Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC - Central Bank of Armenia Workshop on *"External Sector Statistics"*

Dilijan, Armenia, 11-12 June 2018

Imputation techniques for the nationality of foreign shareholders in Italian firms[1]

Andrea Carboni and Alessandro Moro,

Bank of Italy

[1] This paper was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

Imputation Techniques for the Nationality of Foreign Shareholders in Italian Firms

Andrea Carboni[1], Alessandro Moro[2]

Abstract

In order to estimate the Foreign Direct Investments (FDI) item of the Italian Balance of Payments and International Investment Position, the Bank of Italy realises a direct sample survey for the non-financial and insurance companies; the survey's methodology and definitions follow the international standards defined in the IMF BPM6 and OECD BD4 and best practices. In the sampling strategy a stratified sample is used, considering among the other stratification variables the presence or absence for the firm of FDI relationships (inward and outward). In general information on FDI inward is available in administrative data: in fact, the Italian enterprises annually report to the Chambers of Commerce the list of theirs shareholders (the so-called "*Elenco Soci*" in the *Infocamere* database). While this information is used by the Bank of Italy to identify the list of enterprises with FDI inward, it is in many cases incomplete as the nationality of the shareholders is missing. In order to solve this problem, the present paper proposes an algorithm of imputation when the nationality of foreign firms is unknown and the only relevant information is represented by the name of the corporations. The procedure works as follows: firstly, the name of the different firms is decomposed in its elementary words and the most frequent ones are selected; then, for each selected word, a dummy variable is constructed taking value one when that word is included in the name of the firms and zero otherwise; finally, a statistical model is estimated linking the nationality of the firms to those dummy variables. The out-of-sample analysis reveals that this procedure is able to obtain a high percentage of correct classification, with an almost perfect discrimination between Italian and foreign firms.

Keywords: Foreign Direct Investments; Balance of Payments; Machine learning; Statistical learning; Imputation techniques; Classification problems; Logistic regressions; Multinomial models.

JEL classification: C10, C80.

The views expressed in the paper are those of the authors and do not involve the responsibility of the Bank of Italy.

[1] Bank of Italy, Statistical Data Collection and Processing Directorate, External Statistics Division, E-mail: andrea.carboni@bancaditalia.it

[2] Bank of Italy, Statistical Data Collection and Processing Directorate, External Statistics Division, E-mail: alessandro.moro2@bancaditalia.it

Contents

1. Introduction

In order to estimate the Foreign Direct Investments (FDI) item of the Italian Balance of Payments and International Investment Position, the Bank of Italy realises a direct sample survey for the non-financial and insurance companies; the survey's methodology and definitions follow the international standards (IMF, 2009; OECD, 2008). In the sampling strategy a stratified sample is used, considering as stratification variables the dimension, the geographical location and the presence or absence for the firm of FDI relationships (inward and outward).

In particular, for the outward side, strata are given by three dimensional classes (according to the total asset of the enterprises), four geographic areas, and the presence/absence of FDI abroad. For the inward side, instead, the strata are defined considering the three dimensional classes, the four geographic areas and then the two binary variables related to FDI: presence/absence of FDI abroad and presence /absence of at least a foreign shareholder. The two designs are reconciled by means of overlapping techniques, that ensure the possibility to realise simultaneously an efficient allocation for both samples (FDI inward and FDI outward).

In general, information on FDI is available in administrative data. For the FDI outward, the information is reported annually in the supplementary note of the financial statement; the Bank of Italy acquires this information in a structured database provided by the Cerved Group. On the other hand, regarding the FDI inward, the Italian enterprises annually report to the Chambers of Commerce the list of theirs shareholders (this is the so-called "*Elenco Soci*" in the *Infocamere* database). While this information is used by the Bank of Italy to identify the list of enterprises with FDI inward, it is in many cases incomplete as the nationality of the shareholders is missing. The nationality of the foreign shareholders would be an important variable in order to improve the stratification, and consequently the efficiency, of the sampling scheme of the survey. Moreover, this piece of information can be used in the grossing up procedure for calibrating the statistical estimates of FDI since the foreign investments data are disseminated with the geographical detail of the shareholder.[3]

Actually, the current procedure employed by the Bank of Italy only discriminates between Italian and foreign enterprises. In fact, a "dictionary" has been constructed considering all the words contained at least 20 times in the names of a sample of 15,000 foreign firms (extracted from an external source: the Cerved database). If the denomination of the shareholder in our database contains at least one of the words in the dictionary, it is classified as a foreign investor; otherwise, it is considered as Italian. The percentage of correct classification is around 80% but the outcome of the procedure is simply binary (Italian vs foreign enterprise).

In order to solve these problems, the present paper proposes a machine learning algorithm[4] for the imputation of the nationality of foreign shareholders when the only relevant information is represented by the name of the corporations. The procedure works as follows: firstly, the name of the different firms is

[3] For details about the estimation of FDI and sampling techniques adopted in the compilation of the Balance of Payment of Italy, see http://www.bancaditalia.it/statistiche/manuale_bop_19mag16.pdf.

[4] There is a growing interest on the practical applications of machine learning algorithms in central banks. For a recent review, see Chakraborty and Joseph (2017).

decomposed in its elementary words and the most frequent words are selected; secondly, for each selected word, a dummy variable is constructed taking value one when that word is included in the name of the firms; finally, a model is estimated linking the nationality of the firms to those dummy variables.

In order to train the algorithm, it is necessary to use a database in which the nationality of the single enterprise is known. To this scope, the database Orbis, provided by Bureau van Dijk, appears appropriate. In fact, Orbis contains financial statement information about almost all the equity capital enterprises of the world: in total about 300 million of firms from all over the world, merging information from about 80 qualified data providers.

Given the prevalence of Italian shareholders in the *Infocamere* database (around 90%), which is the dataset to whom the algorithm should be applied, an unbalanced set is drawn from the Orbis database using the prior information, evaluated from our past direct surveys, about the proportion of the different nationalities of the shareholders of the Italian enterprises. Since in the obtained sample the Italian firms represent the vast majority of cases, a two-step algorithm has been adopted. In the first step, a logit model (Cox, 1958) is estimated with the aim of identifying the Italian firms. Then, in the second step, on the observations identified as foreign enterprises in the first step, a multinomial model (McFadden, 1974) is fitted.

In sum, the out-of-sample analysis reveals that this algorithm is able to obtain a high percentage of correct classification (about 98%), with an almost perfect discrimination between Italian and foreign firms. The results of the model can be further improved by combining the two-step approach with a popular dimension reduction technique used in text mining, i.e. the Singular Value Decomposition (Deerwester et al., 1990). The paper also shows that the performance of the proposed model is better than that of alternative machine learning algorithms, like the decision trees and random forests (for a review of these techniques, see Friedman et al., 2001).

The procedure described in this paper can enhance the quality of the Bank of Italy's statistics of FDI in two ways: (a) ex-ante, since the imputed nationality of firms in the *Infocamere* database, evaluated with the model, is a relevant stratification variable for the sampling scheme of the direct surveys, much richer than the binary classification obtained with the current procedure; (b) ex-post, since the model can be used in the FDI grossing-up procedure to improve the allocation of the FDI estimates to the different counterpart countries. In fact, this allocation can be performed on the model-based imputed nationality evaluated on the entire *Infocamere* database and not only on the sampled firms, as in the procedure currently adopted. Moreover, the allocation can also be realised in a fuzzy prospective: given the model probabilities that an enterprise is resident in different countries, the equity capital can be split among those countries with weights equal to these probabilities.

The rest of the paper is organised as follows: Section 2 presents the proposed procedure from a methodological point of view; Section 3 describes the data; Section 4 summarises the main results; Section 5 proposes an improved version of the two-step algorithm that combines it with more sophisticated text mining techniques; some robustness checks are described in Section 6, in which the results of the model are compared to those obtained with different sample compositions

and alternative machine learning approaches; finally, Section 7 summarises the main conclusions.

2. The Algorithm

The procedure currently used at the Bank of Italy in order to discriminate between Italian and foreign enterprises is such that if the name of the investor contains a word included in a predefined list of foreign words, it is considered as a non-resident enterprise, otherwise it is considered as an Italian (resident) one. This model has shown an overall accuracy of about 80 percent of correct classification of the shareholders. However, this approach does not provide any information about the nationality of the shareholders and yields only a binary response (resident/ non-resident). Moreover, the response is not based on a rigorous inferential statistical model. Against this background, this work aims at improving the current set up by overcoming the above limits.

The aim of the proposed new algorithm is the identification of the nationality of a given firm when the only observed feature is represented by the name of the corporation. The problem can be formalised by considering a set of *N* firms, each of them characterised by the couple *(Name, Country)*. The final objective is the identification of a predictive model $f(.)$ relating the probability of belonging to a given country $\pi_{Country}$ to the name of the firm:

$$\pi_{Country} = f(Name) \quad (1)$$

After a preliminary data cleaning step, in which the punctuation and the special characters are removed, the procedure begins with the decomposition of the name of each firm in its elementary words. Then, the frequencies of the different words with at least two characters are evaluated and only the most frequent *K* ones are selected.[5] For each selected word, a dummy variable is constructed taking value 1 in correspondence of a given firm when the considered word is included in the name of the firm, and 0 otherwise: therefore, $d_{i,j} = 1$ if the name of the i-th firm includes the j-th word; $d_{i,j} = 0$, elsewhere (Table 1 presents some examples of dummy construction). These dummy variables constitute the regressors of all the subsequent statistical models.

Table 1: Examples of dummy variables

Name	SRL	Societa	SPA	SA	Ltd	GMBH	PTY
Trelpa SA	0	0	0	1	0	0	0
Sud Chemie Australia PTY Ltd	0	0	0	0	1	0	1
Tarigia SRL	1	0	0	0	0	0	0
NGM Verwaltungs GMBH	0	0	0	0	0	1	0

[5] After a sensitivity analysis exercise, *K* is set to a value of 50 words in the next application of the algorithm. In fact, different choices of *K* have also been proved but do not alter the main results. In particular, the inclusion of a higher number of words, and consequently of dummies, does not increase the performance of the procedure.

Since in our database the Italian firms represent the vast majority of cases, the procedure is articulated in two-steps. In the first one, a logit model is estimated with the scope of identifying the Italian firms. In particular, the probability that the nationality of the i-th firm is Italian, $\pi_{i,IT}$, is given by:

$$\pi_{i,IT} = \frac{\exp(\beta'_{IT} d_i^{(1)})}{1 + \exp(\beta'_{IT} d_i^{(1)})} \quad (2)$$

where $d_i^{(1)}$ is the K-dimensional vector of dummy variables observed on the i-th firm, while β_{IT} represents the vector of coefficients specific to the probability of being Italian. If $\pi_{i,IT} > 0.5$, then, the i-th firm is classified as Italian.

On the observations that have been considered as foreign in the first step, the selection of the most frequent K words is repeated and the corresponding dummy variables are created: these new words should be more representative of the foreign countries given the selection of Italian corporations in the first step. Then, a multinomial logit model[6] is estimated, in which the probability that the i-th firm belongs to the h-th nationality is given by:

$$\pi_{i,h} = \frac{\exp(\beta'_h d_i^{(2)})}{\sum_{h=1}^{H} \exp(\beta'_h d_i^{(2)})} \quad (3)$$

in which $d_i^{(2)}$ is the new K-dimensional vector of dummy variables observed on the i-th firm, while β_h represents the country-specific vector of coefficients. The predicted nationality for the i-th corporation is the one to whom is associated the maximum probability.

3. The Sample Selection

As already mentioned, the algorithm is trained using the Bureau Van Dijk's Orbis database, which contains, beyond many balance information, the name and nationality of more than 300 million world enterprises.

From this source a large set of about 1.7 million enterprises is selected from the most developed countries (G-20) and from the main partners of Italy in terms of direct investments abroad. Since every year the Bank of Italy collects from a sample of about 6,000 enterprises individual detailed data about the FDI (both outward and inward), this piece of information has been taken into account in order to update, ex-post, the fitted probabilities when applied to the *Infocamere* database, or, ex-ante, to define directly the sample for the training of the model. Both approaches are tested.

Firstly, balanced samples are used in which each country has the same proportion, namely the same number of units. In this case the model results should be adjusted (ex-post) to take into account the priors probability applying the Bayes Theorem. The alternative approach consists of using the a priori information about the FDI breakdown for building the sample; in this case the units for each country

[6] The model is estimated with the *multinom* function of the nnet package in R (Ripley and Venables, 2016), which is based on neural network techniques.

are proportional to the weights of the nations in the distribution of the equity capital of the Italian enterprises among the different shareholders. Overall, this second option improves the results.[7]

Therefore, in the best training sample, 180,000 enterprises are extracted from the Orbis database with a country-specific probability of inclusion equal to these known priors. In particular, since about 90% of the shareholders in the *Infocamere* database are Italians, the sample is constructed in order to reproduce this proportion. The remaining 10% of the sample has been selected according to the frequencies of firms by country evaluated on our past samples (see Table 2): the most frequent countries are Luxembourg (21% of the foreign shareholders are from that country), Netherlands (18%), France (12%), Germany (10%), Great Britain (10%), while China is the last relevant country considered (1%). Firms from countries members of the OECD and G20, different from those listed in Table 2, are randomly selected in order to cover the residual component, called "Others" (OT).

Table 2: FDI in Italian enterprises by country

Country	FDI in Italian enterprises (Euros)	Percentage (%)
LU	19.701.471.920	21,4
NL	38.562.125.473	17,7
FR	23.375.783.193	11,9
DE	5.921.531.842	10,4
GB	24.979.578.528	9,8
CH	4.838.944.447	6,1
ES	3.362.143.938	3,5
US	3.171.216.814	3,1
BE	5.592.011.842	2,6
AT	915.696.442	2,4
JP	929.862.399	1,6
DK	682.030.864	1,5
SE	774.057.108	1,2
CN	48.917.610	0,7
OT	4.248.537.125	0,1

The information extracted from the Orbis database are simply the name and the country of each company. The reason for this choice is that the main objective of the present work is the application of the algorithm to the list of shareholders from the administrative data of the Chambers of Commerce (*Infocamere* database), in which the only available information is represented by the name of corporations. The model is trained using 80% of the sample (training set) constructed from Orbis and it is validated with the remaining 20% (validation set).

A standard cross-validation technique has also been performed, by splitting the sample in five groups of the same cardinality and rotating the test set (the other

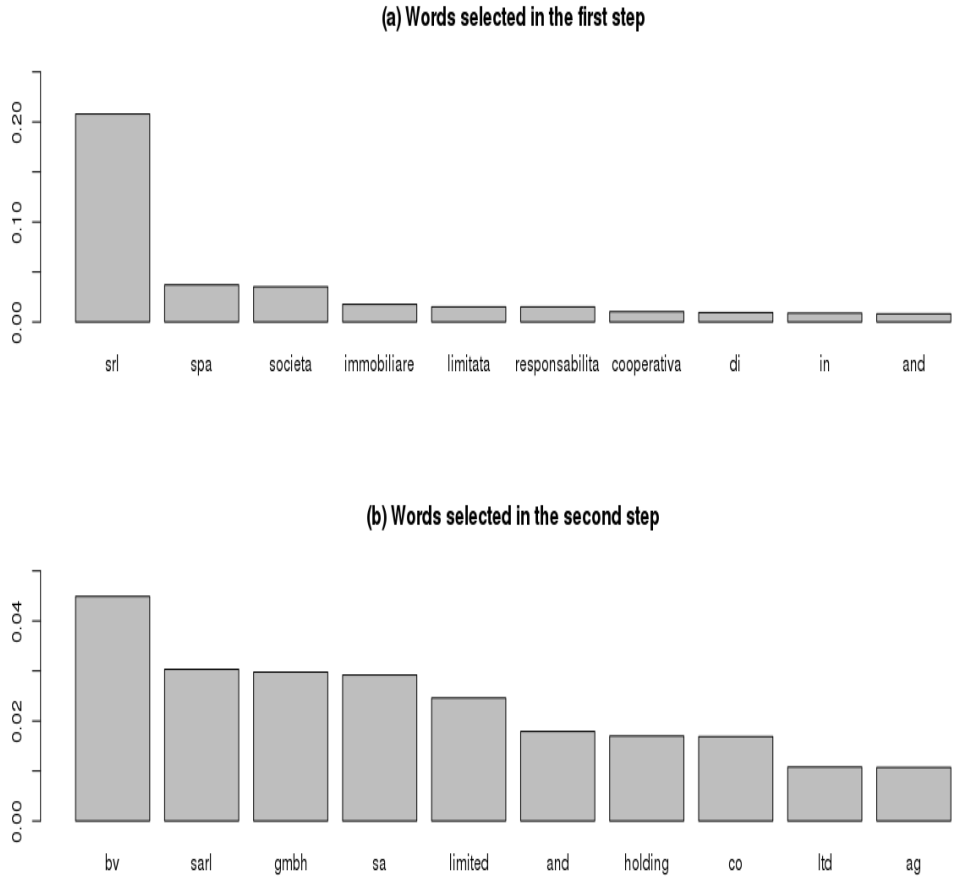
[7] In Subsection 6.1 a robustness check has been carried out with a balanced set.

four groups are used as training set). Results are robust and very similar to the one presented in this document.

4. Results

Given the high proportion of Italian firms in the sample (around 90%), a two-step algorithm that is able to firstly select the Italian units and then to discriminate between different foreign nationalities seems to be an appropriate choice.

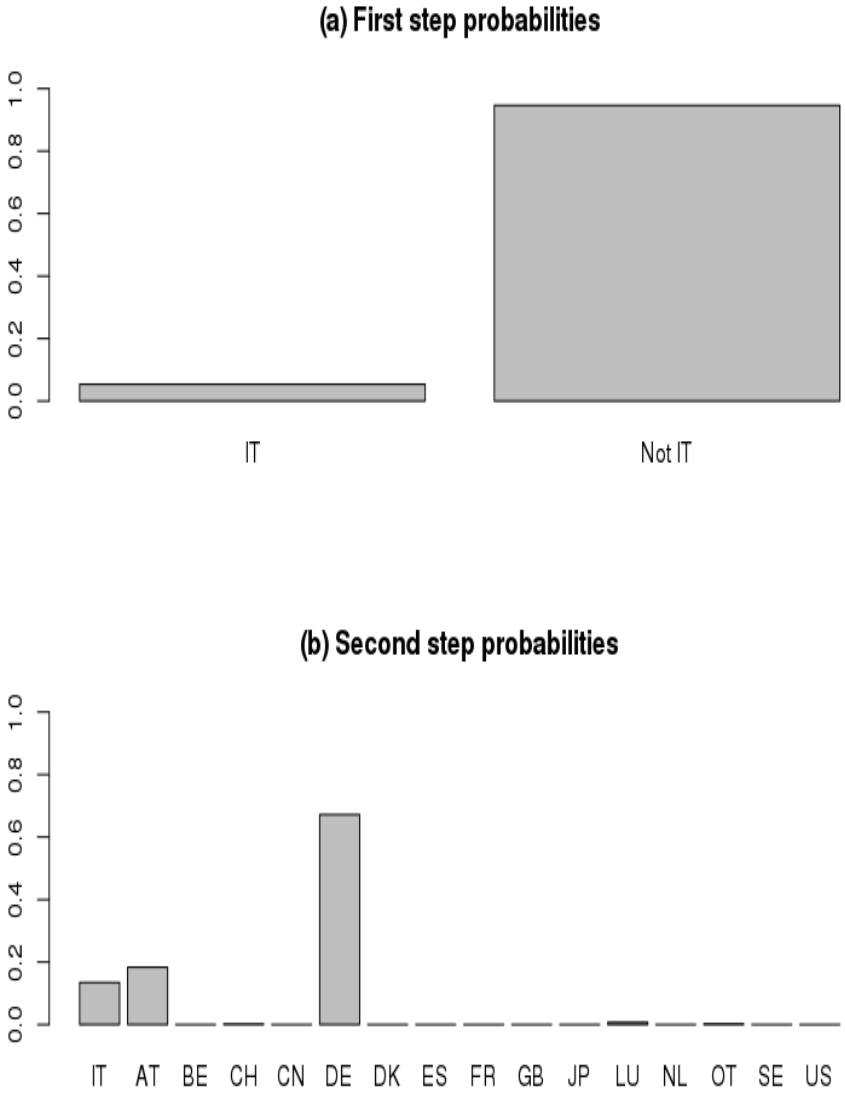
Figure 1: Example of word selection: the first ten most frequent words selected in the first and second step



For a better understanding of how the procedure works, the process of word selection in the two steps is briefly described. The first panel in Figure 1 shows the most frequent ten words selected in the first step of the procedure. As expected, most of them are Italian: they include two abbreviations of legal entities of Italian corporations (SRL and SPA) and other frequent expressions in Italian names of corporations (like "*Società a Responsabilità Limitata*"). Given the specificity of these words, they should be able to identify very well the Italian corporations. Then a logit model is estimated using the dummies associated to these words as regressors.

Once the firms considered as Italian in the first step are filtered out, the process of word selection is repeated: the new most frequent words are displayed in the second panel of Figure 1. In the second step the most frequent words are the abbreviations of legal entities of the firms located in different European countries, such as BV for Netherlands, SARL for Luxembourg, GMBH for Germany and Austria, and frequent English words in international companies, like limited, ltd, holding, and so on. These new words are the inputs in the multinomial model.

Figure 2: Example of a model prediction: first and second step fitted probabilities in the case of a German firm



Once the models in the first and the second step are estimated on the training set, then they are used in order to make out-of-sample predictions on the nationality of any given firm. Figure 2 illustrates the case of the German firm NGM

Verwaltungs GMBH. The first panel of the figure shows the probabilities estimated at the first step from the logit model: it is easy to observe that the model correctly classifies the firm as foreign. In the second step the highest probability predicted by the multinomial model is the one associated to Germany: therefore, the final prediction is correct. It is interesting to notice that the second highest predicted probability is the one of Austria since GMBH is a common legal entity name in that country too.

In order to illustrate the results of the model performance in the validation set, Table 3 shows the confusion matrix: the columns report the true countries, the rows the predicted nationalities. Each column exhibits how often, in percentage terms, a given country is classified correctly and, conversely, when it is confused with another nationality: for example, German firms are classified correctly 90% of cases, 3% of cases they are classified as Swiss, 2% as Austrian, 2% as French and 2% as Italian. It is worth remarking that the highest percentages usually lie on the main diagonal, which means that the model correctly discriminates the different nationalities, especially those with the most relevant FDI in Italian firms (Luxembourg, Netherlands, Germany, Great Britain), that are also more represented in the sample. Moreover, the model presents a perfect discrimination between Italian and foreign firms which explains, given the prevalence of Italian units in the sample, the high performance in terms of overall model accuracy (around 98%).

Table 3: Confusion matrix of the two-step procedure

	TRUE															
FITTED	AT	BE	CH	CN	DE	DK	ES	FR	GB	IT	JP	LU	NL	OT	SE	US
AT	10%	0%	0%	0%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
BE	0%	7%	2%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	2%
CH	0%	0%	41%	0%	3%	0%	1%	6%	2%	0%	0%	2%	0%	2%	27%	0%
CN	0%	2%	0%	14%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%
DE	81%	0%	1%	0%	90%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
DK	0%	0%	2%	0%	0%	82%	0%	1%	1%	0%	0%	1%	0%	1%	0%	0%
ES	0%	0%	0%	0%	0%	0%	59%	1%	0%	0%	0%	1%	0%	3%	0%	0%
FR	0%	76%	24%	7%	2%	9%	4%	66%	5%	0%	3%	1%	2%	21%	0%	45%
GB	0%	0%	1%	7%	0%	0%	0%	0%	80%	0%	6%	0%	0%	18%	0%	0%
IT	10%	9%	16%	7%	2%	9%	10%	11%	5%	100%	13%	6%	1%	10%	8%	6%
JP	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	45%	0%	0%	0%	0%	0%
LU	0%	2%	10%	0%	0%	0%	25%	9%	0%	0%	0%	88%	0%	5%	0%	0%
NL	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	96%	0%	4%	2%
OT	0%	0%	2%	64%	0%	0%	0%	0%	7%	0%	29%	0%	0%	34%	0%	6%
SE	0%	2%	2%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	62%	0%
US	0%	2%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	5%	0%	39%

It is interesting to notice that the low percentages of correct classification are due to cases in which there is a common language and in which the legal entity abbreviations are the same: this explains the reason why the model classifies most of the Austrian corporations as German and most of the Belgian companies as French.

Furthermore, a cross-validation exercise is done rotating the 20% of the sample used for the testing of the model and estimating the coefficients with the remaining 80%. Coefficient estimates and classification results are very stable over these different trials.

5. Further Improvements

Our two-step procedure can be easily improved by considering, instead of the K=50 most frequent words, all the terms in the denominations of firms and by applying a popular dimension reduction procedure used in text mining, i.e. the Singular Value Decomposition (SVD).

The SVD can be interpreted as a sort of Principal Component Analysis applied to the matrix of the dummies associated to the different words. The extracted principal components are then the new regressors of the two-step procedure. Using this method, it is possible to extract information in a more efficient way from the denominations of firms. However, this enrichment of the information set comes with a cost, i.e. the interpretation of the input variables is less clear.

More precisely, the SVD method can be described as follows. Let *D* be the *NxM* matrix of dummies, where *N* is the number of firms and *M* the number of terms, and assume *r* the rank of the matrix. The singular value decomposition of the transpose of *D* is:

$$D' = U\Sigma V' \ (4)$$

where *U* and *V* are orthogonal matrices with dimensions *Mxr* and *Nxr*, respectively, and Σ is a *rxr* diagonal matrix, whose diagonal elements are ordered, by convention, from the largest to the smallest ones.

The power of this representation comes from the property of the SVD regarding approximating matrices. In fact, we can note that expression (4) can be viewed as a sum of rank one matrices:

$$D' = \sum_{i=1}^{r} \sigma_i u_i v'_i \ (5)$$

in which u_i and v_i are the columns of *U* and *V*, respectively. The best rank-*k* approximation of *D*, with $k \leq r$, is given by:[8]

$$D_k{}' = \sum_{i=1}^{k} \sigma_i u_i v'_i = U_k \Sigma_k V_k{}' \ (6)$$

Therefore, it is possible to map the *M*-dimensional vector d_i of dummy variables for enterprise *i* to a lower dimensional subspace using matrix U_k:

$$\hat{d}_i = U_k' d_i \ (7)$$

In fact, $\hat{d}_i$ is a *k*-dimensional vector obtained through a linear combination of the original elements of d_i using as weights the elements of U_k. In this way the new input variables for the subsequent statistical models are obtained.

The revised version of the proposed two-step algorithm becomes:

- application of the SVD decomposition to the matrix of the dummies associated to the words contained in the denomination of all the firms in the sample and use of the extracted components as inputs in the logit model (first step);
- estimation of the logit model and classification of the Italian companies;
- application of the SVD decomposition to the matrix of the dummies associated to the words contained in the denomination of the firms classified as foreign in the first

[8] After a sensitivity analysis exercise, *k* is set to a value of 50 in the next application of the algorithm.

step and use of the extracted components as inputs in the multinomial model (second step);

- estimation of the multinomial model and classification of the foreign companies.

The confusion matrix evaluated on the validation set of this improved version of the two-step procedure is reported in Table 4. It is possible to observe that the percentages of correctly classified enterprises, i.e. those on the main diagonal of the confusion matrix, generally increase and the countries with the highest levels of FDI (Luxembourg, Netherlands, France, Germany, Great Britain) have percentages of correct classification higher than 89%.

Table 4: Confusion matrix of the two-step procedure with SVD

	TRUE															
FITTED	AT	BE	CH	CN	DE	DK	ES	FR	GB	IT	JP	LU	NL	OT	SE	US
AT	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
BE	0%	4%	2%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	2%
CH	0%	0%	49%	0%	2%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%
CN	0%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
DE	97%	0%	1%	0%	91%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
DK	0%	0%	0%	0%	0%	95%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%
ES	0%	0%	0%	0%	0%	0%	67%	0%	0%	0%	0%	0%	0%	2%	0%	0%
FR	3%	89%	33%	7%	6%	5%	4%	92%	9%	0%	13%	3%	1%	31%	27%	45%
GB	0%	0%	2%	7%	0%	0%	0%	0%	89%	0%	3%	0%	0%	18%	0%	0%
IT	0%	4%	1%	0%	0%	0%	1%	3%	2%	100%	0%	1%	0%	0%	0%	4%
JP	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	42%	0%	0%	1%	0%	0%
LU	0%	0%	11%	0%	0%	0%	28%	4%	0%	0%	0%	94%	1%	2%	0%	0%
NL	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	98%	0%	0%	0%
OT	0%	0%	1%	86%	0%	0%	0%	0%	0%	0%	42%	0%	0%	41%	0%	10%
SE	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	73%	0%
US	0%	0%	1%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	2%	0%	39%

6. Robustness Checks

In this section, the robustness of the results is tested with respect to both the sample composition and the choice of the classification algorithm. In particular, a completely balanced set is employed in which the different countries have the same number of units. Moreover, the outcome of the model is compared to an alternative, totally non-parametric, machine learning algorithm, i.e. random forests.

6.1 Balanced Set

Keeping fixed the total number of firms (around 180,000 units), a new sample is constructed in which all the sixteen countries represented (including the residual "Others - OT" category) have the same number of companies. Then, the algorithm is trained using the 80% of this new sample and is tested on the remaining 20%.

The new confusion matrix evaluated on the test set is reported in Table 5. It is possible to notice that the percentages of correct classification increase with respect to the previous unbalanced sample for the countries with low levels of FDI in the Italian companies, like Sweden, China, Spain, United States, which were previously underrepresented. However, the model performance is lower for countries with high levels of FDI. In particular, in some important cases, the choice between a balanced or unbalanced sample implies a clear trade-off: the increase in the accuracy of classification, obtained with a balanced sample, for Austria and Belgium (countries with low levels of FDI) is translated into a decrease of precision for Germany and France (countries with high levels of FDI), respectively.

Table 5: Confusion matrix when all countries have the same frequencies

	TRUE															
FITTED	AT	BE	CH	CN	DE	DK	ES	FR	GB	IT	JP	LU	NL	OT	SE	US
AT	50%	0%	2%	0%	22%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
BE	1%	75%	19%	3%	2%	6%	1%	62%	7%	1%	5%	2%	6%	20%	0%	25%
CH	2%	5%	39%	1%	3%	2%	2%	6%	1%	0%	1%	3%	0%	4%	0%	4%
CN	0%	0%	1%	86%	0%	0%	0%	0%	1%	0%	24%	0%	0%	20%	0%	1%
DE	45%	0%	5%	0%	72%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
DK	1%	5%	2%	0%	0%	83%	0%	3%	1%	0%	0%	1%	0%	6%	0%	4%
ES	0%	2%	4%	0%	0%	1%	77%	4%	0%	0%	0%	5%	0%	6%	1%	1%
FR	0%	5%	2%	0%	0%	1%	0%	11%	0%	0%	1%	1%	0%	2%	0%	2%
GB	0%	0%	3%	1%	0%	0%	0%	0%	81%	0%	1%	0%	0%	10%	0%	0%
IT	0%	0%	1%	0%	0%	0%	2%	0%	0%	97%	0%	0%	0%	0%	0%	0%
JP	0%	0%	1%	2%	0%	1%	0%	0%	0%	0%	57%	0%	0%	2%	0%	1%
LU	0%	2%	15%	0%	0%	0%	15%	7%	0%	0%	1%	87%	0%	4%	0%	2%
NL	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	93%	0%	0%	0%
OT	0%	0%	3%	4%	0%	0%	0%	0%	7%	0%	5%	0%	0%	23%	0%	1%
SE	0%	3%	2%	0%	0%	5%	2%	4%	0%	0%	1%	0%	0%	1%	98%	0%
US	0%	4%	3%	2%	0%	0%	1%	2%	1%	0%	5%	0%	0%	3%	0%	59%

Since the loss of accuracy involves countries with high levels of direct investments in the Italian enterprises, an unbalanced set should be preferred for the training of the algorithm.

6.2 Random Forests

In order to check whether the analytical assumptions underlying our approach are too restrictive to fit the data adequately, the binary and multinomial logit models are replaced by a totally non-parametric classification algorithm, i.e. the random forests, using as input variables the extracted components of the SVD procedure. The exercise is carried out using the unbalanced sample.

The confusion matrix evaluated on the validation set (20% of the sample) is reported in Table 6. It is possible to observe that the results obtained with this approach are similar to those in the two-step algorithm for most of the countries with high levels of FDI in the Italian enterprises, such as Great Britain, Luxembourg, Netherlands; moreover, random forests are also able to discriminate accurately between Italian and foreign firms and show better results for countries with low levels of FDI, such as Switzerland, Spain, Japan, and Sweden. However, the percentages of correct classification are much lower than our proposed two-step approach for two important countries with high levels of direct investments, namely France and Germany.

All in all, we can conclude that the classification performances of the proposed two-step algorithm are at least competitive with respect to other machine learning methods.

Table 6: Confusion matrix of the random forests approach with SVD

	TRUE															
FITTED	AT	BE	CH	CN	DE	DK	ES	FR	GB	IT	JP	LU	NL	OT	SE	US
AT	38%	0%	1%	0%	16%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%
BE	0%	13%	0%	0%	0%	0%	0%	10%	1%	0%	0%	0%	0%	0%	0%	8%
CH	0%	25%	85%	0%	1%	0%	2%	2%	1%	0%	0%	2%	0%	0%	0%	0%
CN	0%	0%	0%	33%	0%	0%	0%	0%	1%	0%	0%	0%	0%	11%	0%	2%
DE	63%	0%	4%	0%	76%	0%	0%	2%	0%	0%	0%	0%	0%	0%	0%	2%
DK	0%	0%	0%	0%	0%	70%	0%	1%	1%	0%	0%	0%	0%	0%	0%	2%
ES	0%	0%	0%	0%	0%	0%	95%	0%	0%	0%	0%	4%	0%	0%	0%	0%
FR	0%	50%	1%	0%	0%	3%	0%	59%	1%	0%	0%	2%	0%	0%	0%	8%
GB	0%	0%	0%	0%	0%	0%	0%	2%	86%	0%	0%	0%	0%	4%	0%	5%
IT	0%	0%	2%	0%	7%	0%	0%	11%	0%	100%	0%	1%	1%	0%	0%	3%
JP	0%	0%	0%	11%	0%	0%	0%	0%	1%	0%	95%	0%	0%	9%	0%	0%
LU	0%	0%	6%	0%	0%	3%	0%	2%	0%	0%	0%	90%	1%	0%	0%	0%
NL	0%	6%	0%	0%	0%	3%	0%	1%	0%	0%	0%	0%	99%	0%	0%	3%
OT	0%	0%	1%	44%	0%	20%	2%	5%	8%	0%	5%	1%	0%	77%	0%	13%
SE	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	2%
US	0%	6%	0%	11%	0%	0%	0%	3%	2%	0%	0%	0%	0%	0%	0%	51%

7. Conclusions

This paper describes a procedure to impute the nationality of foreign shareholders in Italian firms when the only relevant information is represented by the name of the corporations. The interpretation of the model and its predictions is very intuitive.

Despite this simplicity, the analysis has shown that the overall accuracy of the algorithm is very high (around 98%), with an almost perfect discrimination between Italian and foreign firms. Moreover, the proposed approach seems to be able to classify correctly most of the countries with high levels of direct investments in Italy.

The outcome of the classification can be further improved by combining this two-step approach with a popular method for dimension reduction used in text mining applications, i.e. the Singular Value Decomposition, that allows to extract in a more efficient way the information contained in the denominations of firms with a cost related to less interpretable input variables.

Furthermore, it has been proved that it is better to train the algorithm on an unbalanced sample based on the known priors on the distribution of the nationalities in the population than a balanced one, in order to classify with a higher accuracy the countries with the highest levels of FDI in the Italian enterprises.

The analysis has also shown that the results of the proposed procedure are competitive with respect to other, non-parametric, machine learning algorithms, like the random forests.

The adoption of the two-step algorithm presented in this work would improve significantly the procedure currently used by the Bank of Italy, that only discriminates between Italian and foreign firms, with a percentage of correct classification of about 80%. Indeed, the new model presents an overall accuracy equal to 98% and allows to impute the country of residency of the shareholders.

More in detail, the procedure described in this paper can enhance the quality of the Bank of Italy's FDI statistics in two directions: (a) ex-ante, since the imputed nationality of firms in the *Infocamere* database is a relevant stratification variable for the sampling scheme of the direct surveys; (b) ex-post, since the model can be used in the FDI grossing-up procedure to improve the allocation of the direct investments to the different counterpart countries, even in a probabilistic or "fuzzy" way.

References

Chakraborty, C., Joseph, A. (2017). Machine Learning at Central Banks. *Bank of England Staff Working Paper*, 674.

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 215-242.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.

Friedman, J., Hastie, T., Tibshirani, R. (2001). *The Elements of Statistical Learning* (Vol. 1, pp. 337-387). New York: Springer Series in Statistics.

IMF (2009), *Balance of Payments and International Investment Position Manual*, Sixth Edition (BPM6), Washington, D.C.: IMF.
https://www.imf.org/external/pubs/ft/bop/2007/pdf/bpm6.pdf

McFadden, D. L. (1974). Conditional Logit Analysis of Qualitative Choice Behavior, in P. Zaremba, ed., *Frontiers in Econometrics*, Academic Press, New York, 105-142.

OECD (2008). *Benchmark Definition of Foreign Direct Investment*, vol. 4. Paris: OECD.
https://www.oecd.org/daf/inv/investmentstatisticsandanalysis/40193734.pdf

Ripley, B., Venables, W. (2016). Package 'nnet'. R package version 7.3-12.

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC - Central Bank of Armenia Workshop on *"External Sector Statistics"*

Dilijan, Armenia, 11-12 June 2018

Imputation techniques for the nationality of foreign shareholders in Italian firms[1]

Andrea Carboni and Alessandro Moro,

Bank of Italy

[1] This presentation was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

BANCA D'ITALIA
EUROSISTEMA

Imputation Techniques for the Nationality of Foreign Shareholders in Italian Firms

Andrea Carboni, Alessandro Moro

Bank of Italy, Statistical Data Collection and Processing Directorate, External Statistics Division

IFC – Central Bank of Armenia workshop on "External sector statistics"
Dilijan, 11-12 June 2018

Outline

- ✓ Motivations
- ✓ Current Procedure
- ✓ A New Algorithm:
  1. Methodology
  2. Sample Selection
  3. Results
  4. Robustness Checks
- ✓ Conclusions

Motivations (1)

- In order to estimate the **Foreign Direct Investments (FDI)** item of the Italian Balance of Payments and International Investment Position, the Bank of Italy realises a direct sample survey for the non-financial and insurance companies.

- A stratified sample is used, considering among the other stratification variables the presence or absence for the firm of FDI relationships (inward and outward).

- Information on FDI inward is available in administrative data: annually, Italian enterprises report to the Chambers of Commerce the list of theirs shareholders (the so-called "*Elenco Soci*" in the *Infocamere* database).

Motivations (2)

- While this information is used by the Bank of Italy to identify the list of enterprises with FDI inward, quite often **the nationality of the shareholders is missing**.

- This piece of information would:

  1. help us in improving the stratification (and the efficiency) of the sampling scheme of our survey.

  2. allow to correctly attribute the FDI investments to the different countries.

- Hence, the present paper proposes a **Machine Learning Algorithm** of imputation when the nationality of foreign firms is unknown and the only relevant information is represented by the name of the corporations.

Current Procedure

- The current procedure discriminate between Italian and foreign firms as follows:
  - A **«dictionary»** has been constructed considering all the words contained at least 20 times in the names of a sample of 15,000 foreign firms (extracted from an external source: *Cerved* database).
  - If the denomination of the shareholder in our database contains at least one of the words in the dictionary, it is classified as a foreign investor; otherwise, it is considered as Italian.
- The percentage of correct classification is around 80%.
- The outcome of the actual procedure is binary (Italian vs foreign firms).

A New Algorithm (1)

- We propose a Machine Learning Algorithm for the identification of the nationality of shareholders based only on the **name of enterprises**.

- Our problem can be formalised by considering a set of *N* firms, each of them characterised by the couple *(Name, Country)*.

- The **final objective** is the identification of a predictive model relating the country to the name of the firm:

$$Country = f(Name)$$

- The outline of the procedure is:

  a) Preliminary data cleaning step: punctuation and special characters are removed.

  b) Decomposition of the name of each firm in its elementary words.

  c) The frequencies of the different words are evaluated and only the most frequent *K=50* words are selected.

A New Algorithm (2)

d) For each firm *i* and selected word *j*, a dummy variable is constructed:

$D_{i,j} = 1$ if the name of the *i*-th firm includes the *j*-th word;

$D_{i,j} = 0$, elsewhere

Table 1: Examples of dummy variables

Name	SRL	Societa	SPA	SA	Ltd	GMBH	PTY
Trelpa SA	0	0	0	1	0	0	0
Sud Chemie Australia PTY Ltd	0	0	0	0	1	0	1
Tarigia SRL	1	0	0	0	0	0	0
NGM Verwaltungs GMBH	0	0	0	0	0	1	0

These dummy variables constitute the regressors of all the subsequent statistical models

A New Algorithm (3)

- Since in our database the Italian firms are more than 90%, the procedure is articulated in **two steps**.

- **First step**. A logit model is estimated with the aim of identifying the Italian firms. The probability $\pi_{i,IT}$ that the nationality of the *i*-th firm is Italian is given by:

$$\pi_{i,IT} = \frac{\exp(\beta'_{IT} D_i)}{1 + \exp(\beta'_{IT} D_i)}$$

- If $\pi_{i,IT} > 0.5$, the *i*-th firm is classified as Italian. The selection of the most frequent *K* words is repeated on the observations classified as foreign in the first step.

- **Second step**. A multinomial logit model is estimated, in which the probability that the *i*-th firm belongs to the *h*-th nationality is given by:

$$\pi_{i,h} = \frac{\exp(\beta'_h D_i)}{\sum_{h=1}^{H} \exp(\beta'_h D_i)}$$

- The predicted nationality for the *i*-th corporation is the one to whom is associated the maximum probability.

Sample Selection

- The algorithm is trained using the **Bureau Van Dijk's Orbis database**, which contains, beyond many balance information, the name and nationality of more than 300 millions world enterprises.

- We have extracted from the Orbis database around 180,000 enterprises with a country-specific probability of inclusion equal to **our known priors**:

  - In particular, we know that more than 90% of the shareholders in the *Infocamere* database are Italians.

  - The remaining 10% of the sample has been selected according to the frequencies derived from our past samples (Table 2).

Table 2: FDI Investments in Italy by Country

Country	FDI Investments (Euros)	Percentage (%)
LU	19.701.471.920	21,4
NL	38.562.125.473	17,7
FR	23.375.783.193	11,9
DE	5.921.531.842	10,4
GB	24.979.578.528	9,8
CH	4.838.944.447	6,1
ES	3.362.143.938	3,5
US	3.171.216.814	3,1
BE	5.592.011.842	2,6
AT	915.696.442	2,4
JP	929.862.399	1,6
DK	682.030.864	1,5
SE	774.057.108	1,2
CN	48.917.610	0,7
Others	4.248.537.125	0,1

Example of Word Selection

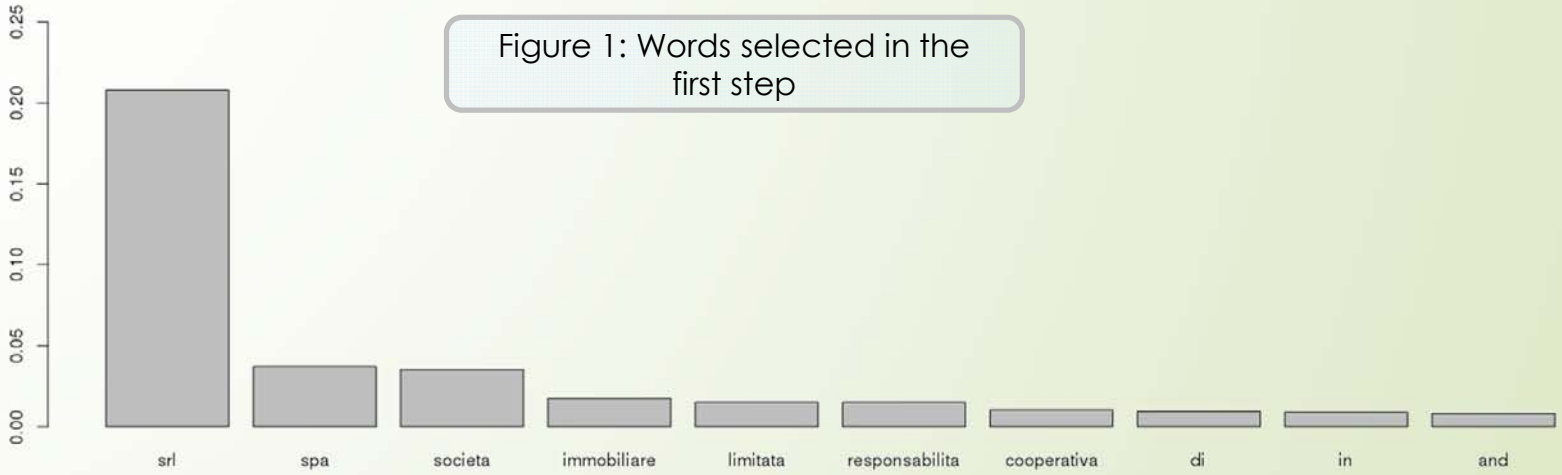


Figure 1: Words selected in the first step

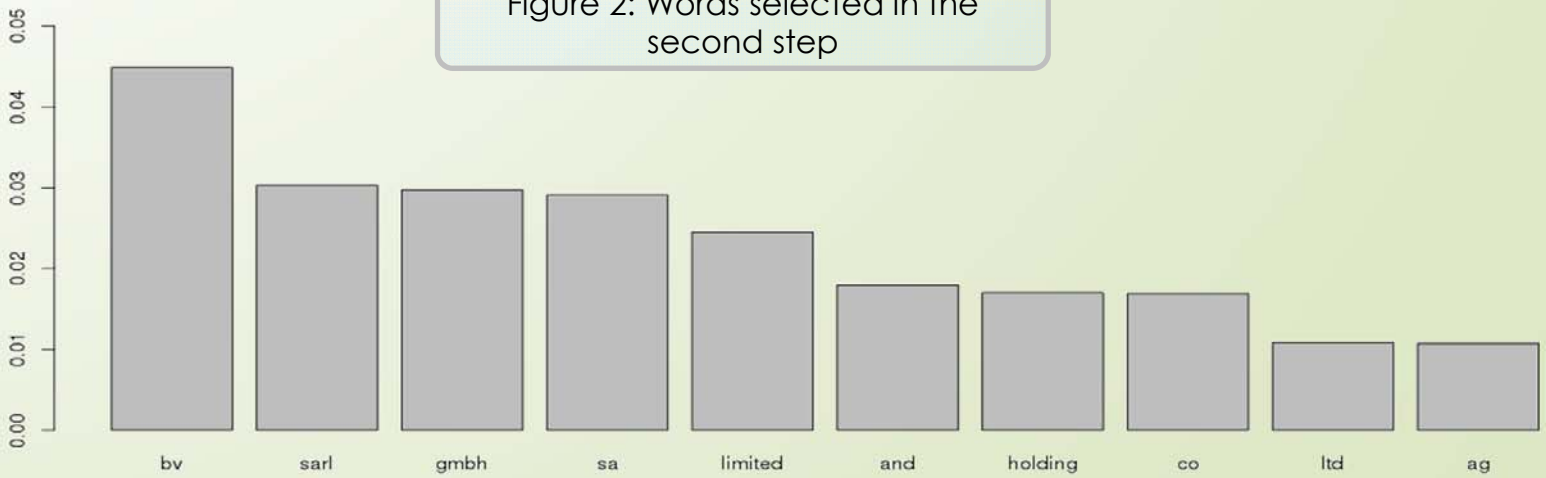


Figure 2: Words selected in the second step

Example of Model Prediction

- First and second step fitted probabilities in a real case: the German firm NGM Verwaltungs GMBH.

Figure 3: First Step Probabilities

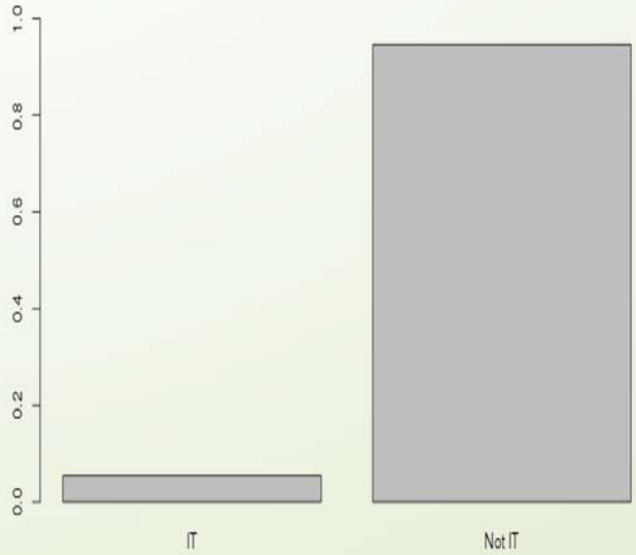
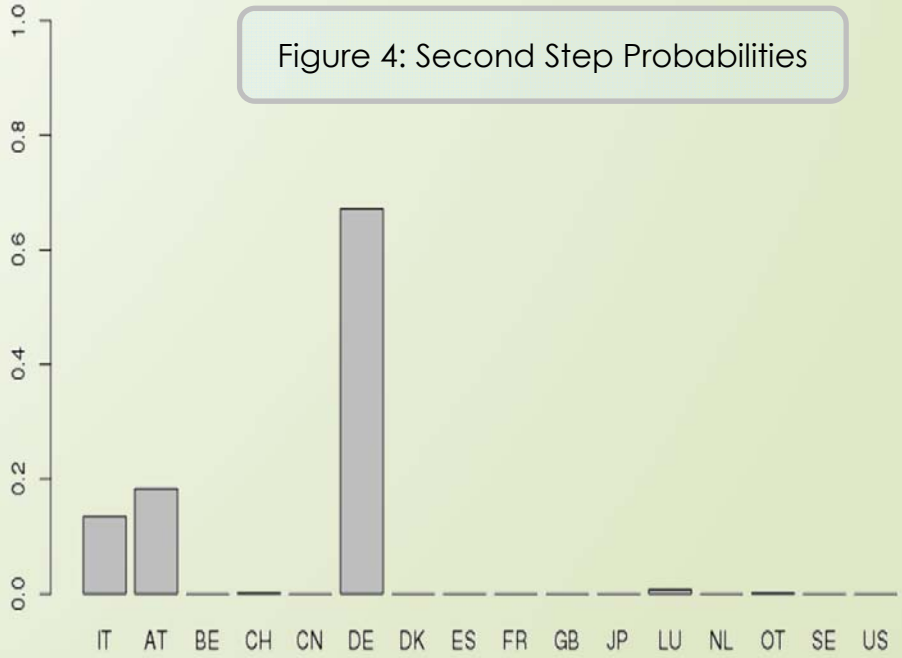


Figure 4: Second Step Probabilities



Results

- The model is estimated on the 80% of the sample and it is validated with the remaining 20%.
- The overall accuracy on the validation set is 98.29%.
- The confusion matrix in the validation set is:

	TRUE															
FITTED	AT	BE	CH	CN	DE	DK	ES	FR	GB	IT	JP	LU	NL	OT	SE	US
AT	10%	0%	0%	0%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
BE	0%	7%	2%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	2%
CH	0%	0%	41%	0%	3%	0%	1%	6%	2%	0%	0%	2%	0%	2%	27%	0%
CN	0%	2%	0%	14%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%
DE	81%	0%	1%	0%	90%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
DK	0%	0%	2%	0%	0%	82%	0%	1%	1%	0%	0%	1%	0%	1%	0%	0%
ES	0%	0%	0%	0%	0%	0%	59%	1%	0%	0%	0%	1%	0%	3%	0%	0%
FR	0%	76%	24%	7%	2%	9%	4%	66%	5%	0%	3%	1%	2%	21%	0%	45%
GB	0%	0%	1%	7%	0%	0%	0%	0%	80%	0%	6%	0%	0%	18%	0%	0%
IT	10%	9%	16%	7%	2%	9%	10%	11%	5%	100%	13%	6%	1%	10%	8%	6%
JP	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	45%	0%	0%	0%	0%	0%
LU	0%	2%	10%	0%	0%	0%	25%	9%	0%	0%	0%	88%	0%	5%	0%	0%
NL	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	96%	0%	4%	2%
OT	0%	0%	2%	64%	0%	0%	0%	0%	7%	0%	29%	0%	0%	34%	0%	6%
SE	0%	2%	2%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	62%	0%
US	0%	2%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	5%	0%	39%

Table 3: Confusion matrix of the proposed algorithm

Robustness Checks (1)

- A cross-validation exercise has been performed (80% training, 20% validation): both the parameter estimates and the percentages of correct classification are constant over the different samples.

- The procedure has been applied to a new sample in which all countries have the same frequencies. The new confusion matrix is:

	TRUE															
FITTED	AT	BE	CH	CN	DE	DK	ES	FR	GB	IT	JP	LU	NL	OT	SE	US
AT	50%	0%	2%	0%	22%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
BE	1%	75%	19%	3%	2%	6%	1%	62%	7%	1%	5%	2%	6%	20%	0%	25%
CH	2%	5%	39%	1%	3%	2%	2%	6%	1%	0%	1%	3%	0%	4%	0%	4%
CN	0%	0%	1%	86%	0%	0%	0%	0%	1%	0%	24%	0%	0%	20%	0%	1%
DE	45%	0%	5%	0%	72%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
DK	1%	5%	2%	0%	0%	83%	0%	3%	1%	0%	0%	1%	0%	6%	0%	4%
ES	0%	2%	4%	0%	0%	1%	77%	4%	0%	0%	0%	5%	0%	6%	1%	1%
FR	0%	5%	2%	0%	0%	1%	0%	11%	0%	0%	1%	1%	0%	2%	0%	2%
GB	0%	0%	3%	1%	0%	0%	0%	0%	81%	0%	1%	0%	0%	10%	0%	0%
IT	0%	0%	1%	0%	0%	0%	2%	0%	0%	97%	0%	0%	0%	0%	0%	0%
JP	0%	0%	1%	2%	0%	1%	0%	0%	0%	0%	57%	0%	0%	2%	0%	1%
LU	0%	2%	15%	0%	0%	0%	15%	7%	0%	0%	1%	87%	0%	4%	0%	2%
NL	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	93%	0%	0%	0%
OT	0%	0%	3%	4%	0%	0%	0%	0%	7%	0%	5%	0%	0%	23%	0%	1%
SE	0%	3%	2%	0%	0%	5%	2%	4%	0%	0%	1%	0%	0%	1%	98%	0%
US	0%	4%	3%	2%	0%	0%	1%	2%	1%	0%	5%	0%	0%	3%	0%	59%

Table 4: Confusion matrix when all countries have the same frequencies

Robustness Checks (2)

- Our procedure has been compared to other text mining approaches (Singular Value Decomposition - SVD) and machine learning methods (random forests).
- The SVD is a sort of PCA applied to the matrix of dummies (considering all the words in the names of firms). The extracted components are then the inputs of a random forest algorithm.
- The confusion matrix of this approach (SVD + random forest) is:

	TRUE															
FITTED	AT	BE	CH	CN	DE	DK	ES	FR	GB	IT	JP	LU	NL	OT	SE	US
AT	2%	0%	1%	8%	2%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%
BE	0%	4%	1%	0%	0%	0%	0%	1%	1%	0%	0%	1%	1%	0%	0%	0%
CH	2%	0%	33%	0%	0%	4%	2%	1%	1%	0%	0%	0%	0%	0%	5%	4%
CN	0%	0%	1%	17%	0%	0%	0%	0%	1%	0%	0%	0%	0%	2%	0%	0%
DE	76%	0%	1%	0%	86%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%
DK	0%	0%	0%	0%	0%	58%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
ES	0%	0%	0%	0%	0%	0%	31%	0%	0%	0%	0%	0%	1%	1%	0%	0%
FR	0%	6%	7%	0%	1%	8%	8%	36%	1%	0%	0%	4%	4%	3%	0%	7%
GB	0%	0%	0%	17%	0%	0%	0%	0%	85%	0%	0%	0%	1%	20%	0%	0%
IT	12%	13%	4%	0%	4%	4%	8%	6%	1%	100%	4%	3%	4%	4%	0%	5%
JP	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	11%	0%	0%	1%	0%	2%
LU	0%	11%	7%	17%	1%	0%	5%	9%	2%	0%	7%	67%	8%	3%	0%	2%
NL	7%	64%	44%	0%	5%	27%	45%	47%	3%	0%	71%	23%	77%	29%	68%	25%
OT	0%	0%	0%	42%	0%	0%	0%	0%	4%	0%	4%	0%	1%	36%	5%	5%
SE	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	23%	0%
US	0%	2%	1%	0%	0%	0%	2%	0%	1%	0%	4%	0%	2%	2%	0%	49%

Table 5: Confusion matrix in the SVD + random forest procedure

Conclusions

- The overall accuracy of the model is very high with an almost perfect discrimination between Italian and foreign firms.
- The proposed approach seems to be able to classify correctly most of the countries with high levels of FDI investments in Italy.
- Estimates and predictions seem robust in different samples, with the same country composition (CV) and with a different one (same number of units per country).
- The results are competitive with respect to other approaches (SVD + random forest).
- The model can help us in two ways:
  - **Ex-ante**: the imputed nationality of firms is a relevant stratification variable for our sampling scheme;
  - **Ex-post**: the model can be used in the FDI grossing-up procedure to attribute the values of FDI investments to the different countries in a probabilistic or "fuzzy" way.