-23 June 2023, MERCURE Paris 19 Philharmonie La Villette, Paris, France

Analysis of R&D projects developed in science and technology parks through topic modeling

Onur Bilgin
Research Assistant, Department of Economics, Kirikkale University, Türkiye
E-mail: onurbilgin@kku.edu.tr

Haci Bayram Isik
Prof, Department of Economics, Kirikkale University, Türkiye
E-mail: hbayram@kku.edu.tr

DOI: https://doi.org/10.19275/RSEPCONFERENCES264

Abstract

To enhance companies' innovation performance, many new Science and Technology Parks (STPs) have been established in many cities in Türkiye in the last 30-35 years. The companies located in these areas develop technology through R&D projects. In this context, understanding the distribution of themes of R&D projects developed in STPs is an important factor in determining the success factors of STPs. This study analyses the R&D documents of companies using a type of unsupervised machine-learning technique called topic modeling. The dataset containing the project documents of the companies was obtained through the Entrepreneur Information System of the Ministry of Industry and Technology of the Republic of Türkiye. The results showed that digital technologies such as artificial intelligence, mobile technologies, and web applications were developed intensively in STPs with high R&D innovation performance.

Keywords: Science and technology parks, R&D projects, topic modeling
Jel codes: O30, O32

1. Introduction

In Türkiye, the idea of developing technology in specific regions was first expressed in government reports and plans in the 1980s. In the 1990s, STP-like structures started to be established and in the 2000s, STPs became an important tool for technological development. Today, there are 81 STPs in operation in Türkiye. The rapid increase in the number of STPs has naturally led to the questioning of their level of innovation. However, there is no study in the literature that analyzes the innovation performance of STPs and the content of technologies developed in STPs as a whole. In this context, the aim of this study is not only to find out the innovation levels of STPs but also to find out what technologies are developed in STPs and how to differentiate these technologies between efficient and inefficient SPTs.

To fulfill the purpose of the research, firstly, the innovation performance of STPs was determined using Data Envelopment Analysis (DEA) with data of 2021. As a result of DEA, STPs with an efficiency value of 1 are classified as efficient STPs, and STPs with an efficiency value less than 1 are classified as inefficient STPs. Efficiency analysis is the first stage of this paper, and then the information and findings of this stage are presented under the title "How is the innovation performance of STPs?". In the second stage of this paper, the technologies developed in these STPs were explored. For this purpose, the R&D project files developed between 2017-2021 in STPs were investigated by structural topic modeling. As a result of this analysis, it was determined how the content of R&D projects differed both as a whole and between efficient and inefficient STPs. The information and findings of the second stage are presented under the title "Which technologies are developed in STPs?".

2. How is the innovation performance of STPs?

According to the Ministry of Industry and Technology, there are 97 STPs in Türkiye. Of these STPs, 16 are under construction. The number of active STPs is 81. All variables need to be known for all STPs in order to conduct the DEA. Therefore, data for 2021 was used. In addition, not all active STPs are advanced enough to be included in the research. Although the results of innovation activities vary across sectors, they generally take many years. Leiponen (2005, p. 319) states that this period takes five years on average. For this reason, STPs established in 2016 and later were excluded from the scope of the study. In addition, 21 different STPs, which were established before 2016, were also excluded from the study, since they were underdeveloped to the extent that they violated the assumption of a similar structure of Decision Making Units (DMU). As a result of these limitations, analyses were conducted with the data of R&D firms in the remaining 37 STPs. Table 1 shows the inputs and outputs used in DEA.

Table 1. Inputs and outputs used in DEA

Inputs	**Abbreviation**	**Description**
Number of R&D Companies in STPs	IN1	Number of active R&D companies in STP in 2021
Total Number of R&D Projects	IN2	In 2021, the number of completed and ongoing R&D projects in STP
R&D Expenditure per Employee	IN3	Total R&D expenditures made by R&D companies in 2021
Outputs	**Abbreviation**	**Description**
Total Domestic R&D Income	OUT1	In 2021, the sum of R&D revenues (excluding exports) generated by R&D companies
Total Export R&D Revenue	OUT2	In 2021, the sum of R&D export revenues generated by R&D companies
Total Intellectual Property Application	OUT3	In 2021, total number of national patents, international patent, copyright (software), industrial design and utility model applications

Once DMUs and variables were selected, the first thing to check was the positive correlation between input and output variables. For DEA to give proper results, there should be a positive relationship between input and output variables (Kutlar & Bakırcı, 2018). The correlation matrix for the variables was presented in the table below. As can be seen from the table, there was no negative correlation between the variables.

Table 2. Correlation Matrix for DEA Variables

	IN1	IN2	IN3	OUT1	OUT2	OUT3
IN1	1					
IN2	0.82	1				
IN3	0.6	0.57	1			
OUT1	0.83	0.74	0.46	1		
OUT2	0.87	0.69	0.63	0.76	1	
OUT3	0.77	0.59	0.37	0.8	0.72	1

Finally, before DEA, descriptive statistics of the data and data sizes can be examined. Descriptive statistics of the variables were presented below.

Table 3. Descriptive statistics for DEA variables

	IN1	IN2	IN3	OUT1	OUT2	OUT3
Min:	18	7	18.244	4.642e+06	12.821	1
Q1:	66	26	94.414	6.311e+07	1.371.399	1
Median:	93	43	123.339	1.834e+08	3.871.625	3
Mean:	112,9	49,27	159.072	5.584e+08	32.087.745	13,46
Q3:	136	56	220.531	7.405e+08	18.053.200	19
Max:	357	167	447.662	2.900e+09	277.681.590	100

As seen from the descriptive statistics, the quantitative magnitudes of the data took values between 10^0 and 10^9. This difference indicated that the data set was unbalanced in terms of the magnitudes of the variables. The most practical way to eliminate this imbalance is to perform mean normalization. Each observation value of the variables is normalized to the mean of the series. After this process, DEA is expected to produce more reliable results (Zhu & Cook, 2007, p. 310).

Efficiency values were calculated using RStudio and the deaR (Coll-Serrano et al., 2023) and DJL (Lim, 2023) packages. The efficiency values were presented below.

Table 4. Results of DEA

STP	**Efficiency**
Ankara STP	0,98
Ankara Teknopark STP	0,47
Ankara Üniversitesi STP	0,78
Batı Akdeniz STP	0,31
Bilişim Vadisi STP	0,44
Boğaziçi Uni. STP	0,21
Celal Bayar Uni. STP	0,08
Çukurova STP	0,11
Cumhuriyet STP	0,12
Dokuz Eylül STP	0,20
Ege Teknopark STP	0,17
Erciyes Üniversitesi STP	0,17
Erzurum STP	0,09
Eskişehir STP	0,20
Fırat STP	0,52
Gazi Teknopark STP	1,00
Gaziantep Uni. STP	0,14
Göller Bölgesi STP	0,29
GOSB Teknopark STP	0,56
Hacettepe Uni. STP	1,00
İstanbul STP	1,00
İ.Ü. and İ.Ü. Cerrahpaşa STP	1,00
İTÜ Arı Teknokent STP	1,00

İzmir Bilim and Tek. Parkı	0,07
İzmir STP	0,24
Kocaeli Uni. STP	0,49
Mersin STP	0,28
Niğde Ömer H. Uni. STP	0,14
ODTÜ Teknokent STP	1,00
Ostim Ekopark STP	0,19
Samsun STP	0,15
Selçuk Uni. STP	0,24
Tokat STP	0,12
Trabzon STP	0,10
Tübitak-MAM STP	0,86
Ulutek STP	0,49
Yıldız Teknik Uni. STP	1,00
Average Efficiency	**0,44**

DEA was applied under the assumption of nonincreasing returns to scale (NIRS). Accordingly, the efficient STPs were Hacettepe University STP, Yıldız Technical University STP, ODTÜ TEKNOKENT STP, ITU Arı Teknokent STP, Istanbul STP, Gazi Teknopark STP, Istanbul University and Istanbul University-Cerrahpaşa STP. Moreover, the average efficiency value of STPs was 0.44. This value showed that there was an average inefficiency of approximately 56% in STPs.

3. Which technologies are developed in STPs?

The innovation performance of STPs was found with DEA. However, do the contents of the technologies developed in these STPs differ? To answer this question, the content of the R&D projects of firms can be examined through topic modeling analysis. The most important stage in text analysis or topic modeling studies is the pre-processing of the data. To be accurate in analysis, the untidy data must first be converted into a tidy format. The following figure summarizes the preparation process of the data set through an example sentence.

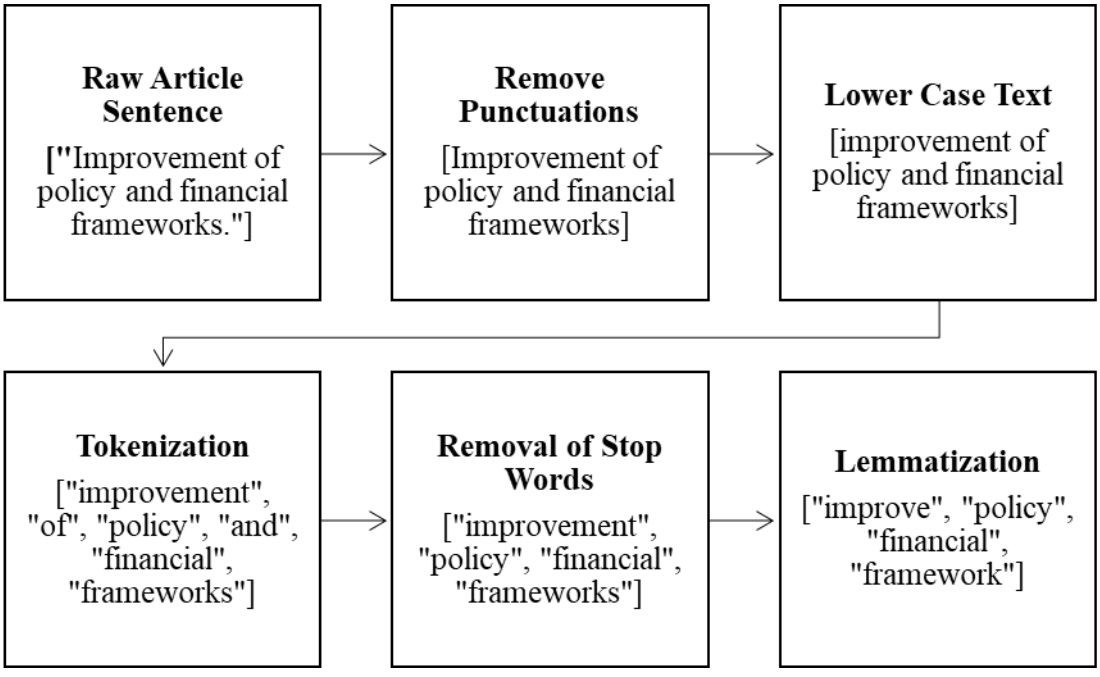

Figure 1. Text processing stages

Source: (Kumar & Ng, 2022, p. 214)

After the data set was prepared, the researcher first determined how many different topics would be researched. This study analyzed more than 23 thousand documents related to R&D projects. In a data set consisting of such a large number of documents, there may be tens or even hundreds of topics. However, as the number of topics increases, it is very difficult for someone who does not know much about the relevant terms to understand the

superstructure expressed by the terms related to the subject. Moreover, it is more important to understand the differentiation of technology levels and trends in STPs in general terms rather than micro-level topics. In this context, researchers can rely on their expertise to determine the number of topics, or they can decide directly by using various statistical methods. In this study, the number of topics was decided based on statistical criteria. In statistical terms, four criteria are generally calculated to determine the number of topics. These are held-out likelihood, semantic coherence, residual dispersion, and lower bound (Roberts et al., 2019).

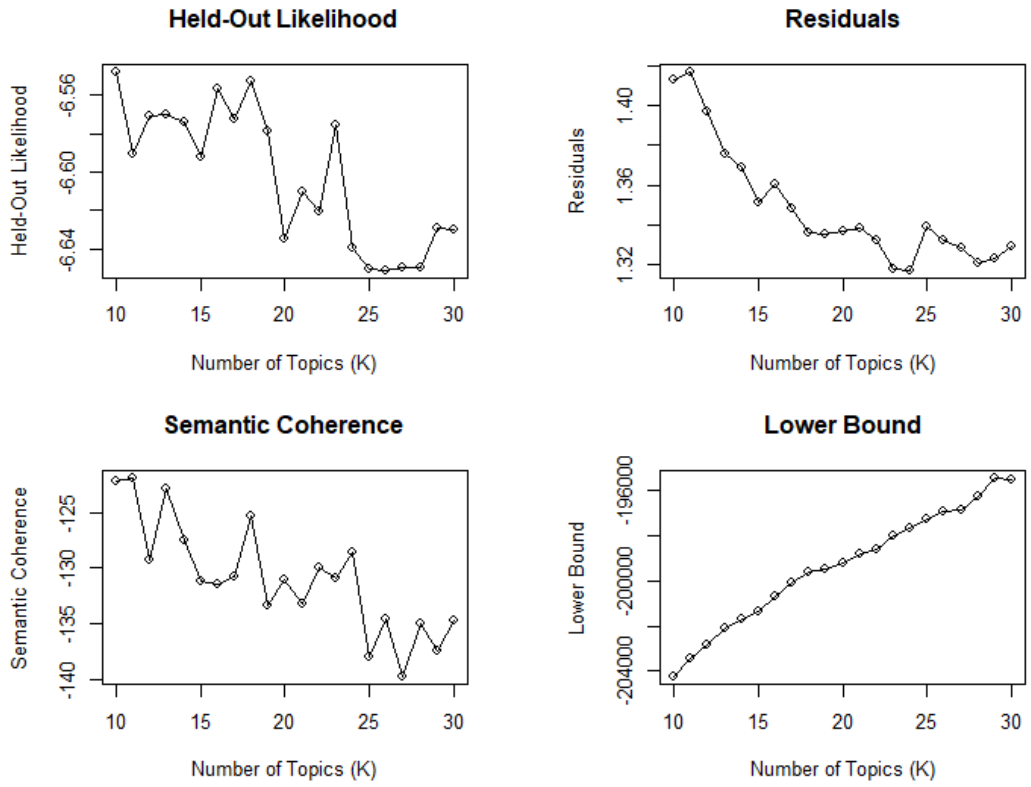

Figure 2. Various statistics for determining the number of topics

In order to find the model that produces the most semantically consistent and diverse topics, calculations were performed for a range of 10-40 topics. The 18-topic model yielded relatively high semantic coherence and held-out likelihood, but relatively low residuals and lower bounds (Figure 2). Analysis was performed on the dataset with 18 topics. The topics obtained as a result of the analysis were presented in Table 5.

Table 5. Topic labels, Key Terms and Proportion

No. Topic Label	Metric	Key Terms	Proportion
1. Health and Treatment	Prob:	health, treatment, patient, clinic, medical, diagnosis, patient's	4.53%
	FREX:	diagnosis, surgery, physician, bone, brain, patient, surgery	
	Lift:	anesthesia, antibodies, antibodies, infant, infants, surgery, doctor	
	Score:	treatment, patient, health, health, patient's, clinical, diagnosis, patients	
2. Process Tracking and Management	Prob:	management, module, processes, processes, of, companies, institutions, processes, of, processes	8.98%
	FREX:	stock, insurance, processes, business, processes, of, companies, processes	
	Lift:	masad, tramer, pension, agency, agencies, agencies, agencies, dask	
	Score:	management, accounting, module, institutions, stock, order, insurance	
3. Machinery and Manufacturing	Prob:	machine, quality, manufacturing, part, process, process, robot, process	5.17%
	FREX:	machine, part, of, machines, painting, robot, machine, mold	
	Lift:	boyahan, painting, dyeing, bending, mold, in, machine, machine, presses	
	Score:	manufacturing, part, press, mold, mold, plastic, machine, parts	
4. Cooling Systems, Electric Motors	Prob:	engine, electricity, vehicle, mechanics, motion, cooling, fuel	5.35%
	FREX:	battery, charging, motor, drive, gear, cooling, vibration	
	Lift:	gear, magnetized, bldc, masts, masts, axial, brushless	
	Score:	engine, electricity, cooling, charging, battery, vibration, fuel	
5. Communication	Prob:	device, measurement, communication, data, remote, hardware, electronics	6.08%
	FREX:	wireless, measurement, malfunction, rfid, devices, devices, sensors	

	Lift:	analyzer, modbus, meter, lorawan, meters, meters, wireless	
	Score:	communication, wireless, device, sensor, measurement, measurement, fault, data	
6. Online Education	Prob:	education, virtual, reality, learning, student, assessment, social	4.03%
	FREX:	student, students, education, exam, students, teacher, student	
	Lift:	secondary education, parents, educational, development, of, lessons, trainings, high school	
	Score:	education, students, student, exam, reality, teaching, learning	
7. Defense and Weapon	Prob:	air, defense, military, electronic, aselsan, systems, mission	6.35%
	FREX:	aselsan, flight, weapon, unmanned, radar, tactical, military	
	Lift:	operation, helicopter, hürkuş, aircraft, combat, ihas, of, ihas	
	Score:	defense, military, aselsan, air, flight, unmanned, radar	
8. Medicines and Pharmaceutical Ingredients	Prob:	drug, referan, substance, containing, active, formulation, in, treatment	1.90%
	FREX:	formulation, oral, zone, referan, analyses, of, drug, convenience	
	Lift:	will remain, guidelines, ease, of, zone, bioequivalence, dosage, pharmacoeconomic	
	Score:	drug, formulation, treatment, pharmacoeconomic, oral, zone, bioequivalence	
9. Payment and Billing Systems	Prob:	payment, electronic, credit, invoice, transactions, bank, transaction	5.90%
	FREX:	credit, jotform, tax, archive, bank, document, invoice	
	Lift:	lawyers, in, banks, bull, invoice, signed, by, law, taxpayers	
	Score:	payment, invoice, credit, bank, banking, jotform, identity	
10. Artificial Intelligence	Prob:	data, artificial, analysis, decision, intelligence, model, support	7.24%
	FREX:	learning, artificial, analysis, intelligence, algorithms, mining, models	
	Lift:	predicting, probabilities, unopro, panoramic, comparison, inference, intelligence	
	Score:	data, intelligence, artificial, learning, learning, machine, analysis	
11. Image Recognition	Prob:	smart, vehicle, image, security, security, traffic, human, recognition	4.98%
	FREX:	recognition, smart, traffic, city, parking, transportation, accident	
	Lift:	gerc, intersection, nic, will, be, beacon, planted, purpose	
	Score:	smart, vehicle, traffic, image, recognition, video, camera	
12. Cloud Computing and Cyber Security	Prob:	data, cloud, security, application, service, management, platform	6.78%
	FREX:	cyber, database, telecom, servic, cloud, will, cloud	
	Lift:	cicd, deliveri, improv, orchestration, result, build, increas	
	Score:	cloud, data, cyber, will, server, servic, crypto	
13. Food, Agriculture and Livestock	Prob:	food, agriculture, natural, plant, chemical, agricultural, biological	4.17%
	FREX:	plant, herbal, nutrient, cosmetic, fruit, fertilizer, agricultural	
	Lift:	plant, fertilizers, field, chicken, amino, amino, aroma, aromatic	
	Score:	plant, food, agriculture, vegetable, agricultural, nutrient, protein	
14. Mobile Application and Game	Prob:	mobile, app, game, digit, commerce, internet, online	7.18%
	FREX:	advertising, game, site, media, media, game, games, games of, games	
	Lift:	campaigns, advertising, ads, retent, instagram, game, in, games	
	Score:	game, mobile, advertising, internet, media, android, onlin	
15. Energy Production and Efficiency	Prob:	energy, electricity, solar, waste, environment, treatment, earthquake	4.69%
	FREX:	earthquake, renewable, coal, mining, energy, wastewater, concrete	
	Lift:	wastewater, lignite, thermal, treatment, earthquake, natural gas, hydrogeological	
	Score:	energy, electricity, waste, solar, waste treatment, earthquake, renewable	
16. Coating (Metal, Composite)	Prob:	material, surface, surface, metal, composite, coating, properties, of, materials	5.04%
	FREX:	coating, polymer, composite, material, material, of, nano, materials	
	Lift:	alloy, epoxy, coatings, carbide, material, fireproofing, alloys	
	Score:	composite, coating, metal, surface, steel, polymer, thermal	
17. Maritime and Ship	Prob:	ship, engineering, marine, package, build, quantity, ulutek	2.28%
	FREX:	ulutek, ship's, ship, ships, ballast, portal, maritime	
	Lift:	shipowner, ships, ship's, mission, class, pakistan, can, be, marketed	
	Score:	ship, marine, ballast, ship's, ulutek, portal, engineering	
18. Non-Technology	Prob:	sectord, value, needs, added, Türkiye, nation, sectord	9.34%
	FREX:	technok, developing, overseas, product, from, domestic, innovation	
	Lift:	accentur, our, experience, in, our, work, from, the, country, where, we, came, from	
	Score:	export, technok, firms, global, added, firms, sector	

The table above presents the words that best represent the topic based on the metrics Prob, FREX, Lift, and Score for each topic. The "Prob" metric represents the probability that a term is related to a particular topic. This metric reflects the probability that the term is an important part of a topic. For example, in topic 1, the term with the highest "Prob" metric is "health". This indicates that the term "health" is an important component of this structured topic and is associated with this topic with a higher probability than other terms.

The FREX (FRequency and EXclusivity) metric indicates the degree to which a term is related to a particular topic. FREX takes into account how often a term is used in a particular topic and how unique it is compared to other topics. In topic 1 above, the term "diagnosis" has the highest FREX value, indicating that it is an important part of this structured topic and has more uniqueness compared to other topics. For example, a person who hears the word “parity” will understand that the topic is related to “economy”. FREX metric tries to find such words.

Lift is a metric that measures how the likelihood of two terms being used together varies relative to the likelihood of those terms being used independently. That is, the Lift metric shows the relationship of one term to another term. High Lift values indicate that terms are more likely to be used together than they are to be used independently. This indicates that the terms are closely related and are used together in a meaningful way. For example, in the table above, the term "anesthesia" has a high Lift value with the terms "surgery" and "doctor's". This indicates that the term "anesthesia" is frequently used with the terms "surgery" and "doctor's" and that there is a strong relationship between these terms.

The score metric shows the likelihood of terms being keywords of a particular topic. For example, the term with the highest score in topic 1 is "treatment". As a result of all these metrics, it is understood that the relevant topic is related to health and treatment technologies. Considering these four metrics and examining the documents in which the topics were found intensively, each topic was labeled.

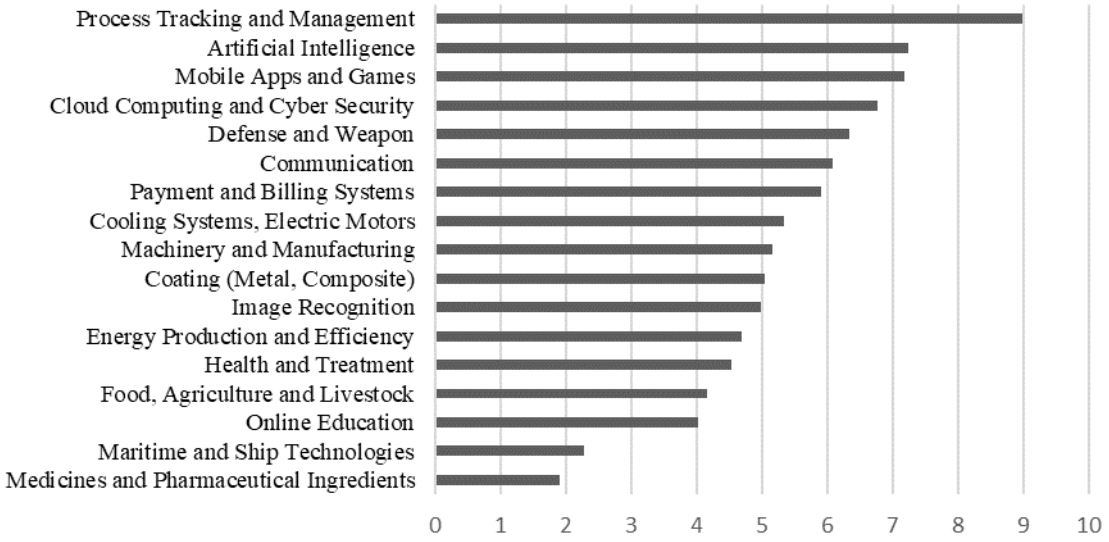

Figure 3. Top Topics

The first topic that stands out proportionally in the projects developed between 2017 and 2021 consists of terms such as "value, need, Türkiye, sector, industry, product, innovation", which are expected in almost all R&D projects. In this sense, topic 18, which corresponds to 9.34% proportionally, is not directly related to technology. Therefore, they are not included in the figure above. The other 17 topics identified as a result of the analysis are easily associated with specific technologies. In this context, the most common topic is related to process tracking and management. Artificial intelligence technology ranks second. Technologies related to mobile applications, web applications and digital games take the third place. Cloud computing and cyber security technologies take the fourth place. As can be seen, software technologies are among the most frequently researched and developed topics in STPs. This is quite normal for the STP ecosystem, which consists of approximately 50% of software companies. However, thanks to the topic modeling, we can see which technologies are developed more frequently by companies operating in the software sector.

In fifth place is the topic that covers words related to various military and defense industry technologies. In sixth place is the development of technologies that form both the hardware and software infrastructure of communication technologies. In seventh place is technologies related to payment and billing systems. In eighth place is cooling systems and electric motor technologies. In ninth place are manufacturing technologies, and in tenth place are coating technologies. In eleventh place are image recognition technologies, one of the application areas of artificial intelligence technology. In twelfth place are technologies for energy production and efficiency, and in thirteenth place are technologies for health and treatment. In fourteenth place is the development of technologies related to food, agriculture and animal husbandry. In fifteenth place are online education

technologies. The sixteenth ranked topic is related to maritime and ship technologies. The least common topic in R&D project files is related to pharmaceuticals and active substance production.

In Figure 3, the results of the topic modeling are presented without any distinction of year or innovation performance. In other words, it shows the topics or latent structures included in the R&D projects developed in all STPs. But, figure 4 shows that how the content of R&D projects different between efficient STPs and inefficient STPs.

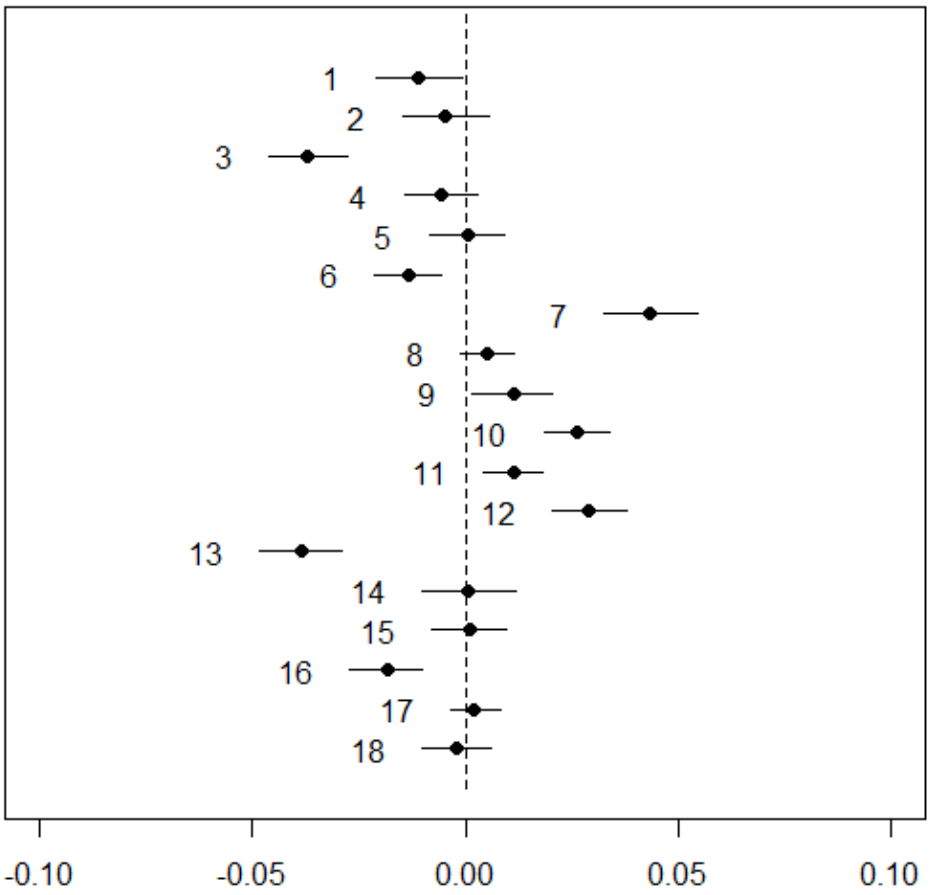

Figure 4. Difference in Topic Proportions

The figure consists of two parts. The further to the left of the dashed line, the more a topic is developed by firms in inefficient STPs. Alternatively, the further to the right of the dashed line, the more a topic is developed by firms in efficient STPs. As can be seen from the figure, the proportion of communication technologies (5), mobile application and gaming technologies (14), energy generation and efficiency technologies (15), maritime and ship technologies (17) and non-technology elements (18) in efficient and inefficient STPs is quite close to each other. On the other hand, food, agriculture and animal husbandry related technologies (13), manufacturing technologies (3) and composite and coating technologies (16) are significantly present in the R&D projects of firms in inefficient STPs. The R&D projects in which defense and arms industry technologies are most common belong to firms in efficient STPs. Other common technologies in efficient STPs are cloud computing and cyber security technologies (12), artificial intelligence technologies (10), image recognition technologies (11), and technologies related to payment and billing systems (9).

As can be seen, there is a difference between the topics of R&D projects in terms of efficient and inefficient STPs. The proportional changes of the relevant technologies in the last 5 years are presented in the graphs below on the basis of STP efficiency value. Figure 5 shows the change between 2017 and 2021 in the technologies that are more common in inefficient STPs. The red line shows the ratio of the relevant technology in inefficient STPs, while the blue line shows the ratio in efficient STPs. Shaded areas show the 95% confidence intervals of the respective lines. The top left graph shows the food, agriculture, and animal husbandry related technologies (13). According to this graph, it is more common in the projects of inefficient STPs and has increased at a similar rate in the projects of efficient and inefficient STPs in the last 5 years. The lower left graph shows the topic related to manufacturing technologies. While this topic has remained almost constant in the projects of efficient STPs, it has decreased in the projects of inefficient STPs over the last 5 years. The top right graph

shows the topic related to composite and coating technologies (16). This topic is slightly more common in inefficient STPs over the years. The bottom right graph shows the topic related to online education (6). According to the graph, while it remained almost constant in the last five years in inefficient STP, the proportion of this topic in projects started to increase in efficient STPs, especially with the onset of the Covid-19 pandemic. However, approximately 2 years later, it is observed that the interest of companies in efficient STPs in this topic decreased.

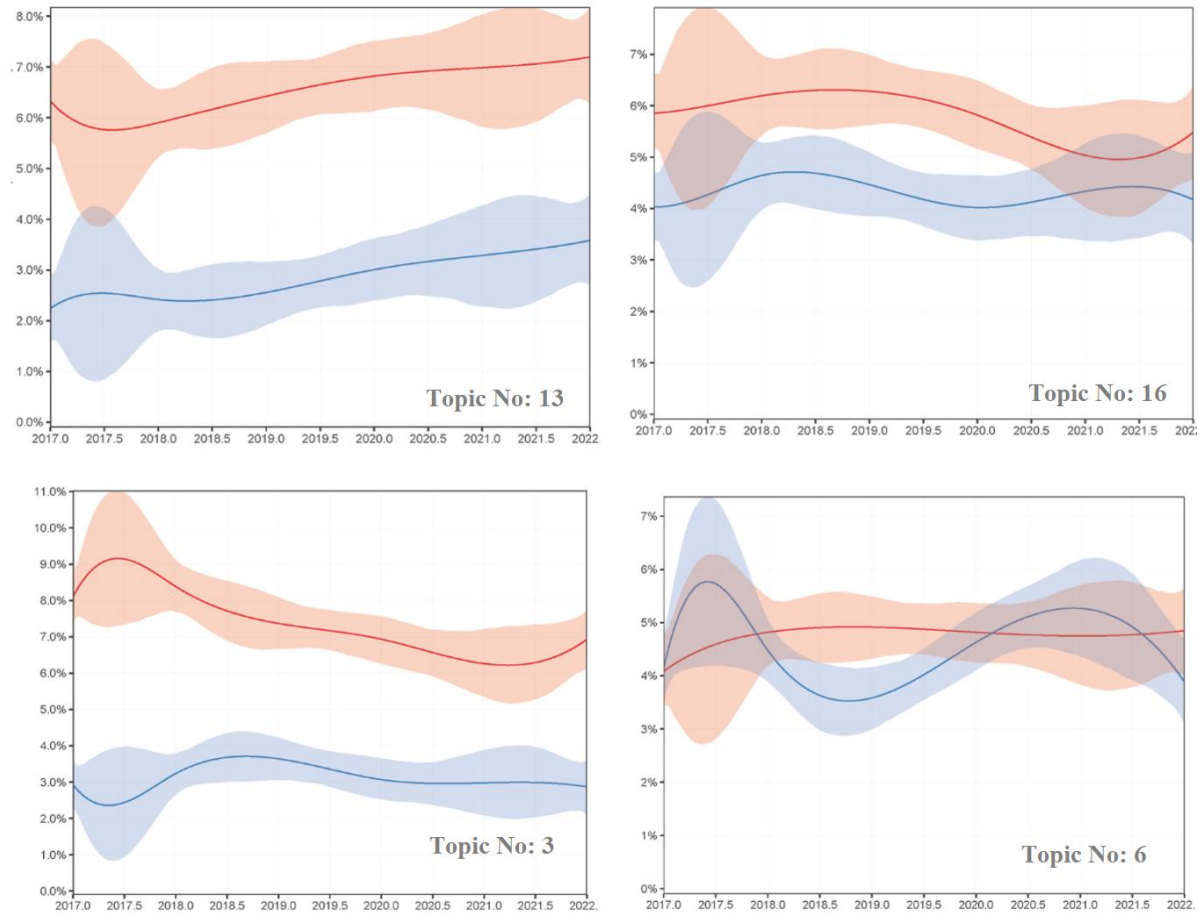

Figure 5. Topics that are more common in inefficient STPs

Figure 6 shows the proportional change between 2017 and 2021 in the technologies that are more common in efficient STPs. On the upper left side of the figure, there is a graph of the topic related to defense and weapon technologies (7). According to the graph, companies in efficient STPs have developed R&D projects on this topic more intensively in the last 5 years. On the bottom left side, there is the topic of payment and billing systems (9). While this topic was similar in both STP groups in 2017, it slowly increased in efficient STPs over the years. It slowly decreased in inefficient STPs. In the upper right part of the figure, there is the topic of artificial intelligence (10). Accordingly, this topic tends to increase for both STP groups, but it was observed more intensively in efficient STPs in all years. The bottom right graph shows the topic of cloud computing and cyber security (12). In recent years, the two STP groups have come closer to each other on this topic.

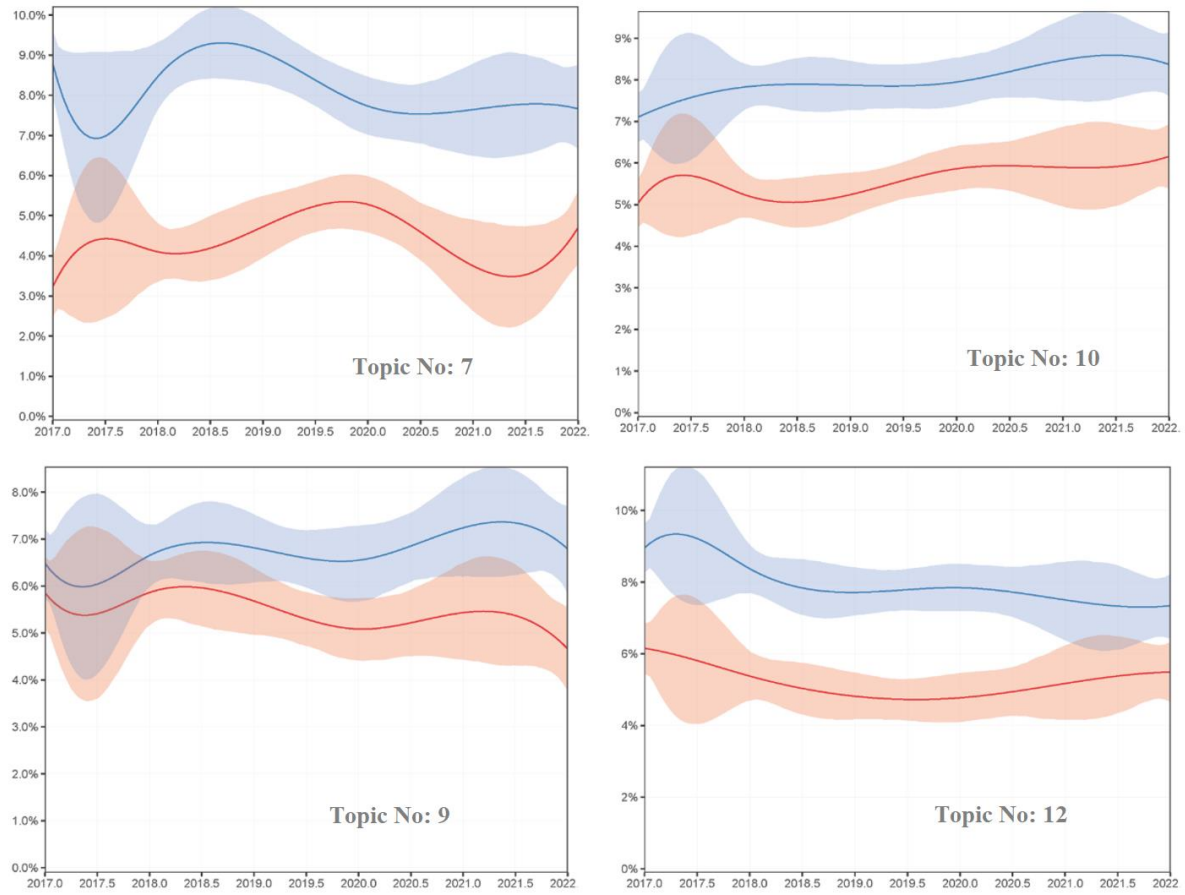

Figure 6. Topics that are more common in efficient STPs

In this part of the analysis process, the topics covered by R&D projects were identified using structured topic modeling, a type of unsupervised machine learning technique, and the differentiation of topics across STPs was analyzed. In inefficient STPs, topics covering more traditional technologies such as food, agriculture, animal husbandry, composites, coatings, engines and mechanics were found to be prominent, while topics related to more knowledge-intensive technologies such as artificial intelligence, image processing, cyber security, cloud computing and defense industry were found to be prominent in efficient STPs.

4. Conclusion

In this study, the efficiency of STPs and the content of R&D projects developed in these STPs were analyzed as a whole. As a result of DEA, it was concluded that 7 STPs are efficient producing innovation. And also, this paper showed that R&D projects developed in efficient STPs focused on more knowledge-intensive technologies such as artificial intelligence, cloud computing and defense technologies. On the other hand, in inefficient STPs, topics related to more traditional technologies such as agriculture, animal husbandry, composites and machinery were more prominent.

References

Coll-Serrano, V., Bolos, V., & Suarez, R. B. (2023). *deaR: Conventional and Fuzzy Data Envelopment Analysis.* Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/deaR/index.html

Kumar, M., & Ng, J. (2022). Using text mining and topic modelling to understand success and growth factors in Global Renewable Energy projects. *Renewable Energy Focus*, *42*, 211–220. https://doi.org/10.1016/j.ref.2022.06.010

Kutlar, A., & Bakırcı, F. (2018). *Veri Zarflama Analizi: Teori ve Uygulama*. Orion Kitabevi.

Leiponen, A. (2005). Skills and innovation. *International Journal of Industrial Organization*, *23*(5-6), 303–323. https://doi.org/10.1016/j.ijindorg.2005.03.005

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm : An R Package for Structural Topic Models. *Journal of Statistical Software*, *91*(2). https://doi.org/10.18637/jss.v091.i02

Zhu, J., & Cook, W. D. (2007). Modeling Data Irregularities and Structural Complexities in Data Envelopment Analysis. *Springer EBooks*.