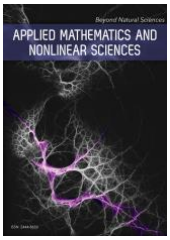
sciendo Applied Mathematics and Nonlinear Sciences, 8(2) (2023) 3521-3544



Applied Mathematics and Nonlinear Sciences

https://www.sciendo.com

Unlocking the secrets of Spain's R&D subsidies: An advanced analysis of applicant companies

Mónica Espinosa-Blasco[1], Gabriel I. Penagos-Londoño[2], Felipe Ruiz-Moreno[3], María J. Vilaplana-Aparicio[4,†]

1. Department of Financial Economics and Accounting, University of Alicante, Alicante, Spain
2. Department of Economics, Pontificia Universidad Javeriana, Bogotá, Colombia
3. Department of Marketing, University of Alicante, Alicante, Spain
4. Department of Communication and Social Psychology, University of Alicante, Alicante, Spain

Submission Info

Communicated by Z. Sabir
Received March 29, 2023
Accepted July 4, 2023
Available online October 31, 2023

Abstract

Innovation is crucial for companies to stay competitive, provide value to customers, and generate profits. Likewise, research and development (R&D) is critical for companies to sustain productivity growth. Spain has lagged behind other countries in terms of R&D investment, with only 1.4% of its GDP allocated to R&D, well below the European average. To improve this situation, the government offers subsidies to stimulate R&D in Spanish companies. This study examines the profile of subsidized companies in Spain. The aim is to provide insight into the support for companies that apply for innovation subsidies by analyzing the profile of subsidized companies and identifying key variables influencing the success of obtaining innovation grants. The study is based on advanced estimation methods. Natural language processing (NLP), artificial neural network (ANN) techniques, and clustering are used to perform rigorous and robust analysis of the profile of subsidized companies in Spain. The study thus contributes to knowledge in the field of innovation subsidies

Keywords: Innovation subsidies; public funds; R&D; innovation strategy; natural language processing; neural network; finite mixture model
AMS 2020 codes: 00000

†Corresponding author.
Email address: maria.vilaplana@ua.es

ISSN 2444-8656
https://doi.org/10.2478/amns.2023.2.01144

1 Introduction

Innovation is crucial for companies aiming to remain competitive and provide value to customers. It involves improving existing elements, ideas, or protocols or creating new ones that positively affect companies' market performance while generating profits. Therefore, companies must ensure that their ideas provide value or benefits once implemented. Every year, thousands of companies strive to differentiate themselves by offering increasingly complex and unique products. Research and development (R&D) is a crucial tool to sustain productivity growth [1-3].

Investment in R&D is substantial in many countries worldwide. Respectively, the United States, Japan, and China allocate 3.45%, 3.26%, and 2.40% of their gross domestic product (GDP) to R&D. In Europe, Sweden, Austria, Belgium, and Germany are leading investors in R&D, with investment ranging from 3.13% to 3.35% [4]. In contrast, Spain's investment is concerning. Spain invests only 1.4% of GDP in R&D. This figure is well below the European average of 2.27%. Furthermore, the number of companies in Spain conducting R&D has decreased over the last decade, mostly in the small and medium-sized enterprise (SME) segment [5].

Spain's situation is gradually improving, thanks in part to funds from the EU's Next Generation program. This program prioritizes innovation in Spain's Recovery, Transformation, and Resilience Plan. However, official figures from La Vanguardia [6] reveal that R&D investment increased by only 0.03% from 2020 to 2021. This minimal increase highlights the need for further efforts in this area. Another issue is the distribution of innovative companies in Spain. Such companies are not evenly spread across the country, with four of 19 Spanish regions accounting for over 70% of innovative start-ups. Therefore, the government is justified in prioritizing the creation of technology-based companies that focus on R&D. The reason is that such companies have been found to have a positive effect on the Spanish economy. The scientific literature indicates that subsidies play a key role in stimulating R&D in Spanish companies [7-13].

Spain has an extensive support system for businesses. This system is based on direct and indirect assistance. This assistance includes nonrefundable aid, subsidized loans, tax deductions, and social security rebates [5]. The Centre for the Development of Industrial Technology (CDTI) is a public entity operating under the auspices of the Ministry of Science and Innovation. It allocates a large amount of direct R&D and innovation aid. In 2022, the CDTI planned to launch R&D and innovation projects worth 2,469 million euros. In total, 1.366 billion euros were specifically allocated for grants and loans, and approximately 1.103 billion euros were made available from international programs such as Horizon Europe [14].

The NEOTEC program is one of the CDTI's most important grants. It has been funded by Next Generation EU funds in recent calls for proposals. This program is designed to support new business projects based on the use of technologies, knowledge derived from research, or technology development. NEOTEC targets small, young, innovative companies that develop technology that can be applied to companies from any region or sector in Spain. It uses a competitive concurrence selection process. Such a process means that all projects are evaluated but only those with high scores receive support.

The main aim of this study is to provide comprehensive analysis of the profile of subsidized companies. This aim is motivated by the importance of promoting the creation of innovative companies in Spain. Despite the relevance of this topic, research on this specific issue in Spain is scant. Advanced estimation methods were employed to provide accurate results and thus fill this gap in the literature. Natural language processing (NLP) techniques were used. Specifically, topic modeling was applied to transform the titles of submitted projects into two additional variables that

were added to the set of regressors. A neural network was used to measure the importance of the predictive variables. This neural network led to the selection of the most important variables to cluster subsidy applicants. Overall, these methods provide a robust, rigorous approach to investigate the profile of subsidized companies in Spain. The study thus contributes to the knowledge on innovation subsidies.

2 Literature review

Innovation drives economic growth, provides solutions to the challenges facing society, and aids competitiveness in advanced economies [15-16]. It is so important that it is included in the United Nations Sustainable Development Goals (SDGs) [17]. Consequently, promoting investment in innovation through R&D spending by private companies has become a major goal of innovation policies worldwide. Public subsidies can encourage higher private investment by firms, leading to more skilled employment [18, 19], productivity, exports, and public-private partnerships [20]. Public support for R&D is crucial given the market failures that result in a gap between the social and private benefits of R&D [21] and the issues surrounding innovation appropriation [22]. Innovation can be costly and uncertain, making it unappealing to firms. Thus, public agencies finance R&D projects that would not otherwise be carried out, particularly in critical or strategic areas for society. According to Hall and Lerner [23] and Gök and Edler [24], innovation policies should encourage investment, even though they may involve costs, particularly regarding appropriation failures. Kiman and Jongmin [25] showed that public financial support alone cannot fully mitigate market failure beyond a certain level of public intervention. Nevertheless, public financing programs aim to impact the learning processes and innovation capacity of companies [26].

Most countries encourage private R&D through support programs with diverse objectives. Blanes and Busom [27] revealed some common features of the companies that participated in these programs. For instance, mostly smaller and younger firms were interested, suggesting that company size and age are key factors. Pisár et al. [28] also found that funding was most effective in younger firms in a sample of Slovakian companies. Participation is positively associated with previous R&D experience. Therefore, defining the public sector evaluation criteria and selection methods of R&D projects and funding procedures is crucial.

Numerous studies have analyzed the effectiveness of R&D programs and the impact of public support on private investment and innovation behavior [19, 29-33]. However, these studies have provided inconclusive results due to factors such as national or regional context, sector, application, design of specific instruments, and country development status [34-37]. Montmartin et al. [38] showed that the regional impact of public R&D funding depends on policy design and local economic setup. The importance of industry in innovation is also debated in the literature. Some studies suggest that firms in the same sector have similar patterns of innovation activities [39-42]. Others have shown considerable differences within sectors. Hence, sector has a limited ability to explain differences between firms' innovation behaviors [43-45].

The literature has extensively examined how innovation contributes to firms. Studies have evaluated the effects of incentivizing R&D on variables such as skilled employment, tangible assets, and R&D investment [18, 19]. Research has shown that public R&D subsidies stimulate R&D spending, especially in small firms [46]. High-tech SMEs with patents already meet some requirements to secure grants covering future R&D innovation activity [47]. Previous patent activity has been observed to increase the likelihood of participation in funded projects in the UK [48]. However, public subsidies do not appear to affect the probability of patenting or the number of patents a firm would have filed in the absence of such funding.

The selection of beneficial projects for public funding has been widely studied. Governments have used varying criteria [49-51]. The selection of R&D projects for subsidies is complex given the diverse objectives and stakeholders involved. Santamaría et al. [52] studied the selection criteria used by the Spanish government for cooperative R&D projects. They analyzed factors determining project selection and funding program goals, including why two financial tools (loans and grants) were implemented within the same call for proposals. They found that project type was essential in the selection process and concluded that differences depended on sector and the year of the call. Acosta Ballesteros and Modrego Rico [53] and Blanes and Bosum [54] focused on the decisions of companies when applying for R&D grants. They sought to identify key variables in the selection process of cooperative R&D projects.

According to the literature, several company characteristics are related to public funding of R&D. For example, Afcha [54] identified the determinants of R&D innovation strategies that influence the receipt of R&D subsidies. They used a sample of Spanish firms from 1998 to 2005, finding that R&D efforts in previous years, technological cooperation, foreign capital, exports, and hiring of qualified personnel determine access to public subsidies. In Germany, Cantner and Kösters [55] used logistic regression to study R&D subsidy allocation to start-ups. They found that start-up capital and the working team had the most influence on securing public funds.

Company size has been widely studied in relation to public R&D funding. However, results are mixed. Some studies suggest that firm size positively affects the receipt of public funds. However, this finding seems to contradict the aim of many programs that seek to support SMEs. Some authors have concluded that firm size does not influence access to subsidies or is not significant in regional support [56, 57]. In contrast, several studies suggest that there are significant differences in the amount of public funding received by firms of varying sizes [7, 10, 12, 58, 59, 60]. According to Vila et al. [59], large companies benefit the most from public aid for innovation. Segarra and Teruel [56] reported that younger firms tend to receive more aid and that larger firms are less likely to receive funding [61]. Additionally, agencies tend to favor smaller firms [27]. However, Almus and Czarnitzki [62] suggested that firm size, along with other factors such as the existence of an R&D department and a foreign presence, positively affects the ability to attract public funds. Mardones and Zapata [63] found that firm size determines the receipt of public funds for innovative activities. They used both pseudo-panel and cross-sectional data, finding that company size was also related to accessing subsidies at different levels (regional, national, or European). García and Afcha [58] found that central government funding is typically given to larger companies, whereas regional funds tend to target SMEs. Blanes and Busom [27] argued that regional and central administrations have different objectives in terms of innovation policies. From the perspective of regional or central R&D subsidy programs, Gao et al. [64] suggested that local governments interact more efficiently and effectively with subsidy recipients, leading to superior performance of local R&D subsidy programs in China.

Receiving past grants may influence the chances of obtaining new R&D funding, according to several studies. Duguet [65] found that the debt ratio, ratio of private R&D investment to sales, and past public aid increased the likelihood of receiving a subsidy. Antonelli and Crespi [66] also noted that prior funding increased the probability of securing additional funding. Experience with R&D projects was found to increase the amount received, according to Duch-Brown et al. [67]. Hussinger [68] used two-stage selection models to show that grant award history was among the variables used by the German government to identify the most suitable candidates for funding.

3 Data collection and sample characteristics

The study used data on 516 Spanish companies that applied for R&D subsidies between 2018 and 2019 [69]. Of these companies, 157 were granted subsidies worth a total of 37,886,242 euros. The remaining 359 companies did not receive any subsidy due to noncompliance with requirements, low innovation project scores, or budget constraints. The study examined the NEOTEC Program. This program provides public subsidies for new business projects of innovative companies. It is managed by the CDTI. The study period spanned the period of economic growth in Spain just before the crisis caused by the COVID-19 pandemic.

The final resolution of the NEOTEC Program is published by CDTI during the year following the application for subsidies, which in our case corresponds to the years 2019 and 2020. This resolution shows both approved and rejected applications. It also gives access to valuable information regarding applicant companies. This information includes company name and identification, a short description or title of the presented innovation project, the result of the grant (approved or rejected), the amount granted to approved projects, and the reasons for rejection. The basic information extracted from this publication was used to create a list of applicants. This list was cross-checked with the SABI financial database (Sistema de Análisis de Balances Ibéricos; https://sabi.bvdinfo.com) to provide to additional data on the applicants.

Numerous variables were collected for each company, including a brief description of the innovation project in the form of the project title. NEOTEC subsidies are intended to support innovative projects for young companies. Therefore, the variable Age_{it} was included to reflect the age of the company in days. $Profits_{it}$ represented the company's profits in year t. $Assets_{it}$ referred to the total assets of firm i in year t. $Pp\&e_{it}$ represented the value of property, plant, and equipment or long-term tangible assets of company i in year t. To capture the intangible assets of the company, $Brands_{it}$ captured the number of registered brands owned by company i in year t. $AggVal_{it}$ indicated the value added of company i in year t. $Totequ_{it}$ referred to the value of assets less liabilities of company i in year t. The variable $Internat_{it}$ captured whether company i was operating outside its home country borders. Finally, $Activity_{it}$ was the general industrial classification of company i in year t (1 for service companies and 0 otherwise). These 10 features, together with the project title, were chosen as explanatory variables. The dependent variable was a dichotomous variable that took the value 1 if the innovation project was approved and 0 otherwise.

The final data set consisted of $n = 516$ observations and $p = 12$ variables. The data set consisted of standardized numerical, quantitative data and qualitative, categorical data. The data set also included unstructured data corresponding to the titles of the documents. These unstructured data were converted into structured data.

4 Method

This section describes the estimation techniques used in this study. The primary goal of this method was to identify the latent cluster structure of the data. Specifically, the aim was to detect K groups. The first step was to process the unstructured data extracted from the innovation project titles to create a new variable in the data set. This step required transformation of the unstructured data into structured data using NLP techniques. The chosen technique was topic modeling. After preparing the data, a feature selection approach reduced the number of predictors because some variables might not have been relevant to this aim. A filtering method was developed based on an artificial neural network (ANN).

Topic modeling natural language processing

The corpus of documents was used to draw a vocabulary of terms.[1] These documents were regarded as being composed of a fixed number, *K*, of topics. These topics were defined as probability distributions over the vocabulary. Each document in the corpus contained each topic in a given proportion. Formally, a hidden variable model was assumed, where the observed data (i.e., the words in documents) were driven by hidden random variables (i.e., the topics) [70].

The relationship between documents and topics was mediated by a probabilistic generative model. This model was a topic model, from which the documents were created. Each word in a document was chosen in a double random experiment. First, a topic was selected at random from a distribution over the topics. A word was then drawn at random from that topic. The order or position of the words was not considered in this process. This assumption is known as the "bag-of-words" assumption.

Consider a corpus of *D* documents generated from *T* topics. Each document d comprises N_d words, giving a total of $N = \sum_d N_d$ words out of a vocabulary of *W* words. The aim in this study was the inverse problem: to determine the topics. Hence, the aim was to determine the probability distributions β over the words that generated the documents, as well as the probability distribution θ of the topics over the documents. In simpler terms, the aim was to recover these two probability distributions from the data.

The Latent Dirichlet Allocation (LDA) model [71] was developed as an extension of the probabilistic Latent Semantic Indexing method proposed by Hofmann [72]. In LDA, the topic distribution over documents, θ, is assumed to follow a Dirichlet distribution, which is a conjugate prior for the multinomial distribution. LDA is a mixed-membership model, allowing for any document to be a combination of more than one topic, in contrast to a traditional mixture model [70]. The Dirichlet density distribution is given as follows:

$$p\,(\theta|\alpha) = \frac{\Gamma(\sum_{j=1}^{T}\alpha_j)}{\prod_{j=1}^{T}\Gamma(\alpha_j)}\prod_{j=1}^{T}\theta_j^{\alpha_{j-1}}$$

Here, Γ is the gamma function. The hyperparameters α_j, $\alpha = (\alpha_1, \ldots, \alpha_T)$ can be interpreted as the number of times a topic j is sampled in a document before any observation is made (i.e., a prior pseudocount) [73]. A symmetric Dirichlet distribution is achieved when α_j is a constant value α.

The Dirichlet distribution is a multivariate generalization of the beta distribution. It is defined on the $T-1$ simplex, meaning that the input variables $\theta_j \epsilon\, \lfloor 0,1 \rfloor$ must sum to one, $\sum_{j=1}^{T}\theta_j = 1$. This property is consistent with the interpretation of θ_j as a probability. The simplex contains all the probability distributions over the topics or words depending on the underlying distribution, as its points. Thus, the domain of the Dirichlet distribution, which is the set of points on the simplex, consists of discrete probability distributions. These distributions can be either the distribution of the topics over the documents or the distribution of the words in the topic.

The topic distributions (i.e., the discrete distributions of words in each topic) are modeled as Dirichlet distributions $\beta_t, t \in \{1, \ldots, T\}$, with hyperparameter $\delta = (\delta_1, \ldots, \delta_w)$. Meanwhile, the proportions of topics on the documents are modeled as Dirichlet distributions $\theta_d, d \in \{1, \ldots, D\}$, with hyperparameter $\alpha = (\alpha_1, \ldots, \alpha_T)$ [74].

[1] A word was defined as a sequence of letters from a given alphabet, a corpus as a collection of documents, and a document as a set of words (a "*bag of words*").

The topic model has an intuitive geometric interpretation [73]. For W words in the vocabulary, consider a W-dimensional Euclidean space in which each axis encodes the probability of observing each word. As stated earlier, each point in the $W-1$ simplex represents a probability distribution, in this case over the words, with the simplex containing all probability distributions of this kind.

For example, consider a vocabulary of $W=3$ words, as shown in Figure 1. The 2-simplex is represented by a triangle intersecting the three axes at 1. Each point within this triangle is a probability distribution over the three words of the vocabulary, and all possible distributions are contained in the simplex. Similarly, a topic is a point on the simplex. Any document generated by the model is a convex combination of the T topics contained in both the $W-1$ simplex and the $T-1$ simplex expanded by the T topics. When the number of topics is $T=2$, the space expanded by the topics is a segment with the two topics as its endpoints (i.e., a 1-simplex). This segment is contained in the triangle, a larger 2-simplex. In this context, the hyperparameters δ are interpreted as *forces* that move the topic locus away from the corners of the simplex.

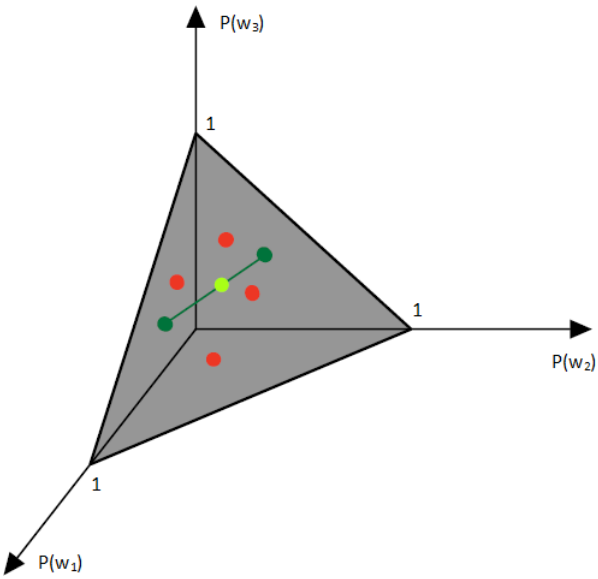


Figure 1. Topic model geometric interpretation

Note: The shaded area within the triangle, which is the 2-simplex, contains all possible probability distributions represented by points for a 3-word vocabulary. The red points in this figure represent examples of documents, whereas the green ones are two topics. The space expanded by the topics is a 1-simplex, which is the green segment on the 2-simplex. The points on the green segment, such as the light green one, represent a document generated with the help of the two topics.

To set up the LDA model, the number T of topics must be specified a priori. However, this value is unknown in advance. Therefore, techniques are used to discover this hyperparameter. Typically, to determine T, the pairwise dissimilarity between topics for a given T is calculated. The optimal number of topics $\hat{T}$ is the one with the maximum overall level of dissimilarity [75]. A heuristic was derived to determine the optimal value $\hat{T}$, given by:

$$\hat{T} = arg\max_{T} \frac{1}{T(T-1)} \sum_{(t,t')\in TOP_T} D\ (t\|t')$$

Here, T represents the hyperparameter counting the number of topics, t and t′ are any two topics from the set Top_T of T topics resulting from LDA, and D is a measure of information divergence. The number $\hat{T}$ of topics resulting from LDA is the one that maximizes the overall dissimilarity according

to D [75]. The Jensen-Shannon divergence, which is a symmetric version of the Kullback-Leibler divergence, is commonly used as a dispersion measure in this methodology. It is defined as follows:

$$D(t\|t') = \frac{1}{2}KL(t\|s) + \frac{1}{2}KL(t'\|s)$$

Here, KL is the Kullback-Leibler divergence, and $s = {}^{(t+t')}/_{2}$.

Cao et al. [76] proposed another helpful method to determine the optimal number of topics. A stability measure is introduced for a topic structure with *T* topics. The average cosine distance is defined as follows:

$$D(T) = \frac{2}{T(T-1)}\sum_{t\neq t' \in Top_T} corr\ (t, t')$$

In this study, a cluster-based method was used to determine the optimal number of topics, T. The topics were considered as semantic clusters. Intra-cluster similarity should be large, indicating that the topic had specific meaning. In contrast, inter-cluster similarity should be small, suggesting that the structure was stable [76]. The method was performed iteratively until the average cosine distance and model cardinality converged. Model cardinality, denoted as $Cardinality(Top_T, n)$ for a topic model Top$_T$, is the number of topics in Top_T contained in the open ball of radius n with respect to the average cosine distance.

Filtering methods

The aim in this study was to identify the latent clusters that partition the entire set of *n* observations into subsets. The data set consisted of *p* variables. Some of these variables may have low or no association, either linear or non-linear with the output. Therefore, they are uninformative as predictors.

Some models perform poorly with superfluous variables, and others perform poorly with correlated variables [77]. A variable should be excluded if it does not provide information (i.e., it is uninformative) as a predictor or is not related with the output. Reducing the complexity of the model is another reason to exclude a variable. In general, a trade-off exists between the predictability and interpretability of a model. To achieve a meaningful subset of variables, the number of variables should be reduced.

Some models such as LASSO can implicitly select variables. However, in most cases, an external procedure is required to reduce the number of variables before processing them. In filtering methods, a supervised selection of features leads to a set of core variables used to fit the learning model. This search is performed once. The resulting variables are then fed into the processing model. Although these methods are fast and capture large trends, they may potentially select more predictors than necessary. This selection of too many predictors results in *false positives*, meaning that some predictors are not strongly related to new data [77]. An ANN is recommended as a filter to address these shortcomings. An ANN was estimated using cross-validation, which reduced overfitting and the occurrence of false positives.

Artificial neural networks

An ANN is a mathematical model that resembles the way the brain works. In an actual neuron, information is transmitted away from the neuron through the axon and is received by other neurons through the dendrites. The information flows from neuron to neuron across the *synapses* [78]. In

artificial intelligence, an ANN is a network of nodes (*neurons*) interconnected by weighted links (*synapses*) [79, 80].

An ANN is defined by its topology. In a feedforward neural network, the model is arranged in layers of nodes interconnected by links. The input layer has as many neurons as input variables. The output layer has as many neurons as output variables. The remaining layers are the hidden layers, which are not directly connected to the exterior [79]. The information flows from the input layer, which receives the input variables, through the hidden layers to the output layer without feedback connections. In general, a neural network with multiple hidden layers is said to be *deep*. It is called a deep feedforward network.

In an ANN, each neuron is a function of incoming signals from previous neurons. These signals are combined as an average of the input values weighted by the link's proportions. The resulting value is then fed into a nonlinear activation function that outputs the *activation* values [81]. The activation values for the input layer are the ANN inputs. The McCulloch-Pitts network uses binary signals, the activation function only outputs zeros or ones, and the links are unweighted. In 1958, psychologist Frank Rosenblatt introduced the perceptron. This innovation expanded the kinds of problems that can be dealt with by incorporating weighted links. A neural network with more than one hidden layer is called a multilayer perceptron (MLP) [82].

The most common nonlinear activation functions *g* are the Sigmoid functions:

$$g(z) = \frac{1}{a+e^{-z}}$$

the Hyperbolic tangent:

$$g(z) = \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}$$

and ReLU (rectified linear unit):

$$g(z) = ReLU(z) = \max(0, z)$$

The latter is the most commonly used for deep learning because it is the most efficient for computations given its simplicity.

In a single hidden layer ANN, as depicted in Figure 2, the inputs to a neuron $\{x_1, \dots, x_s, \dots, x_p\}$ are averaged with the weights associated with each synapse $\omega_{ts}^{(1)}$. Here, the subscript *t* indexes the hidden neuron, and $s \in \{0, 1, \dots, p\}$ is the index for the input variable. The weight associated with $s = 0$ is used for the bias. The superscript (1) indicates that the weights map the inputs from layer 1 (the input layer) to layer 2 (the hidden layer). Therefore, the inputs for the first hidden layer (the activations) are the input variables $x_1, \dots, x_p$.

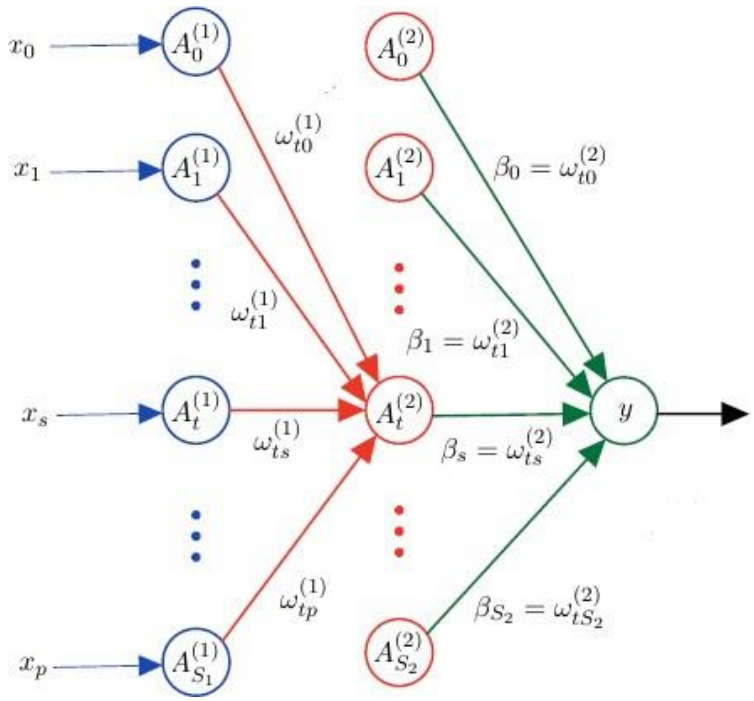


Figure 2. Single-layer neural network

Note: The blue nodes comprise the input layer, layer (1), which has one node for each input variable for a total of p variables. Each neuron, s, of this layer outputs the same input value, the activation $A_s^{(1)} = x_s$. *The red nodes comprise the hidden layer, layer (2), where each node t inputs the weighted average of the previous layer's activations. The activation s is multiplied by* $\omega_{ts}^{(1)}$, *and the results are summed. The average value,* $\bar{x}_t$, *is fed into a nonlinear activation function to produce the layer activations* $A_t^{(2)} = g(\bar{x}_t)$. *Notably,* $x_0 = 1$, *and the associated weight,* $\omega_{t0}^{(1)}$, *is the bias. The output layer activation function is linear. The output of the single neuron is the average of the hidden layer activations. The weight for activation* $A_t^{(2)}$ *is now* β_t.

The output of each neuron, *t*, of the hidden layer, the activation $A_t^{(2)}$, is calculated as follows:

$$A_t^{(2)} = g\left(\omega_{t0}^{(1)} + \sum_{s=1}^{p} \omega_{ts}^{(1)} x_s\right), t \in \{1, \dots, S_2\}$$

Here, S_2 is the number of neurons in the hidden layer (layer 2). The weighted average of the inputs, the total excitation $\bar{x} = \omega_{t0}^{(1)} + \sum_{s=1}^{p} \omega_{ts}^{(1)} x_s$, is compared with a threshold θ before the activation function is applied. If $\bar{x} \geq 0$, the unit fires. In this case, the function g is applied to $\bar{x}$. Otherwise, when $\bar{x} < 0$, the output is zero and the neuron is silent. The threshold θ can be seen as a new input to the neuron whose link has weight $-\theta$ and input $x_0 = 1$. It is known as bias and is generalized to a weight $\omega_{t0}^{(1)}$.

In a multilayer perceptron, as shown in Figure 3, the layer $l \in \{1, \dots, l_O\}$ has S_l nodes. The term l_O represents the total number of layers including the input layer. The activations, $t \in \{1, \dots, S_t\}$, are computed as $A_t^{(l)} = g\left(\omega_{t0}^{(l-1)} + \sum_{s=1}^{S_{l-1}} \omega_{ts}^{(l-1)} A_s^{(l-1)}\right)$.

The activation function in the output layer (l_O) neurons is the identity function. Therefore, the output for quantitative data is given as follows:

$$A_t^{(2)} = g\left(\omega_{t0}^{(1)} + \sum_{s=1}^{p} \omega_{ts}^{(1)} x_s\right), t \in \{1, \ldots, S_2\}$$

which is analogous to a linear regression if $\beta_s = \omega_{ts}^{(lo-1)}, \xi_s = A_s^{(lo-1)}$. Noticing that $y_t = A_t^{(lo)}$, then $y_t = \beta_0 + \sum_{s=1}^{S_{lo-1}} \beta_S \xi_S$.

For qualitative outputs, there are as many output neurons as levels, *m*. The activation function for the output is the softmax $f_t(x) = P(y = t|x) = \frac{e^{A_t(lo)}}{\sum_{t=0}^{m} e^{A_t(lo)}}$.

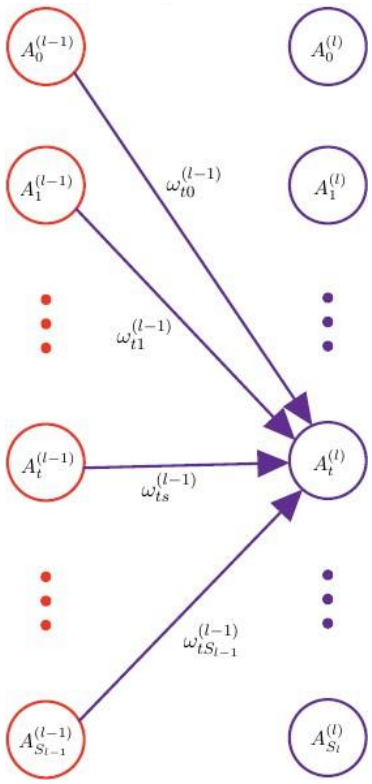


Figure 3. Multi-layer neural network

Note: Red nodes comprise the hidden layer $(l-1)$ *for a total of* S_{l-1} *neurons. Purple nodes comprise the hidden layer* (l) *for a total of* S_l *neurons. Neuron* t *in layer* l *receives the weighted average of the layer* $l-1$ *activations,* $\bar{x}_t$*, where* $\omega_{tS}^{(l-1)}$ *is the weight of the activation* $A_S^{(l-1)}$*. The average is fed into a nonlinear activation function to give the output of this neuron, which is the activation for layer* (l)*,* $A_t^{(l)} = g(\bar{x}_t)$ *. Note that* $A_0^{(l-1)} = 1$*, and* $\omega_{t0}^{(l-1)}$ *represents the bias.*

In summary, a neural network is a function f that takes p explanatory variables, denoted $x = (x_1, \ldots, x_p)$, as the input. The function then outputs a nonlinear function on x, $y = f(x|\Theta_{NN})$, which predicts the output y. The term Θ_{NN} is the complete set of weights in the neural network.

To estimate the hyperparameters Θ_{NN} of an ANN by learning them from the data, a performance measure is minimized. This measure can be the mean square error (MSE) for regression problems or the negative multinomial log-likelihood, cross-entropy, for classification problems [80]. The problem is not convex in parameters, meaning that it has local optima and is hard to solve due to the high number of parameters involved. Backpropagation is the standard procedure to estimate the parameters

of an ANN. In backpropagation, the weights are initialized randomly and updated until the objective function is minimized. This procedure involves calculating the gradient of the objective function with respect to the parameters. These parameters are then adjusted through gradient descent.

The most common strategy to resolve overfitting is tuning the architecture of the network. In the present study, it involves determining the hyperparameters such as the number of neurons in each hidden layer. This issue is typically addressed with cross-validation. Different network topologies are tested, and the one with the lowest error on the testing set is selected. Another strategy, weight decay, bounds the size of the link weights to favor generalization of the ANN. This is achieved by adding the additional following term to the objective function: $\lambda \sum_j \Theta_{NN,j}^2$ for a parameter λ that can be adjusted with cross-validation. In minimizing the previous expression, the weights are bounded to low values, similar to ridge regularization. This procedure controls overfitting.

Neural networks have been criticized for being black boxes. This criticism stems from the difficulty in interpreting the parameters and thus the impact of variables on the result. In a trained ANN, large weights are associated with high signal transfer, positive weights are excitatory, and negative weights are inhibitory [83]. This observation suggests that weights can be used to assess variable importance. Olden et al. [84] developed an algorithm to calculate variable importance. This algorithm provides insight into the role of the inputs in the output. The connection weight method sums the product of the raw input-hidden weights and the raw hidden-output weights across all hidden neurons. The resulting signed values account for the impact strength of input variables on the output.

Clustering

Once the most important variables had been selected, an unsupervised clustering technique was employed to group them. This methodology partitions observations into *K* different groups or clusters based on similarity in terms of some specific criterion such as the distance between them. Similar observations are placed in the same group, and dissimilar observations are placed in different groups. The main purpose of this methodology is to reveal the underlying associations among observations within each cluster in the form of shared characteristics given by the variables.

The data set is represented by a matrix of dimension $n \times d, x \in \mathrm{X}^{n \times d}$, where n is the number of observations and d is the number of variables. Each observation i, x_i, is a d-dimensional vector containing the realized values for the d variables $x_{ij} \in \mathrm{X}_j$ for $j = 1, \dots, d$. Here, X_j is the real space $\mathbb{R}$ for continuous data, and $\{1, \dots, m_j\}$ for categorical data, where m_j is the number of levels of the categorical variable j. The term d_{cont} denotes the dimension of the continuous variables set, and d_{cat} denotes the dimension of the categorical variables set. Applying clustering to the data set gives the membership label vector $z \in \{1, \dots, \mathrm{K}\}^n$, where z_i is the cluster to which the observation x_i belongs, that is, a new categorical feature. The observations are assumed to be realizations of a finite mixture probability density function:

$$f(x_i|\Theta_c) = \sum_{k=1}^{K} p_k \varphi(x_i|\theta_k)$$

Here, $0 < p_k < 1$ are the mixture weights, which sum to 1, $\varphi(\cdot\,|\theta_k)$ is a distribution parametrized by θ_k, and $\Theta_c = (p_1, \dots, p_K, \theta_1, \dots, \theta_K)$ is the complete set of parameters. For continuous variables, φ is a multivariate Gaussian with mean $\mu_k \in \mathbb{R}^{d_{cont}}$ and covariance matrix $\Sigma_k \in \mathbb{R}^{d_{cont} \times d_{cont}}$. In other words, $\theta_k = (\mu_k, \Sigma_k)$. In this case the distribution density $\varphi(\cdot\,|\theta_k)$ had the following form:

$$\varphi(x_i|\theta_k) = (2\pi)^{-d_{cont}/2} |\Sigma_k|^{-1/2} \exp\left\{-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k)\right\}$$

Here, ^{T} is the transpose operator and $|\cdot|$ is the determinant. For categorical variables, φ is a multivariate multinomial distribution with parameters $\alpha_k^j \in \mathbb{R}_{++}^{m_j}$ for $j \in \{1, \ldots, d_{cat}\}$. In other words, $\theta_k = (\alpha_k^j)_{j\in\{1,\ldots,d_{cat}\}}$. In this case the density distribution $\varphi(\cdot\,|\theta_k)$ had the following form:

$$(x_i|\theta_j) = \prod_{j=1}^{d_{cat}} \prod_{h=1}^{m_j} (\alpha_k^{jh}) x_s^{jh}$$

where superscript j runs over the d_{cat^t} variables, and superscript *h* runs over the m_j levels. Parameter estimation was conducted by maximizing the log-likelihood on Θ_c, using the expectation maximization (EM) algorithm.

Procedure

The variables in the data set were preprocessed by type. Numerical variables were standardized using the R function *scale*. Categorical variables were converted into factors. For the unstructured variable (project title), the latent topics were identified, and each observation (project title) was decomposed into these latent topics. The optimal number of topics was determined by applying the function *FindTopicsNumber* from the R library *ldatuning* with the Gibbs method, 100,000 iterations, and a burnin of 10,000 iterations. The number of topics was assessed with the metrics "CaoJuan2009" and "Deveaud2014". This analysis revealed that three topics were optimal. The project title variable was preprocessed using the LDA function of the R package *topicmodels*, with the Gibbs method and 100,000 iterations. Three numerical, 516-dimensional vectors $\theta = (\theta_1, \theta_2, \theta_3)$ were obtained, one for each topic. Those vectors were standardized with the R function *scale* and included in the data set. Finally, the original variable was removed from the data set.

The preprocessed data set was used to fit an ANN using the *mlpWeightDecayML* function. This function is a multi-layer perceptron with weight decay from the R package *caret*. It was run using the *train* function. The number of neurons was varied from 1 to 10 in each of the three hidden layers. The weight decay parameter was varied in the set $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. Cross-validation was performed with the *repeatedcv* method, setting the parameters number = 10 and repeats = 10. Variable importance was assessed with the function *olden* from the R package *NeuralNetTools*. The optimal ANN was found to be a three hidden layer with the number of neurons {12,1,1} and a weight decay of 0.4. The analysis was completed by performing clustering using the *mixmodCluster* function from the R package *Rmixmod* [85]. The NEC criterion was applied.

5 Results and discussion

Table 1 presents descriptive statistics for the research variables. The analysis indicates that 30.4% of applicants were approved for the subsidy. The service sector accounted for 82.4% of companies in the sample. Only 4.1% had international operations. The projects were proposed by companies from different regions: 27.9% from Madrid, 21.3% from Catalunya, 12.6% from Basque Country, and 8.9% from Valencia. The remaining projects (29.3%) were proposed by companies from other Spanish regions. In contrast, the regional distribution of start-ups in Spain is as follows: Madrid (25%), Catalonia (22%), Andalusia (16%), and Valencia (10%) [87]. Therefore, Andalusian companies are not proportionally represented in NEOTEC according to the proportion of start-ups.

Table 1. Mean and standard deviation (in parentheses) of the main variables

Brands	**Age***	**PPE****	**Totequ****	**Assets****	**AggVal****	**Profits****
0.8	487.9	133.4	75.5	256.4	33.8	-31.2
(1.2)	(239.4)	(317.2)	(207.2)	(503.9)	(141.2)	(156.4)

* In days
** In thousands of euros

The first step in the proposed method was to convert the unstructured data (the titles of the innovation projects) to structured data using the NLP technique of topic modeling. Following analysis of the data, the optimal number of topics was found to be three. Table 2 presents the most frequent keywords in each of the three topics. This information can help understand the title contents.

Table 2. Topic modeling of key words in project titles

	Main words
Topic 1	intelligence; base; artificial; application; process
Topic 2	development; new; system; plan; platform
Topic 3	tech; business; data; digital; solution

The results of the topic modeling in Table 2 suggest that these three topics are relevant for companies seeking to obtain R&D subsidies. Topic 1 seems to be relevant to innovation projects involving the implementation of artificial intelligence technologies, data-driven applications, and automated processes. This topic would cover projects in areas such as machine learning, NLP, robotics, and business process automation. Topic 2 seems to be related to projects to develop new or enhanced systems such as software development, R&D, and systems engineering projects. Finally, Topic 3 seems to be associated with business technology projects, data analytics, and digital solutions. This topic would cover projects such as enterprise software development, data analytics, and digital solutions development. Overall, these groupings of words provide insight into the types of innovation projects pursued by companies seeking R&D subsidies.

To establish the strength of association between each project title and these three topics, the words in each project title were grouped with each topic based on the probability of appearing in that topic. This analysis created three variables that indicate the degree of relevance of each topic to each project. These three variables were included in the data set with the variables discussed earlier.

ANNs are increasingly recognized as powerful tools for determining the importance of variables in classification problems. When predicting whether a company will receive an R&D subsidy, ANNs can help identify the most relevant variables that are strongly related to the outcome of interest. In this study, the outcome was approval or rejection of the subsidy application. Selecting the most important variables reduced the variance and improved the accuracy of the model. ANN estimation revealed that the variables TotEqu, PPE, and Profits were the most important factors in predicting the success or failure of a company's innovation subsidy application (as shown in Table 3). Although Activity, Assets, Age, Internationalization, and number of registered brands may also be important factors, the discussion focuses only on those with importance indicators greater than 1.

Table 3. Importance of variables

Importance	VARIABLES
-1.56	Totequ
-0.85	Activity
-0.51	Assets
-0.49	Age
-0.38	Topic 2
0.01	Topic 3
0.23	AggVal
0.26	Topic 1
0.42	Brands
0.76	Internat
1.01	Profits
2.15	PPE

To explore the data further, a finite mixture model was used to identify the cluster structure of the 516 companies based on these three most important variables. The remaining variables were not included in the clustering process. Analysis based on the integrated completed likelihood (ICL) criterion revealed that the optimal number of distinct groups of companies was four (i.e., K = 4). External validation was conducted to confirm the validity of the clustering results using company characteristics represented by the variables Assets, Age, Brands, AggVal, Activity, and Internat. External validation showed that the clustering results were valid and provided valuable insight into the characteristics of the four groups. This approach offered a robust and reliable method of understanding the relationships between variables and identifying the key drivers of a company's likelihood of receiving an innovation subsidy.

The clustering process was effective in identifying four distinct clusters of companies based on their characteristics (as shown in Table 4). The means and standard deviations were well separated for the four distinct company clusters. The three most important variables, TotEqu, PPE, and Profits, were used to cluster the sample based on significant differences between clusters. Cluster 1 comprised 200 companies with the lowest levels of two variables (PPE = 16.0; TotEqu = 19.9) but higher profits (0.8) than companies in the other clusters. Cluster 2 consisted of only 17 companies with high levels of PPE (1210.1) and TotEqu (599.1) but the lowest negative profits (-537.6). Cluster 3 comprised 119 companies with moderate levels of PPE (275.4) and TotEqu (145.3), as well as negative profits (-39.4). Finally, Cluster 4 consisted of 180 companies with low levels of PPE (68.1) and TotEqu (41.7), as well as negative profits (-13.4). The results are therefore consistent with those of Cantner and Kösters [55], who suggested that start-up capital is a key factor in accessing this type of subsidy.

The clustering structure was validated using four variables that had not been used in the process. This external validation method confirmed that the clusters were correctly identified and provided an insightful characterization of the clusters, revealing the unique features of each cluster. For example, companies in Cluster 2 were the oldest and had the most registered brands, as well as the greatest assets and lowest aggregate value. The study showed that the identified clusters differed in terms of companies' effectiveness in securing innovation grants. Specifically, 31% of companies in Cluster 1, 17% of companies in Cluster 2, 34% of companies in Cluster 3, and 28% of companies in Cluster 4 successfully applied for innovation grants. This finding offers insight into the relevance of the identified clusters for innovation subsidy applicants. The clustering process and subsequent analysis provide a comprehensive understanding of the characteristics of different clusters of innovation grant applicants and their effectiveness in receiving these grants.

Table 4. Mean values and standard deviation (in parentheses) for the four clusters of applicants

Variable	CL1 N = 200	CL2 N = 17	CL3 N = 199	CL4 N = 180	Total	Signific.
Tangible assets (PPE)	16.0 (14.1)	1,210.1 (1,114.9)	275.4 (216.9)	68.1 (62.2)	133.4 (317.2)	<.0001
Total equity	19.9 (11.2)	599.1 (862.5)	145.3 (180.3)	41.7 (40.1)	75.5 (207.2)	<.0001
Profits	0.8 (8.2)	-537.6 (621.8)	-39.4 (114.3)	-13.4 (35.0)	-31.2 (156.4)	<.0001
External validation						
Brands	0.51 (0.7)	1.82 (2.8)	0.97 (0.9)	0.94 (1.5)	0.81 (1.2)	<.0001
Assets	75.8 (184.0)	1,858.4 (1,457.4)	490.5 (491.8)	151.0 (125.9)	256.4 (503.9)	<.0001
Aggregate value (AggVal)	23.0 (46.8)	-177.1 (386.1)	75.6 (189.8)	38.1 (110.8)	33.8 (141.1)	<.0001
Age (in days)	433.9 (241.5)	665.8 (198.3)	542.6 (210.8)	494.9 (242.8)	487.8 (239.4)	<.0001

Overall, the use of ANN and finite mixture models proved to be a valuable approach for identifying the most important variables in predicting successful innovation subsidy applications and understanding the cluster structure of the data. These findings have implications for policymakers and companies seeking to increase their chances of receiving innovation subsidies.

6 Conclusions and implications

This study aimed to provide a better understanding of the profile of subsidized companies in Spain. It was hoped that it would contribute to knowledge in the field of subsidies for technology-based companies. The study used advanced estimation methodologies, namely NLP, ANN techniques, and clustering. The results reveal a disparity between regions with the most start-ups and those with the most companies accessing NEOTEC funds. Further research is necessary to determine whether this disparity occurs because regional support may be more attractive to these start-ups or because certain regions are underrepresented. This finding is important because it may have political implications. It may suggest a need to allocate a greater budget to regions that are not proportionally represented.

This study used NLP techniques to examine the titles of grant applications. Three optimal topics were identified. Classification of the project title keywords into these topics enabled characterization of the core focus of each project. Topic 1 refers to projects centered on the implementation of artificial intelligence technologies, data-driven applications, and automated processes. Topic 2 focuses on the development of new or enhanced systems, such as software development, R&D, and systems engineering projects. Finally, Topic 3 refers to business technology projects, data analytics, and digital solutions. These findings are important because they offer valuable insight into the areas of interest for innovation projects. They can assist policymakers and researchers in identifying trends in the types of projects that receive grants. One potential policy implication would be to define strategic thematic areas and allocate the budget accordingly. Such a system could prevent projects focused on artificial intelligence or software development from cannibalizing one another. There are already other grants in Spain that support R&D projects in artificial intelligence, other digital technologies, and their integration into the value chain [88].

In the next phase of analysis, ANNs were used to identify the most relevant variables in explaining the approval or rejection of grant applications. Total equity (*TotEqu*), long-term tangible assets (*PPE*),

and profits (*Profits*) were the most important factors in determining the probability of success or failure in applying for a subsidy.

A finite mixed model was employed to identify the cluster structure of the sample observations based on these variables. Clustering identified four clusters of companies based on their characteristics. The three most important variables, total equity (*TotEqu*), long-term tangible assets (*PPE*), and profits (*Profits*), were used to cluster the sample based on significant differences between clusters. Cluster 1 comprised 200 companies with the lowest values for two variables (*PPE* and *TotEqu*) but higher *profits* than companies in the other clusters. Cluster 2 consisted of only 17 companies with high levels of *PPE* and *TotEqu* but the lowest negative *profits*. Cluster 3 comprised 119 companies with moderate levels of *PPE* and *TotEqu* but negative *profits*. Finally, Cluster 4 consisted of 180 companies with low levels of *PPE* and *TotEqu* and negative *profits*. In short, the model suggests that approved projects require low investment in tangible assets and are likely to generate short-term returns. However, it is unclear whether this trend reflects a key factor in project selection or whether it is driven by the nature of the approved projects, which are inherently in need of low initial investment in assets. Further investigation is required to determine the underlying reasons.

To validate the clustering process and characterize the resulting clusters, four additional variables were selected. This external validation method confirmed that the clusters were correctly identified. It also revealed the unique characteristics of each cluster. The identified clusters differed in their effectiveness in securing innovation grants. This finding provides valuable insight into the relevance of the identified clusters for applicants of innovation grants. The clustering process and subsequent analysis provide a comprehensive understanding of the characteristics of different clusters of innovation grant applicants and their effectiveness in securing these grants.

In summary, the study highlights the effectiveness of using ANNs in combination with finite mixture models to identify the most important variables for securing innovation grants and to gain an understanding of the cluster structure of the data. This insight is crucial for grant applicants and policymakers seeking to allocate grants effectively. Nonetheless, the study is subject to some limitations. For instance, project classification was based solely on the project title, which may not always accurately reflect the nature of the project.

References

[1] Acharya, V., & Xu, Z. (2017). Financial dependence and innovation: the case of public versus private firms. *Journal of Financial Economics*, *124*(2), 223-243. https://doi.org/10.1016/j.jfineco.2016.02.010

[2] Zhang, J., & Guan, J. (2018). The time-varying impacts of government incentives on innovation. *Technological Forecasting Social Change*, 135, 132-144. https://doi.org/10.1016/j.techfore.2018.04.012

[3] Shinkle, G. A., & Suchard, J. A. (2019). Innovation in newly public firms: the influence of government grants, venture capital and private equity. *Australian Journal of Management*, *44*(2), 248-281. https://doi.org/10.1177/0312896218802611

[4] Eurostat (2022). *Gross Domestic Expenditure on R&D*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=R%26D_expenditure&oldid=551418#Gross_domestic_expenditure_on_R.26D

[5] Vilaplana Aparicio, M. J., & Martín Llaguno, M. (2022). Systematic review on R&D&I aid in Spain. *Intangible Capital*, *18*(3), 331-349. https://doi.org/10.3926/ic.1339

[6] Gispert, B. (24 November 2022). La inversión española en I+D crece hasta el 1,43% del PIB pero sigue lejos de la media de la UE. *La Vanguardia*. https://www.lavanguardia.com/economia/20221124/8620424/inversion-espanola-i-d-crece-1-43-pib-sigue-lejos-media-ue.html

[7] Busom, I. (1999). An empirical evaluation of the effects of R&D subsidies. *Economics of innovation and new technology*, *9*(2), 111-148. https://doi.org/10.1080/10438590000000006

[8] Callejón, M., & García-Quevedo, J. (2005). Public subsidies to business R&D: do they stimulate private expenditures?. *Environment and Planning C: Government and Policy*, *23*(2), 279-293.

[9] González, X., Jaumandreu, J., & Pazó, C. (2005). Barriers to innovation and subsidy effectiveness. *RAND Journal of Economics*, 930-950.

[10] Marra Domínguez, M.A. (2006). Efectos de las subvenciones públicas sobre la inversión en I+ D de las empresas manufactureras españolas. *Revista Galega de Economía*, *15*(2).

[11] González, X., & Pazó, C. (2008). Do public subsidies stimulate private R&D spending?. *Research Policy*, *37*(3), 371-389. https://doi.org/10.1016/j.respol.2007.10.009

[12] Huergo, E., & Moreno, L. (2017). Subsidies or loans? Evaluating the impact of R&D support programmes. *Research Policy*, *46*(7), 1198-1214. https://doi.org/10.1016/j.respol.2017.05.006

[13] Segarra Blasco, A. (2018). Subsidies, Loans and Tax Incentives for Business R&D in Catalonia. *Investigaciones Regionales*, (40), 109-140.

[14] CDTI (2022). Plan Operativo Anual del CDTI 2022. https://www.cdti.es/index.asp?MP=99&MS=929&MN=1&TR=A&IDR=1&iddocumento=6759

[15] Mei, L., Rodríguez, H., & Chen, J. (2020). Responsible Innovation in the contexts of the European Union and China: Differences, challenges and opportunities. *Global Transitions*, 2, 1-3. https://doi.org/10.1016/j.glt.2019.11.004

[16] Wanzenböck, I., & Frenken, K. (2020). The subsidiarity principle in innovation policy for societal challenges. *Global Transitions*, 2, 51-59. https://doi.org/10.1016/j.glt.2020.02.002

[17] United Nations (27 November 2022). Sustainable Development Goals. https://sdgs.un.org

[18] Becker, B. (2015). Public R&D policies and private R&D investment: A survey of the mpirical evidence. *Journal of Economic Surveys*, *29*(5), 917-942. https://doi.org/10.1111/joes.12074

[19] Zúñiga-Vicente, J. A., Alonso-Borrego, C., Forcadell, F. J., & Galán, J. I. (2014). Assessing the effect of public subsidies on firm R&D investment: A survey. *Journal of Economic Surveys*, *28*(1), 36-67. https://doi.org/10.1111/j.1467-6419.2012.00738.x

[20] Bravo-Ortega, C., Benavente, J. M., & González, A. (2014). Innovation, exports and productivity: learning and self-selection in Chile. *Emerging Markets Finance and Trade*, *50*(1), 68-95. https://doi.org/10.2753/REE1540-496X5001S105

[21] Arrow, K. J. (1962). Economic welfare and the allocation of resources for invention. In R. Nelson (Ed), *The Rate and Direction of Inventive Activity* (pp. 609-626). Princeton University Press.

[22] Leiponen, A., & Byma, J. (2009). If you cannot block, you better run: Small firms, cooperative innovation, and appropriation strategies. *Research Policy*, *38*(9), 1478-1488. https://doi.org/10.1016/j.respol.2009.06.003

[23] Hall, B. H., & Lerner, J. (2010). The financing of R&D and innovation. In B. H. Hall & N. Rosenberg (Eds.), *Handbook of the Economics of Innovation* (pp. 609-639). Elsevier. https://doi.org/10.1016/S0169-7218(10)01014-2

[24] Gök, A., & Edler, J. (2012). The use of behavioural additionality evaluation in innovation policy making. *Research Evaluation*, *21*(4), 306-318. https://doi.org/10.1093/reseval/rvs015

[25] Kiman, K., & Jongmin, Y. (2022). Linear or nonlinear? Investigation an affect of public subsidies on SMEs R&D investment. *Journal of the Knowledge Economy*, *13*(3), 2519-2546. https://doi.org/10.1007/s13132-021-00823-9

[26] Clarysse, B., Wright, M., & Mustar, P. (2009). Behavioural additionality of R&D subsidies: a learning perspective. *Research Policy*, *38*(10), 1517-1533. https://doi.org/10.1016/j.respol.2009.09.003

[27] Blanes, J. V., & Busom, I. (2004). Who participates in R&D subsidy programs? The case of Spanish manufacturing firms. *Research Policy*, *33*(10), 1459-1476. https://doi.org/10.1016/j.respol.2004.07.006

[28] Pisár, P., Ďurčeková, I., & Křápek, M. (2021). Effectiveness of Public Support for Business Innovation from EU Funds: Case Study in Slovakia. *NISPAcee Journal of Public Administration and Policy*, *14*(1), 261-283. https://doi.org/10.1007/s11356-021-14555-5

[29] Meyer-Krahmer, F., & Montigny, P. (1989). Evaluations of innovation programmes in selected European countries. *Research Policy, 18*(6), 313-332. https://doi.org/10.1016/0048-7333(89)90020-6

[30] Ormala, E. (1989). Nordic experiences of the evaluation of technical research and development. *Research Policy, 18*(6), 333-342. https://doi.org/10.1016/0048-7333(89)90021-8

[31] Roessner, D. (1989). Evaluating government innovation programmes: lessons from the U.S. experience. *Research Policy, 18*(6), 343-359. https://doi.org/10.1016/0048-7333(89)90022-X

[32] Dimos, C., & Pugh, G. (2016). The effectiveness of R&D subsidies: a meta-regression analysis of the evaluation literature. *Research Policy*, *45*(4), 797-815. https://doi.org/10.1016/j.respol.2016.01.002

[33] Xiang, D., Zhao, T., & Zhang, N. (2021). Does public subsidy promote sustainable innovation? The case of Chinese high-tech SMEs. *Environmental Science and Pollution Research*, *28*(38), 53493-53506. https://doi.org/10.1007/s11356-021-14555-5

[34] Marino, M., Lhuillery, S., Parrotta, P., & Sala, D. (2016). Additionality or crowding-out? An overall evaluation of public R&D subsidy on private R&D expenditure. *Research Policy*, *45*(9), 1715-1730. https://doi.org/10.1016/j.respol.2016.04.009

[35] Crespi, G., Giuliodori, D., Giuliodori, R., & Rodríguez, A. (2016). The effectiveness of tax incentives R&D+i in developing countries: the case of Argentina. *Research Policy, 45*(10), 2023-2035. https://doi.org/10.1016/j.respol.2016.07.006

[36] Berrutti, F., & Bianchi, C. (2019). Effects of public funding on firm innovation: transforming or reinforcing a weak innovation pattern?. *Economics of Innovation and New Technology, 29*(5), 522-539. https://doi.org/10.1080/10438599.2019.1636452

[37] Laplane, A. (2021). Market co-creating and shaping through investments in innovation: a comparative analysis of two public funding programmes in Brazil. *Innovation and Development.* https://doi.org/10.1080/2157930X.2021.1989646

[38] Montmartin, B., Herrera, M., & Massard, N. (2018). The impact of the French policy mix on business R&D: how geography matters. *Research Policy, 47*(10), 2010-2027. https://doi.org/10.1016/j.respol.2018.07.009

[39] Pavitt, K. (1984). Sectoral patterns of technical change: towards a taxonomy and a theory. *Research Policy*, *13*(6), 343-373. https://doi.org/10.1016/0048-7333(84)90018-0

[40] Malerba, F. (2002). Sectoral systems of innovation and production. *Research Policy, 31*(2), 247-264. https://doi.org/10.1016/S0048-7333(01)00139-1

[41] Malerba, F. (2005a). Sectoral systems: how and why innovation differs across sectors. In J. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds), *The Oxford Handbook of Innovation* (pp. 380-406). Oxford University Press.

[42] Malerba, F. (2005b). Sectoral systems of innovation: a framework for linking innovation to the knowledge base, structure and dynamics of sectors. *Economics of Innovation and New Technology, 14*(1-2), 63-82. https://doi.org/10.1080/1043859042000228688

[43] Leiponen, A., & Drejer, I. (2007). What exactly are technological regimes? Intra-industry heterogeneity in the organization of innovation activities. *Research Policy*, *36*(8), 1221-1238. https://doi.org/10.1016/j.respol.2007.04.008

[44] Srholec, M., & Verspagen, B. (2012). The voyage of the beagle into innovation: explorations on heterogeneity, selection and sectors. *Industrial and Corporate Change*, *21*(5), 1221-1253. https://doi.org/10.1093/icc/dts026

[45] Coad, A. (2009). *The growth of firms: a survey of theories and empirical evidence*. Edward Elgar Publising Limited.

[46] Aiello, F., Albanese, G., & Piselli, P. (2019). Good value for public money? The case of R&D policy. *Journal of Policy Modeling*, *41*(6), 1057-1076. https://doi.org/10.1016/j.jpolmod.2019.02.006

[47] Sinadinos, C. (2022). Cyclical IPR-public Grant Engine Driving R&D Innovation in Small Research-intensive Private Enterprises. *Journal of Innovation Management*, *10*(1), I-X. https://doi.org/10.24840/2183-0606_010.001_L001

[48] Vanino, E., Roper, S., & Becker, B. (2019). Knowledge to money: Assessing the business performance effects of publicly-funded R&D grants. *Research Policy*, *48*(7), 1714-1737. https://doi.org/10.1016/j.respol.2019.04.001

[49] Hsu, Y. G., Tzeng, G. H., & Shyu, J. S. (2003). Fuzzy multiple criteria selection of government-sponsored frontier technology R&D projects. *R&D Management*, *33*(5), 539-551. https://doi.org/10.1111/1467-9310.00315

[50] Lee, M., & Om, K. (1996). Different factor considered in project selection at public and private R&D institutes. *Technovation*, *16*(6), 271-275. https://doi.org/10.1016/0166-4972(96)00006-5

[51] Lee, M., & Om, K. (1997). The concept of effectiveness in R&D project selection. *International Technology Management*, *13*(5-6), 511-524. https://doi.org/10.1504/IJTM.1997.001684

[52] Santamaría, L., Barge-Gil, A., & Modrego, A. (2010). Public selection and financing of R&D cooperative projects: Credit versus subsidy funding. *Research Policy*, *39*(4), 549-563. https://doi.org/10.1016/j.respol.2010.01.011

[53] Acosta Ballesteros, J., & Modrego Rico, A. (2001). Public financing of cooperative R&D projects in Spain: The Concerted Projects under the National R&D Plan. *Research Policy*, *30*(4), 625-641. https://doi.org/10.1016/S0048-7333(00)00096-2

[54] Afcha, S. (2012). Analyzing the interaction between R&D subsidies and firm's innovation strategy. *Journal of Technology Management & Innovation*, *7*(3), 57-70. http://dx.doi.org/10.4067/S0718-27242012000300006

[55] Cantner, U., & Kösters, S. (2012). Picking the winner? Empirical evidence on the targeting of R&D subsidies to start- ups. *Small Business Economics*, 39, 921-936. https://doi.org/10.1007/s11187-011-9340-9

[56] Segarra-Blasco, A., & Teruel, M. (2016). Application and success of R&D subsidies: what is the role of firm age?. *Industry and Innovation*, *23*(8), 713-733. https://doi.org/10.1080/13662716.2016.1201649

[57] Alarcón, S., & Arias, P. (2018). The public funding of innovation in agri-food businesses. *Spanish Journal of Agricultural Research*, *16*(4). https://doi.org/10.5424/sjar/2018164-12657

[58] García Quevedo, J., & Afcha Chávez, S. (2009). Assessing the impact of public funds on private R&D: A comparative analysis between state and regional subsidies. *Investigaciones Regionales-Journal of Regional Research*, (15), 277-294.

[59] Vila Alonso, M., Ferro Soto, C., & Guisado González, M. (2010). Innovación, financiación pública y tamaño empresarial. *Cuadernos de gestión*, *10*(1), 75-87.

[60] Labeaga Azcona, J.M., & Martínez-Ros, E. (2012). Evaluation of the Impact of R&D Tax Incentives on the Propensity to Innovate. *Proceedings of the 9th International Conference on Innovation & Managem*ent, 966-974.

[61] Cuenca, L., & Boza, A. (2015). Public funding in R&D projects: Opportunities for companies. *INTED2015 Proceedings*, 2194-2199.

[62] Almus, M., & Czarnitzki, D. (2003). The effects of public R&D subsidies on firms´ innovation activities. *Journal of Business & Economic Statistics*, *21*(2), 226-236. https://doi.org/10.1198/073500103288618918

[63] Mardones, C., & Zapata, A. (2019). Determinants of public funding for innovation in Chilean firms. *Contaduría y Administración, 64*(1), 1-16. http://dx.doi.org/10.22201/fca.24488410e.2018.1602

[64] Gao, Y., Hu, Y., Liu, X., & Zhang, H. (2021). Can public R&D subsidy facilitate firms' exploratory innovation? The heterogeneous effects between central and local subsidy programs. *Research Policy*, *50*(4), 104221. https://doi.org/10.1016/j.respol.2021.104221

[65] Duguet, E. (2003). *Are R&D subsidies s substitute or a complement to privately funded R&D? Evidence from France using Propensity Score Methods for non-experimental data* (I Cahier de la MSE EUREQua Working Paper No. 2003.75). https://dx.doi.org/10.2139/ssrn.421920

[66] Antonelli, C., & Crespi, F. (2013). The "Mathew effect" in R&D public subsidies: The Italian evidence. *Technological Forecasting and Social Change*, *80*(8), 1523-1534. https://doi.org/10.1016/j.techfore.2013.03.008

[67] Duch-Brown, N., García-Quevedo, J., & Montolio, D. (2011). The link between public support and private R&D effort: what is the optimal subsidy? *SSRN Electronic Journal*. https://dx.doi.org/10.2139/ssrn.1864192

[68] Hussinger, K. (2008). R&D and subsidies at the firm level: an application of parametric and semiparametric two-step selection models. *Journal of Applied Econometrics*, *23*(6), 729-747. https://doi.org/10.1002/jae.1016

[69] CDTI (10 September 2022). Proyectos Neotec. https://www.cdti.es/index.asp?MP=100&MS=818&MN=2

[70] Srivastava, A. N., & Sahami, M. (Eds.). (2009). *Text mining: Classification, clustering, and applications*. CRC press. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Taylor & Francis

[71] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

[72] Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 50–57, New York, NY, USA.

[73] Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2013). *Handbook of latent semantic analysis*. Psychology Press.

[74] Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of statistical software*, *40*, 1-30. https://doi.org/10.18637/jss.v040.i13

[75] Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, *17*(1), 61-84. https://doi.org/10.3166/DN.17.1.61-84

[76] Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7-9), 1775-1781. https://doi.org/10.1016/j.neucom.2008.06.011

[77] Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman & Hall/CRC Data Science Series. CRC Press.

[78] Kuhn, M., & Johnson, K. (2019). Feature engineering and selection: A practical approach for predictive models. n Taylor & Francis Group.

[79] Woodruff, A. (20 October 2022). What is a neuron? https://qbi.uq.edu.au/brain/brain-anatomy/what-neuron

[80] Zell, A., Mamier, G., Vogt, M., Mache, N., Hübner, R., Döring, S., ... & Wieland, J. (1995). *SNNS: Stuttgart Neural Network Simulator. User Manual, Version 4.2*. Institute for Parallel and Distributed High Performance Systems, Technical Report, (6/95).

[81] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer. https://doi.org/10.1007/978-0-387-21606-5

[82] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York.

[83] Rojas, R. (1996). Modular Neural Networks. *Neural Networks: A Systematic Introduction*, 411-425. https://doi.org/10.1007/978-3-642-61068-4_16

[84] Olden, J. D., & Jackson, D. A. (2002). Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling*, *154*(1-2), 135-150. https://doi.org/10.1016/S0304-3800(02)00064-9

[85] Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological modelling*, *178*(3-4), 389-397. https://doi.org/10.1016/j.ecolmodel.2004.03.013

[86] Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G., & Govaert, G. (2015). Rmixmod: The R package of the model-based unsupervised, supervised, and semi-supervised classification Mixmod library. *Journal of Statistical Software*, *67*, 1-29. https://doi.org/10.18637/jss.v067.i06

[87] Vazquez, D. (2022). El mapa de las startups en España: en qué comunidades autónomas hay más startups activas. https://www.businessinsider.es/madrid-cataluna-quedan-casi-mitad-startups-1134409

[88] Red.es (December 2022). Convocatoria de ayudas 2021 destinadas a proyectos de investigación y desarrollo en inteligencia artificial y otras tecnologías digitales y su integración en las cadenas de valor. https://sede.red.gob.es/es/procedimientos/convocatoria-de-ayudas-2021-destinadas-proyectos-de-investigacion-y-desarrollo-en

About the Authors

Mónica Espinosa-Blasco holds a PhD in Economics and Business Administration. She is currently an Associate Professor in the Department of Financial Economics and Accounting at the University of Alicante and Academic Coordinator of External Internships in the Faculty of Economics at the University of Alicante. She has participated in numerous research projects subsidized by public agencies. She has also partaken in major R&D contracts with companies and public administrations. Her research focuses on the relevance of the disclosure of accounting information for valuation. She has presented papers at numerous national and international conferences (European Accounting Association, European Financial Management Association, Accounting Research Workshop or Workshop on Empirical Research in Financial Accounting). She has published articles in national and international journals such as European Accounting Review, Journal of Business Finance and Accounting, Revista Española de Financiación y Contabilidad, and Investigaciones Económicas.

Gabriel Ignacio Penagos-Londoño is an Assistant Professor of Finance and Mathematical Economics at Pontificia Universidad Javeriana (Colombia). He has been Visiting Scientific Researcher at the Fields Institute for the Thematic Program on Variational Problems in Physics, Economics and Geometry. After his master's degree, he took a position as a portfolio manager and trader at a broker. He then switched to a risk management role in market risk at various financial institutions, including brokers, mutual funds, and banks. His role entailed developing mathematical models to assess risk exposure. His current research interests are machine learning, stochastic calculus, and Bayesian inference for marketing and economic modeling. He has published academic articles in journals such as Journal of Economic Interaction and Coordination, Journal of Modelling in Management and Journal of Mathematical Finance.

Felipe Ruiz Moreno is an Associate Professor of Marketing at the University of Alicante. Prior to his current position, he conducted research as a visiting scholar at several institutions in the United States and Poland. His research focuses on various areas of marketing, including competitive strategy, banking marketing, international marketing, sustainability, and segmentation. He has published numerous articles in high-profile academic journals, including Strategic Management Journal, Journal of Business Research, and Journal of Destination Marketing & Management.

María J. Vilaplana-Aparicio holds a PhD from the University of Alicante (with a special mention), a degree in Advertising and Public Relations from the University of Alicante, and a master's degree in Business Management. She currently lectures in the Department of Communication and Social Psychology at the University of Alicante. For nearly 15 years, she has worked in a consultancy firm specializing in the management of subsidies and tax incentives, mainly in the area of research and development and innovation. She has delivered more than 50 lectures on public aid in different Spanish regions and at the national level.

This page is intentionally left blank