Selecting Features that Drive Internationalization of Spanish Firms

Alfredo Jiménez[1] and Álvaro Herrero[2]

[1] Department of Management, KEDGE Business School, Bordeaux, France.

alfredo.jimenez@kedgebs.com

[2] Grupo de Inteligencia Computacional Aplicada (GICAP), Departamento de Ingeniería Civil, Escuela Politécnica Superior, Universidad de Burgos, Av. Cantabria s/n, 09006, Burgos, Spain.

ahcosio@ubu.es

Abstract. Internationalization is nowadays common for many large enterprises and as a result Foreign Direct Investment (FDI) flows are increasing. Bilateral psychic distance stimuli between home and host countries and vicarious experience are analyzed, together with other firm- and country-level FDI determinants, by means of a feature selection model to gain deep knowledge about the internationalization strategy of Spanish large firms. Wrapper feature selection based on a genetic algorithm is applied to identify the most important features leading to internationalization decision. Additionally, obtained results are compared to those obtained by logistic regressions (Standard, Rare Event, and Conditional).

Keywords: Wrapper Feature Selection, Genetic Algorithms, Support Vector Machine, Random Forest, Internationalization.

1 Introduction

Managing international operations is a critical component of many firms' strategy nowadays. Among the many decisions involved in internationalization, one of the first ones to take is the choice of location. Literature in the International Business domain has long study the determinants of Foreign Direct Investment (FDI), showing the relevance of both firm-level (size, sector, performance, etc.) and country-level factors (GDP growth, population, unemployment, etc.) which can act as incentives or barriers to international ventures. Among those determinants, two of them are attracting lately a higher amount of scholarly interest: distance and experience. In particular, researchers have warned against the typical oversimplification in the conceptualization of these two determinants, and instead offer a finer-grained analysis in recent theoretical and empirical papers where their multi-dimensional nature is highlighted [1].

Recent advances in the study of the determinants of FDI has pointed out the relevance of the distance between the country of origin of the investor and the destination country. These differences have been operationalized though a wide variety of measures, highlighting not only the relevance, but also the multi-dimensional nature of distance. In fact, managing effectively the distance of the firm´s operations abroad has become the essence of the international management field [2]. Traditionally, most studies have focused on the cultural aspect of distance. Arguably, the most influential work was the one from Hofstede (1980), who originally proposed four dimensions of cultural distance, although in subsequent work with colleagues [3] extended the framework to include two additional dimensions. However, this framework has been subject to criticisms for multiple reasons [4]. Among those criticisms, scholars have underlined that the real determinant of investment choices is usually not cultural distance, but psychic distance [5].

Psychic distance refers to *"the sum of factors preventing the flow of information from and to the market. Examples are differences in language, education, business practices, culture, and industrial development"* [6]. While the exact psychic distance should be measured at the individual level and right at the time of making the investment decision, it has been shown that psychic distance perceptions are influenced by macro-level factors, which shapes the individual values of psychic distance [7] [8]. Previous research has found many examples of the significant effect of psychic distance stimuli on a wide range of managerial decisions, including also FDI decisions [9]. To measure these psychic distance stimuli, one of the most accepted and validated measures are those of proposed by Dow and Karunaratna (2006), who propose six different stimuli of national differences in education, industrial development, language, democracy, social system, and religion.

Furthermore, recent studies in international business have focused on another classic determinant of FDI, in this case experience, drawing the attention over the fact that, contrary to common practice, this is also a multi-faceted construct [1, 10]. Thus, drawing on Organizational Learning Theory [11, 12], firms might derive useful knowledge not only from their own experience, but also from the experience of other firms (i.e. vicarious experience). Firms sharing a feature, for example the nationality, the sector, or the type of technology used, represent a valuable source of information about the host country [13]. The information derived from vicarious experience allow the firm to identify in advance potential challenges in the host country, replicating good practices and avoiding mistakes [14]. Moreover, vicarious experience facilitates the identification of opportunities and complementary partners abroad [15] and also increases the legitimacy of the decisions to overcome the resistance from possible stakeholders who oppose the strategy [16].

In present paper, a soft-computing tool is proposed to support enterprise managers by identifying those pieces of data that are key for taking internationalization decisions. When a manager has to take decisions that concern internationalization issues, it would be very helpful to develop a notion of those data features that deserve deeper attention. Only a few of the whole set of features about the enterprise itself and the destination countries can be analyzed in a timely manner. The proposed solution obtains those features, by applying Wrapper Feature Subset Selection, as described in Section 2.

The rest of this paper is organized as follows: the applied methods are described in section 2, the setup of experiments and the dataset under analysis are described in section 3, together with the results obtained and the conclusions of present study that are stated in section 4.

2 Genetic Feature Selection

The main target of present proposal is to obtain the most relevant features of the original data, which will provide enterprise managers with the information to take decisions on internationalization. To do so, Feature Selection (FS) is applied to identify those features driving to positive or negative internationalization decisions.

As defined by [17], FS involves a learning algorithm that approaches the problem of selecting a subset of features upon which it can focus its attention. The remaining features from the original dataset are not considered important and consequently ignored. There is also an induction algorithm that, as in general terms for supervised learning, tries to minimize classification error, being trained on different subsets of features taken from the original data. According to [18], two different levels of relevance (weak and strong) can be defined regarding each one of the features in the original dataset. One of such features is considered as strongly relevant if its removal causes a deterioration in the performance of the induction algorithm. In present paper, this is the underlying idea to identify key features that drive internationalization (whether positive or negative) decisions.

There are mainly three different types of FS methods, namely: Wrapper, Filter and Embedded. In the former, as opposed to the other two, it is the learning algorithm that "wraps" the induction algorithm and may be equated with a "black box" and is run on different subsets of features taken from the original data. Wrapper FS [18] has been selected in present study.

As the induction algorithm under the Wrapper FS perspective, two classifiers are considered and described below: Support Vector Machines and Random Forest. To generate the different subsets of features that are provided to these two classifiers, standard Genetic Algorithms (GAs) [19] have been applied as preliminary results had suggested that they are a powerful mean of reducing the time for finding near-optimal subsets of features from large datasets [20]. These are computational methods based on natural selection and natural genetics, used for searching solutions to a given problem. More precisely, a GA can be seen as an heuristic-based method for global optimization.

In order to optimize solutions to a given problem, these are codified as binary strings. In present case, for feature selection, a bit is assigned to each one of the features in the original dataset; 1 means that the features is included in the given subset and 0 means that it is not included. As a result, vectors of length *n* (being *n* the number of features in the original dataset) are constructed as solutions to the feature-subset selection problem.

In GA, there is a fitness function that measures the "quality" of the generated. Its design is part of the modelling process of the whole optimization approach [21]. In present paper, as per feature selection, the fitness function for was defined as the highest negative error rate (lowest classification error) obtained by applying the above-mentioned classifiers to the generated subset of features, when trying to classify the testing data to forecast the internationalization decision. As usual, selection, mutation and crossover operators are applied, according to certain parameters (different values have been tested as described in section 3.2). All in all, the applied GA is defined as follows:

Table 1. Pseudocode of the FS GA applied in present paper.

```
input: a feature selection problem
set the generation counter $g = 0$
for $i := 1$ to population size do
    create a random combination of feature subset (solution)
end for
while the number of generations is not reached do
    generate child solutions by applying the crossover operator (with a certain probability)
    generate child solutions by applying the selection and crossover operator (with a certain probability)
    compute fitness values by training and testing SVM on each individual (subset of features)
    apply the mutation operator
    selection of best child solutions (tournament)
    replace the worst member of the population by the child solutions
    $g = g + 1$
end while
output: the best subset of features for the given problem
```

2.1 Support Vector Machines

Support Vector Machines (SVMs) [22, 23], based on Statistical Learning Theory, face classification problems from the Structural Risk Minimization perspective as opposed to many other models that are based on the Empirical Risk Minimization paradigm. As a result, they show good generalization performance so they have been applied to wide range of real-life problems [24].

For this problem of finding an hyperplane to separate two classes (one-class classification), SVMs tries to find the optimal hyperplane that not only separates the classes with no error but also maximizes the distance to closest point (for either class).

SVMs can be seen as classifiers where the loss function is the Hinge function, defined as:

$$L[y, f(x)] = max[0{,}1 - yf(x)] \qquad (1)$$

Being x an observation from input features, y the class x belongs to, and $f(x)$ the output of the classifier.

Once trained, that is the support vectors are identified and the margin is maximized, a SVM can be seen as:

$$f(x) = \sum_{i \epsilon S} \alpha . y_i . \langle x_i, x \rangle + \beta_0 \qquad (2)$$

Being S the set of support vectors, α the classifier coefficients, and β the predictor coefficients.

SVMs have been proven as top classifiers and have been also applied to multi-class classification although they were initially designed for single-class classification. In present work, class information is the internationalization decision that was taken in each case: whether investing on a foreign country or not. Hence, SVMs face a one-class classification in present study.

To estimate the accuracy of the SVM for each individual (subset of features), N-fold cross-validation, taking value of 10, was applied.

2.2 Random Forest

Classification trees [25] are well-known and inductive learning methods. Within they inner (tree) structure there are two types of nodes; leaf nodes that are those for taking the final decision (prediction) and internal nodes that are those associated to differentiate responses (branches) for a given question regarding the values of a feature from the original training dataset. All internal nodes have at least two child nodes. Labels are assigned to both archs connecting a node to one of its child nodes (their content is related to the responses to the node question) and leaf nodes (their content is one of the classes in the training dataset). They have proved to be valuable tools for many interesting and challenging tasks such as description, classification and generalization of data [26].

In present research, Random Forest (RF) algorithm [27] has been selected to calculate the fitness function of the GA, from the variety of classification trees. RF can be seen as an aggregation of a number of classification trees such that each one of them depends on the values of a random vector. This vector is sampled independently and with the same distribution for all trees in the forest. One of the main advantages, when compared to a single classification tree schema, is the reduction of variance. In the case of RF, the prediction is obtained for a new data by aggregating (through majority voting) the predictions made by all the single trees. That is, the new data is assigned to the class that was most often predicted by the individual trees.

In order to select an appropriate number of trees (that is one of the requirements of Random Forests), conclusions from recent work [28] have been taken into account. According to it, for the binary classification problem that is addressed in present paper, no tuning has been carried out and the number of trees has been set to 100. To estimate the accuracy of the RF for each individual (subset of features), the Out-of-Bag (OOB) error rate is measured and analysed. It is calculated by predicting the class for each training sample by using only the trees for which this observation was not included in the bootstrap sample. That is, training data are not used to calculate the OOB error rate of a given tree.

3 Experiments & Results

As previously described, feature selection has been applied to a real-life dataset that is described below, together with the obtained results.

3.1 Dataset

The dataset analyzed in present study is based on a sample of all Spanish MNEs registered with the Foreign Trade Institute (ICEX) and from the website www.oficinascomerciales.es, both managed by the Spanish Ministry of Industry, Tourism, and Trade. In order to analyze a representative sample of companies with sufficient autonomy, we restricted the sample to keep only those large and independent enough to conduct and decide their own internationalization strategy. Thus, following a well-established cut-off point in International Business literature, we dropped from the sample those with less than 250 employees. We also dropped those firms with a foreign majority owner controlling more than half of the capital.

It is also important to note the huge impact of the financial crisis on the Spanish economy, which forced many multinational enterprises to sell or postpone international operations in order to focus on the problems of the home market. To avoid distortions in the results due to this exogenous effect, we took the year 2007 as our base year. Overall, the sample consists of 164 firms investing in 119 countries worldwide. Unfortunately, Afghanistan, Andorra, Puerto Rico, and São Tomé and Príncipe are not included in the sample due to a lack of data. In addition, Serbia, Montenegro, and Kosovo are included as a group because at the time of the study they constituted a single country.

For the firms and countries above mentioned, the following data about each one of the cases of international presence were collected (further details about the different features can be found in [9]):

- Firm sector: 5 binary features stating the economy sector the firm belongs to (manufacturing, food, construction, regulated and others).
- Firm product diversification: 3 binary featuring (non-diversified, related or unrelated diversification).
- Other firm characteristics: Number of employees, Return on Equity, number of countries where the firm operates, and whether or not the firm is included in a stock market.
- Host country characteristics: Gross Domestic Product (GDP) growth, total inward FDI, population, unemployment.
- Geographic and psychic distance stimuli between home and host countries [7]. The education distance stimulus is based on differences on literacy rate and enrolment in second and third-level education. The industrial development stimulus takes into account differences in ten dimensions such as in energy consumption, vehicle ownership, employment in agriculture, number of telephones and televisions, etc. The language stimulus is based on the differences between the dominant languages and the bilateral influence of each country's major language in the other country. The democracy stimulus includes differences in political rights, civil liberties and POLCON and POLITY IV indices. The political ideology stimulus is based on the ideological leanings of the chief executive's political party and the largest political party in the government. Finally, the religion stimulus is calculated based on the differences between the dominant religions and the bilateral influence of each country's dominant religion in the other country.

As a result, a dataset containing 10004 samples was obtained, including 25 features that are described in Table 2.

Table 2. Complete list of features (related to the host country and the firm itself) in the original data set.

	#	Feature Name		#	Feature Name
Country	1	Vicarious Experience	Country	14	Psychic Distance - Social System
	2	Vicarious Experience Same Sector		15	Psychic Distance - Religion
	3	Vicarious Experience Different Sector		16	Unemployment
	4	Manufacturing		17	FDI/GDP
	5	Food		18	GDP Growth
	6	Construction		19	Population (Log)
	7	Regulated	Firm	20	Employees
	8	Financial		21	ROE
	9	Geographic Distance (Log)		22	Stock Market
	10	Psychic Distance - Education		23	Related Diversification
	11	Psychic Distance - Industrial Development		24	Unrelated Diversification
	12	Psychic Distance - Language		25	Number of Countries
	13	Psychic Distance - Democracy			

3.2 Results

As described in Section 2, Wrapper FS has been performed by using standard GA as searching method. For it, the most usual parameters were tuned (as stated in Table 3) in 81 different combinations.

Table 3. Set values for the GA parameters during experimentation.

Parameter	Set values
Number of Generations	10, 15, 20
Population Size	10, 20, 30
Crossover Probability	0.3, 0.6, 0.9
Mutation Probability	0.033, 0.06, 0.1
Selection Scheme	Tournament

Additionally, to improve reliability or results, the genetic algorithm was executed 10 times (iterations) with the same values for the parameters above. The fitness function (error rate of the classifier) was applied as the criterion to select the best feature subset. Accordingly, the subset of features with the lowest error rate (highest classification performance) has been selected in each experiment.

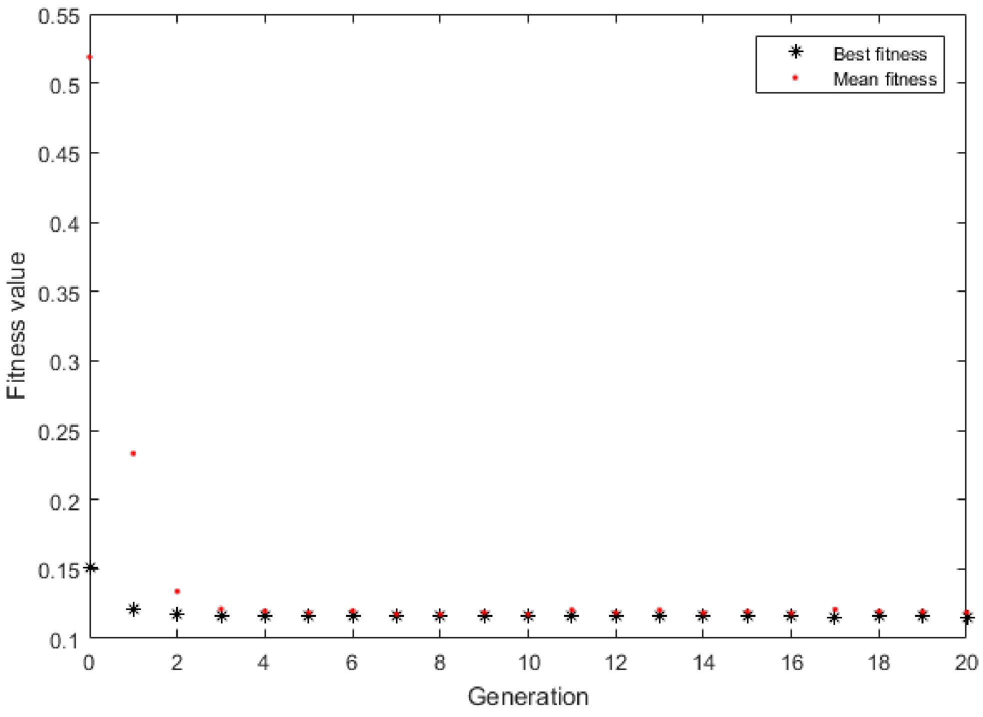
After running the experiments, the best individuals were identified, obtained with the parameter values in the table below. Obtained classification error in the case of SVM was 0.114, while in the case of RF it was 0.109.

Table 4. Set values for the parameters of the GA obtaining the best individual.

	Set values	
Parameter	**SVM**	**RF**
Number of Generations	20	10
Population Size	30	20
Crossover Probability	0.9	0.9
Mutation Probability	0.033	0.1
Selection Scheme	Tournament	Tournament

The fitness values (classification error) for the different generations in the execution of the GA that found the best individual in the case of SVM are shown in Figure 1, for both the best individual and the mean value of each generation.

Fig. 1. Evolution of the fitness values in the GA execution (SVM classifier).



The best individual, in the case of SVM, is composed of the following features: “Vicarious Experience”, “Vicarious Experience Same Sector”, “Vicarious Experience Different Sector”, “Manufacturing”, “Food”, “Geographic Distance (Log)”, “Psychic Distance - Industrial Development”, “FDI/GDP”, “Employees”, “Related Diversification”, and “Number of Countries”. In the case of RF, it is composed of the following features: “Manufacturing”, “Food”, “Construction”, “Geographic Distance (Log)”, “Psychic Distance – Language”, “Psychic Distance – Democracy”, “Psychic Distance - Social System”, “Psychic Distance – Religion”, “FDI/GDP”, “GDP Growth”, “Population (Log)”, “Employees”, “ROE”, “Stock Market”, “Related Diversification”, “Unrelated Diversification”, and “Number of Countries”. The subset of features that are present in these two best individuals are: “Manufacturing”, “Food”, “Geographic Distance (Log)”, “FDI/GDP”, “Employees”, and “Related Diversification”.

From the enterprise management perspective, it can be said that these results are in line with recent contributions to the International Business which emphasize the relevance of distance [2], notably geographic but also some, but not all, dimensions of psychic distance. However, it is also worth noting that not all psychic distance stimuli seem to be particularly significant, and only the Industrial Development psychic distance stimulus is included among the best features. Besides, it is worth noting the relevance of several firm-level features such as employees, related diversification and two sector features, manufacturing and food. In contrast, the ratio FDI/GDP, is the only host country-level feature that appears in the common list of best individuals. These results emphasize how human resources are a critical factor to internationalize. The results also show that a related product diversification strategy is a relevant determinant, as the complementarities between products can make it easier for companies to achieve synergies and enter new markets. Further, in addition to the distance between the home and the host country, a key characteristic of the local economy is the degree of openness to foreign investments. Finally, it is worth noting that the results are not consistent regarding the relevance of vicarious experience. On the one hand, the results from SVM demonstrate the relevance and multi-dimensional aspect of experience [1] and the benefits that firms can obtain from observing other firms from the same nationality [13]. In contrast, the results from RF do not highlight any dimension of vicarious experience among the best individuals.

Regarding the number of features in the obtained searches, the best subset comprises 11 and 17 for SVM and RF respectively. On the other hand, the average number of features in the best individuals obtained in all the iterations for the same parameters (Table 4) amounts to 13.1 (SVM) and 15.7 (RF). It is worth highlighting that the proposed solution significantly reduces the amount of features to be considered from the 25 features in the original dataset (reduction of 56% and 37.2% respectively).

On the other hand, for a comprehensive analysis, each one of the original features has been considered, from an individual standpoint. To do so, it has been calculated the percentage of best solutions (from all the 10 iterations) that includes each one of the original features, both for SVM and RF. Additionally, the sum of percentages has also been calculated, as is shown in Table 5.

Table 5. Inclusion rate (%) of each individual data feature in the obtained best individuals for all the iterations. Feature number (in the second column) corresponds to that introduced in Table 2 (#).

#	Feature Name	%		
		SVM	**RF**	**SUM**
20	**Employees**	**80**	**100**	**180**
2	**Vicarious Experience Same Sector**	**100**	**80**	**180**
25	**Number of Countries**	**100**	**70**	**170**
23	**Related Diversification**	**80**	**90**	**170**
4	**Manufacturing**	**90**	**70**	**160**
24	**Unrelated Diversification**	**80**	**70**	**150**
9	Geographic Distance (Log)	40	90	130
10	Psychic Distance - Education	60	70	130
17	FDI/GDP	60	60	120
7	Regulated	60	60	120
21	ROE	20	100	120
22	Stock Market	50	70	120
5	Food	50	60	110
18	GDP Growth	40	70	110
12	Psychic Distance - Language	50	60	110
3	Vicarious Experience Different Sector	70	40	110
16	Unemployment	50	50	100
6	Construction	0	90	90
13	Psychic Distance - Democracy	50	40	90
1	Vicarious Experience	70	20	90
15	Psychic Distance - Religion	30	50	80

8	**Financial**	**10**	**60**	**70**
19	**Population (Log)**	**20**	**40**	**60**
14	**Psychic Distance - Social System**	**20**	**40**	**60**
11	**Psychic Distance - Industrial Development**	**30**	**20**	**50**

From the results in the table above, the most significant features (identified by the highest sum of rates of inclusion in best individuals) can be selected, in order to take them into account for an internationalization decision. Complementary, same results let us identify the features to be discarded due to their low relevance (lowest inclusion rate), too. According to that, the most important features (being included in most subsets) are (in decreasing order of importance): “Employees”, “Vicarious Experience Same Sector”, “Number of Countries”, “Related Diversification”, “Manufacturing”, “Unrelated Diversification”. Consistent with the previous results, the pivotal role of human resources is identified as one of the most important characteristics. In turn, and somewhat different from the previous results, the inclusion of vicarious experience in the best feature subset shows that firms can also learn good practices and avoid mistakes by observing the internationalization of other firms [14]. While firms can learn from firms in other sectors, the experience of other firms in the same sector is likely to be easier to assimilate and fit the managers´cognitive models [13]. Also, vicarious experience from firms in the same sector tends to be easier to integrate because of the similarities in resources and technologies [29], which also facilitates finding suitable partners abroad [15]. Finally, overcoming the resistance from stakeholders who oppose to some decisions is also likely to be easier when the strategy of the firm is based on what other firms in the sector are doing [16]. The amount of countries where the firm operates increases the international exposure to different environments that allow the firm to benefit from learning opportunities and a pool of knowledge that can be applied to develop capabilities that can be used in different contexts [30, 31]. Another important features, also consistent with our previous results are that of the manufacturing sector, and product diversification. Interesting, in the case of the latter, not only related diversification but also unrelated diversification are highlighted as most significant features.

On the other hand, those features that are least important for an internationalization decision have also been identified. That is the case of “Financial”, “Population (Log)”, “Psychic Distance - Social System”, and “Psychic Distance - Industrial Development”, that have been included by very few of the best individuals. These two set of results combined show that the majority of the body of Spanish firms internationalizing are located in the manufacturing sector. This, however, is not incompatible with the fact that there are a few well-known banks from Spain who are internationally very active, for instance Banco Santander or BBVA. While these two financial institutions have many investment throughout the world (29 and 22 respectively), many other banks from Spain have operations only in one or two neighbouring countries (most notably in Portugal) such as CaixaNova, Caja de Badajoz or Caixa Galicia. Similarly, while there are some firms conducting international operations of construction (ACS being the best example), for many firms in this sector activities in the home country are more important and the share of internationalization is much lower than in the case of manufacturing, an industry with powerhouses such as Inditex (the owner of the brand Zara) with operations in 67 countries or Mango in 89, but also other less-known firms such as Camper (50 countries), Valdepesa (43 countries), Teka (40 countries), Maxam (35 countries) or Tous (34 countries).

When comparing results from each one of the classifiers (SVM and RF) it can be said that in general terms, they agree on the most important features. However, some disagreement can be identified in the case of features “ROE” and “Construction”. The first one was included in only 20% of the best individuals according to SVM but in all of the best individuals according to RF. Similarly, “Construction” was not included in any of the best individuals obtained by SVM but it was included in almost all (90%) individuals obtained by RF. There is not any feature in the opposite circumstances (highest inclusion rate in the case of SVM but lowest inclusion rate in the case of RF).

3.3 Comparison

As previously stated, FS (both SVM- and RF-based) results are compared to those previously obtained by applying Standard, Rare Event, and Conditional Logistic regression, that are available in a previous paper [9]. The main results presented in that paper show that distance plays a critical role in the internationalization strategy of Spanish MNEs. Specifically, all psychic distance stimuli and also geographic distance are significant determinants of Spanish FDI. Besides, the results of that paper also show that vicarious experience, either from firms in the same sector, in different sector, or the total amount, are also significant in the models. Furthermore, larger population, unemployment and GDP growth are significant country-level determinants of FDI decisions. Finally, the number of employees and the international experience measured by the number of countries where the firm runs operations are significant firm-level determinants of FDI decisions.

Our FS results are to some extent consistent with this analysis, showing the critical relevance of human resources on internationalization. Also, both sets of results emphasize the relevance of multiple, but not all, dimensions of experience, both the firm´s own experience and vicarious experience from other firms with the same nationality. Finally, the results are again consistent in underlining the relevance of distance. However, some interesting differences have also emerged from our FS analysis. In particular, the regressions showed that both geographic and most dimensions of psychic distance were critical factors of internationalization. Our FS results clearly identify geographic distance as a key determinant, but psychic distance plays a much minor role and only the psychic distance stimuli in education appears among the features with a higher inclusion rate. Further, the country-level determinants of FDI in the original regressions, population, unemployment and GDP growth, have a minor role here and do not show much explanatory power compared to other variables. In contrast, the FS results draw the attention on the critical role of product diversification, especially related diversification, and notably the sector of the firm, a factor that was not consistently significant in the regressions. Thus, the FS results show the high relevance of the Manufacturing sector, where some of the flag-ships of Spanish internationalization can be found with investments in 60+ countries, but also many other less famous firms also operating in several dozens of locations. In contrast, a sector with a very large weight in the Spanish economy and dominated by very large firms such as Financial, has a much less relevance as determinants of internationalization. While these large corporations are indeed multinationals and have even become global leaders in their respective sectors, the level of concentration of international operations is very high and only found in these specific cases, whereas the rest of firms in the sector remain largely domestic or with operations in a very limited number of destinations.

4 Conclusions and Future Work

From the results presented in section 3.2, it can be concluded that some data related to the host country as well as the firm itself are key features related to the internationalization strategy of Spanish firms. According to that, the most important features driving the internationalization decision of managers have been identified, as well as those features with a lowest level of relevance. Obtained results are consistent with previous work on this same dataset and with the state of the art.

In future work, some other feature selection models will be applied to the same dataset to better understand its nature and gaining deep knowledge of the internationalization strategy of Spanish firms. Additionally, data from some other home countries will also be analysed to compare results in a transnational study.

Acknowledgments

The work was conducted during Álvaro Herrero's research stay at KEDGE Business School in Bordeaux (France).

References

1. Jiménez, A., et al., *The Multi-faceted Role of Experience Dealing with Policy Risk: the Impact of Intensity and Diversity of Experiences.* International Business Review, 2018. **27**(1): p. 102-112.
2. Zaheer, S., M.S. Schomaker, and L. Nachum, *Distance without Direction: Restoring Credibility to a Much-loved Construct.* Journal of International Business Studies, 2012. **43**(1): p. 19.
3. Hofstede, G., G.J. Hofstede, and M. Minkov, *Cultures and Organizations: Software of the Mind*. 2010, McGraw-Hill: New York.
4. Shenkar, O., *Cultural Distance Revisited: Towards a More Rigorous Conceptualization and Measurement of Cultural Differences.* Journal of International Business Studies, 2001. **32**(3): p. 519-536.
5. Tung, R.L. and A. Verbeke, *Beyond Hofstede and GLOBE: Improving the Quality of Cross-cultural Research.* Journal of International Business Studies, 2010. **41**(8): p. 1259-1274.
6. Johanson, J. and J.-E. Vahlne, *The Internationalization Process of the Firm: a Model of Knowledge Development and Increasing Foreign Market Commitments.* Journal of International Business Studies, 1977. **8**(1): p. 23-32.
7. Dow, D. and A. Karunaratna, *Developing a Multidimensional Instrument to Measure Psychic Distance Stimuli.* Journal of International Business Studies, 2006. **37**(5): p. 575-577.
8. Dow, D. and S. Ferencikova, *More than just national cultural distance: Testing new distance scales on FDI in Slovakia.* International Business Review, 2010. **19**(1): p. 46-58.

9. Jiménez, A. and D. de la Fuente, *Learning from Others: the Impact of Vicarious Experience on the Psychic Distance and FDI Relationship.* Management International Review, 2016. **56**(5): p. 633-664.

10. Clarke, J.E., R. Tamaschke, and P.W. Liesch, *International experience in international business research: A conceptualization and exploration of key themes.* International Journal of Management Reviews, 2013. **15**(3): p. 265-279.

11. Cyert, R.M. and J.G. March, *A behavioral theory of the firm*. Vol. 2. 1963, Malden, MA: Blackwell. 169-187.

12. Levitt, B. and J.G. March, *Organizational learning.* Annual Review of Sociology, 1988. **14**(1): p. 319-338.

13. Jiang, G.F., G.L. Holburn, and P.W. Beamish, *The impact of vicarious experience on foreign location strategy.* Journal of International Management, 2014. **20**(3): p. 345-358.

14. Terlaak, A. and Y. Gong, *Vicarious learning and inferential accuracy in adoption processes.* Academy of Management Review, 2008. **33**(4): p. 846-868.

15. Meyer, K.E. and H.V. Nguyen, *Foreign investment strategies and sub‐national institutions in emerging markets: Evidence from Vietnam.* Journal of Management Studies, 2005. **42**(1): p. 63-93.

16. Guillén, M.F., *Structural inertia, imitation, and foreign expansion: South Korean firms and business groups in China, 1987–1995.* Academy of Management Journal, 2002. **45**(3): p. 509-525.

17. John, G.H., R. Kohavi, and K. Pfleger. *Irrelevant Features and the Subset Selection Problem*. in *11th International Conference on Machine Learning*. 1994. Morgan Kauffman.

18. Kohavi, R. and G.H. John, *Wrappers for Feature Subset Selection.* Artificial Intelligence, 1997. **97**(1–2): p. 273-324.

19. Goldberg, D.E., *Genetic Algorithms in Search, Optimization, and Machine Learning*. 1989: Addison-Wesley.

20. Siedlecki, W. and J. Sklansky, *A Note on Genetic Algorithms for Large-scale Feature Selection.* Pattern Recognition Letters, 1989. **10**(5): p. 335-347.

21. Kramer, O., *Genetic Algorithm Essentials*. Studies in Computational Intelligence. 2017, Cham: Springer International Publishing. 3-10.

22. Boser, B.E., I.M. Guyon, and V.N. Vapnik. *A training algorithm for optimal margin classifiers*. in *5th Annual Workshop on Computational Learning Theory*. 1992. ACM.

23. Cortes, C. and V. Vapnik, *Support-Vector Networks.* Machine Learning, 1995. **20**(3): p. 273-297.

24. Byun, H. and S.-W. Lee. *Applications of Support Vector Machines for Pattern Recognition: A Survey*. 2002. Berlin, Heidelberg: Springer Berlin Heidelberg.

25. Safavian, S.R. and D. Landgrebe, *A Survey of Decision Tree Classifier Methodology.* IEEE Transactions on Systems, Man and Cybernetics, 1991. **21**(3): p. 660-674.

26. Sreerama, K.M., *Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey.* Data Mining and Knowledge Discovery, 1998. **2**(4): p. 345-389.

27. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.

28. Probst, P. and A.-L. Boulesteix, *To Tune or Not to Tune the Number of Trees in Random Forest?* Journal of Machine Learning Research, 2018. **18**(181): p. 1-18.

29. Ingram, P. and T. Simons, *The Transfer of Experience in Groups of Organizations: Implications for Performance and Competition.* Management Science, 2002. **48**(12): p. 1517-1533.

30. Powell, K.S. and M. Rhee, *Experience in Different Institutional Environments and Foreign Subsidiary Ownership Structure.* Journal of Management, 2016. **42**(6): p. 1434-1461.

31. Zhou, N. and M.F. Guillén, *From home country to home base: A dynamic approach to the liability of foreignness.* Strategic Management Journal, 2015. **36**(6): p. 907-917.