BANCA D'ITALIA
EUROSISTEMA

Imputation Techniques for the Nationality of Foreign Shareholders in Italian Firms

Workshop on "Big Data & Machine Learning Applications for Central Banks"
Rome, 21-22 October 2019

Andrea Carboni & Alessandro Moro
Bank of Italy, Statistical Analysis Directorate,
External Statistics Division

Outline

- ✓ Motivations
  - ✓ Old Procedure
  - ✓ The New Algorithm:
    1. Methodology
    2. The Sample Selection
    3. Results
    4. Robustness Checks and Improvements
  - ✓ Conclusions

Motivations (1)

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

- Bank of Italy collects information about FDI through a direct survey; 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).

FDI: the direct investor owns at least 10% of the voting power of the direct investment enterprise. (*OECD –FDI Benchmark Definition*)

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:
  - help us in improving the stratification (and the efficiency) of the sampling scheme of our survey;
  - allow to correctly attribute the FDI investments to the different countries;
  - Verify the correctness of the responds received by the enterprises selected in our survey (direct reporting).

- Hence, we propose 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.

Results of the ML Algorithm

- The results of the proposed procedure are:
  - **99.9%** of correct classification of Italian vs foreign firms (the old procedure ensured a correct classification equal to 80%).
  - Very good performance in the classification of countries with high levels of FDI in Italy (LU, NL, FR, DE, GB).
- The model has already been used:
  - for selecting the 2019-2020 direct reporting sample,
  - in the grossing-up procedure,
  - and for checking the questionnaire replies.

Old Procedure

- The previous procedure discriminated 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, i.e. Cerved database).
  - If the denomination of the shareholder contained at least one of the words in the dictionary, the shareholder was classified as a foreign investor; otherwise, it was considered as Italian.
  - The percentage of correct classification was around 80%.
  - The outcome of the procedure was binary (Italian vs foreign firms).

The New Algorithm (1)

- We propose a ML 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 $f(.)$ relating the probability of belonging to a given country $\pi_{Country}$ to the name of the firm:

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

- The outline of the procedure is:
  - Preliminary data cleaning step: punctuation and special characters are removed.
  - Decomposition of the name of each firm in its elementary words.
  - The frequencies of the different words are evaluated and only the most frequent *K* words are selected.

The New Algorithm (2)

- 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.

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 the classification model.
- **Improvements:** selection of dummies with LASSO or SVD applied to the matrix of dummies.

The 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:

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

- 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:

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

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

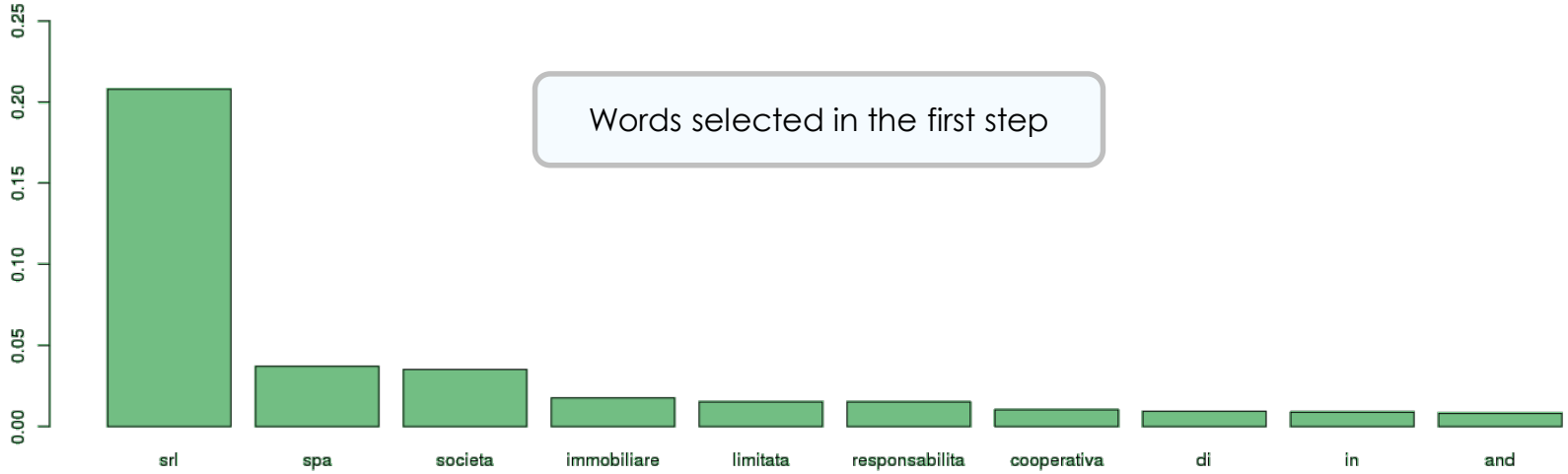
Sample Selection

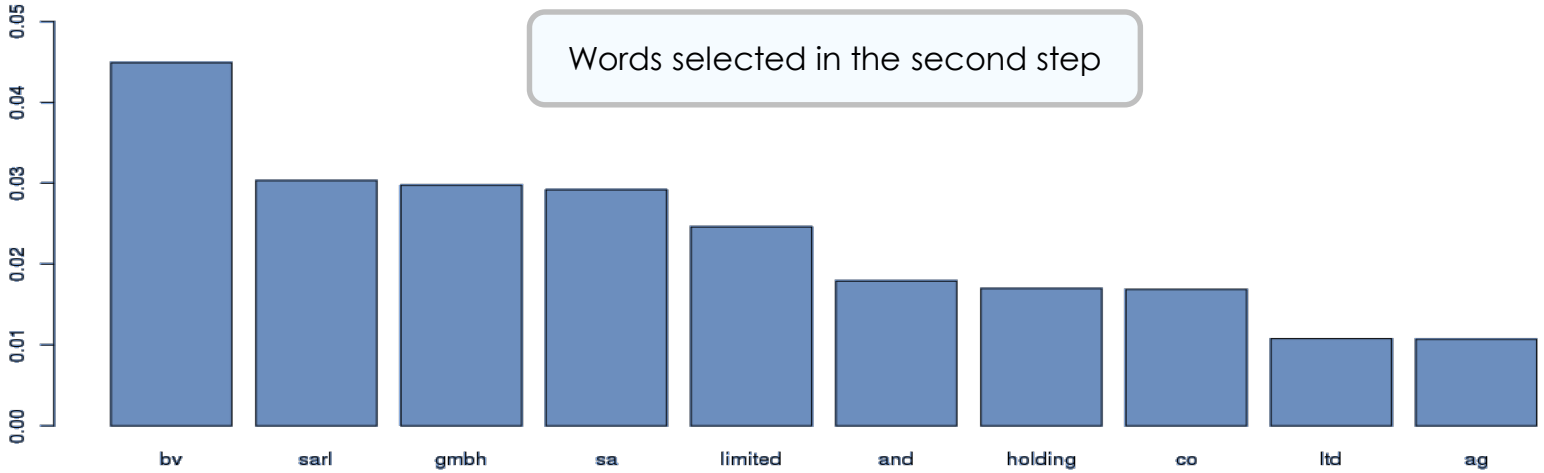
- The algorithm is trained using the **Bureau Van Dijk's Orbis database**, which contains, beyond many balance information, the name and nationality of more than 300 millions world enterprises.
- We have extracted from the Orbis database around 200,000 enterprises with a country-specific probability 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.

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 (OT)	4.248.537.125	0,1

Example of Word Selection

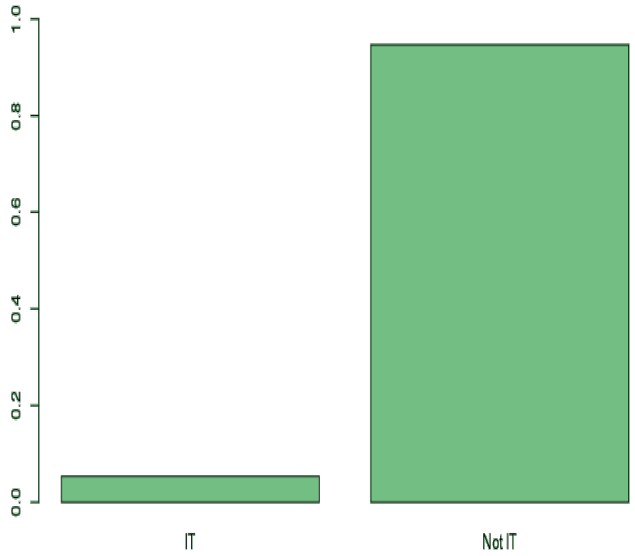




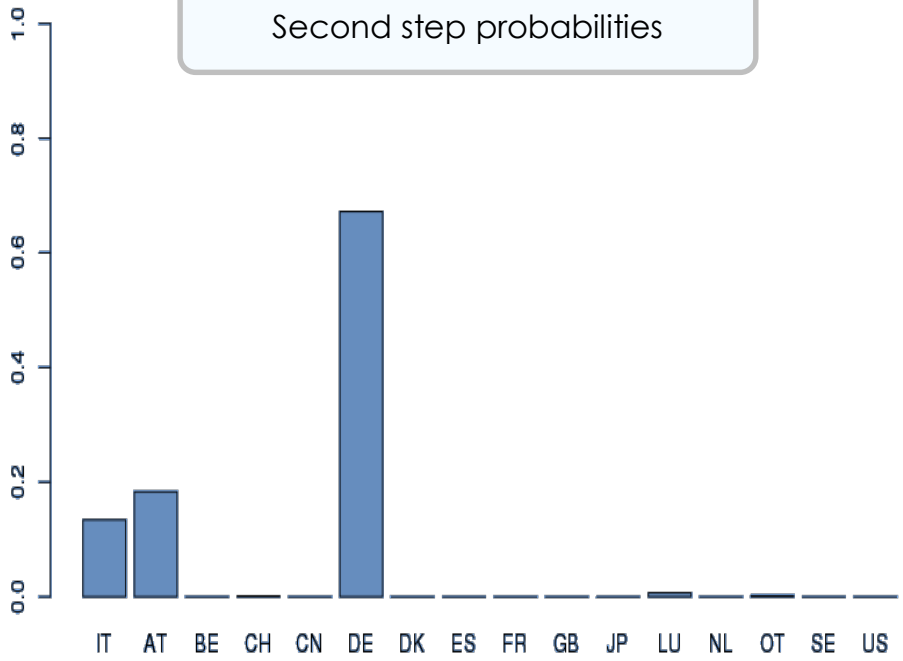
Example of Model Prediction

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

First step probabilities



Second step probabilities



Classification 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 98.3%.
- The confusion matrix in the validation set is:

Confusion matrix

	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%

Classification Results (SVD)

- Our procedure can easily be improved using the Singular Value Decomposition (SVD).
- The SVD is a sort of PCA applied to the matrix of dummies (considering all the words in the names of firms).
- The overall accuracy on the validation set is 98.8%.

Confusion matrix

	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%

Classification Results (LASSO)

- Our procedure can easily be improved using LASSO, to automatically select the most relevant variables, forcing the insignificant model coefficients to zero.
- This method can be easily incorporated in the proposed two-step algorithm by adding the LASSO penalty $(\sum_j |\beta_j|)$to the likelihood function associated to the logit (first step) and multinomial (second step) model.
- The overall accuracy on the validation set is 98.9%.

Confusion matrix

	TRUE															
FITTED	AT	BE	CH	CN	DE	DK	ES	FR	GB	IT	JP	LU	NL	OT	SE	US
AT	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
BE	0%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
CH	0%	7%	71%	0%	7%	0%	0%	6%	1%	0%	0%	1%	0%	1%	0%	2%
CN	0%	0%	0%	21%	0%	0%	0%	0%	0%	0%	3%	0%	0%	1%	0%	0%
DE	97%	0%	1%	0%	90%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%
DK	0%	0%	0%	0%	0%	64%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
ES	0%	0%	0%	0%	0%	0%	65%	0%	0%	0%	0%	0%	0%	5%	0%	0%
FR	0%	76%	10%	7%	0%	32%	4%	77%	2%	0%	3%	2%	3%	20%	0%	53%
GB	0%	0%	2%	7%	0%	0%	0%	0%	95%	0%	3%	0%	0%	7%	0%	0%
IT	0%	0%	2%	0%	3%	0%	4%	5%	0%	100%	3%	2%	1%	3%	0%	0%
JP	0%	2%	0%	7%	0%	0%	0%	0%	0%	0%	52%	0%	0%	0%	0%	0%
LU	0%	2%	12%	0%	0%	0%	26%	7%	1%	0%	0%	95%	0%	4%	0%	2%
NL	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	95%	0%	0%	0%
OT	0%	2%	2%	57%	0%	0%	0%	0%	1%	0%	35%	0%	0%	54%	0%	2%
SE	0%	4%	1%	0%	0%	5%	0%	4%	0%	0%	0%	0%	0%	0%	100%	0%
US	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	3%	0%	41%

Conclusions (1)

- 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.

- **The model has been used for the 2019-2020 direct reporting sample:**
  - **to stratify the enterprises before selecting the sample and in the grossing up procedure;**
  - **for checking purpose of the response of the direct reporters.**

Conclusions (2)

- Too many machine learning papers?
- about **100 ML papers per day** (looking to *Arxiv*, a popular public repository of research papers); 33.000 ML papers per year
- Not so much are produced in the Central Banks
- Very few application of big data or innovative techniques are used in the production of the official statistics

BANCA D'ITALIA
EUROSISTEMA

Imputation Techniques for the Nationality of Foreign Shareholders in Italian Firms

Thank you !!!

Andrea.carboni@bancaditalia.it

Alessandro.moro2@bancaditalia.it

Workshop on "Big Data & Machine Learning Applications for Central Banks"
Rome, 21-22 October 2019

Andrea Carboni & Alessandro Moro
Bank of Italy, Statistical Analysis Directorate,
External Statistics Division