

Estimation of Probability of Undeclared Employment in Italian Building Industry

Maria Felice Arezzo - Giorgio Alleva

Dept. of Studi geoeconomici, linguistici, statistici, storici per l'analisi regionale - Sapienza University of Rome
(e-mail: mariafelice.arezzi@uniroma1.it; giorgio.alleva@uniroma1.it)

Abstract: Vladimir Spidla, the EU's employment chief, in 2008 said Europe's black market economy "undermines the financing" of national social security programs and hinders efforts to boost economic growth. "There are no signs that the phenomenon is decreasing. Indeed in certain sectors and certain forms of work it appears to be growing," Spidla said. "We need to step up our approach and take more decisive action across the EU." This paper goes in the direction of shedding light on the phenomenon by using statistical models to detect which companies are more likely to hire irregular workers.

Keywords: Classification, CART, Logistic regression, Moonlighting, Administrative data base integration

1. Introduction

Moonlighting could be referred either as a multiple job worker or as a single job worker not regularly registered as an employ. In this paper ¹ we focus on the latter estimating the probability of moonlighting. The idea is that the probability of interest could be estimated from National Social Security Institute (INPS in the following) and Department of Labor inspections outputs. The outputs are expressed in terms of presence/absence of undeclared work. To understand which factors has a major impact on moonlighting, we relate the presence/absence variable to a wide variety of firm's characteristics (such as firm localization, labor productivity, labor cost, economic activity of the firm and so forth). The first researches on moonlighting in Italy go back at the early 70's. In one of the most important study, due to Censis in 1976, an estimate of about three million moonlighters was found. To explain the phenomenon Censis pointed out on one side at the off shoring of manufacturing firms and on the other side to the increased labor cost and to the improved labor conditions. In the 80's and 90's moonlighting has slowly but constantly increased until a turning point in 2002, probably due to Bossi-Fini's law, when it has decreased to the same level of 1992. The entity of moonlighting and its persistency in Italian economy leads us to think that it cannot be regarded as temporary and it highlights the importance of having tools able to track down firms with higher probability to hire irregular workers. Agriculture and Tertiary are traditionally the most exposed sectors (Istat estimated that in 2006 the percentage of moonlighters was 18,9% and 13,7% respectively). Within tertiary, commerce, transportation and tourism are far more affected than other sectors. In Industry

¹The research was realized thanks to ISTAT - ISFOL convention. We thank dott. Marco Centra e dott. Manlio Calzaroni that make it possible

moonlighting is marginal (3,7%) except for building and construction companies where the average irregularity percentage in 2006 was 11%. The paper is organized as follows. In section 2 we illustrate the main characteristics of the archives we had access to and we will explain which ones we used and how we merge them. In section 3 we'll show the main results we found using logistic and classification and regression trees models (CART in the following).

2. Dataset creation and variables used

To detect moonlighting, we need to pool firms through common behavioral patterns. In other words we have to find which characteristics construction companies which use irregular employee have in common. We need two kind of information: the first one is whether a firm has or hasn't moonlighters, the second are all its characteristics. It does not exist at the moment an informative system that gather all the information needed; therefore we had first of all to identify useful informative sources and then to combine them in a single dataset. The first kind of information are in INPS and Dpt of Labor inspections archives whereas the second ones are available in different archives. The following table illustrates the main characteristics of datasets we decided to use:

Table 1: *Main characteristics of all archives used*

<i>Owner</i>	<i>Content</i>	<i>Individual</i>	<i>Dimension</i>
INPS	Inspections outputs (2005 - 2006)	Inspection	31.658 inspections on 28.731 firms
Dpt of Labor	Inspections outputs (2005 - 2006)	Inspection	62.047 inspections
Revenue Agency	Sectoral studies (2005-2006). Models: TG69U, TG75U (SG75U), TG50U (SG50U and SG71U), TG70U	Firm	Universe of firms with at most 5 million euros of income
ISTAT	Asia Archives (2005-2006)	Firm	Universe of firms
ISTAT	Balance sheets of Stock and Ltd companies (2005-2006)	Firm	Universe of company/corporation

After a thorough data quality assessment, Dpt of Labor archives was temporarily discarded because information is available only for 16 regions out of 20 and because about 1/3 of VAT number are missing. Also balance sheets data were not used because they are not available for sole proprietorship business which have higher probability of using irregular workers. Data assessment showed that satisfactory territorial and sector coverage are guaranteed in the archives used i.e INPS, ASIA and Revenue Agency. Record linkage was done using VAT numbers and/or tax codes. The original variables were used to build economic indicators which can be grouped in the following 8 different firm's facets: a) 9 indicators for economic dimension, b) 13 for organization, c) 6 for structure, d) 6 for management, e) 11 for performance, f) 38 for labor productivity and profitability, g) 3 for contracts award mode, h) 7 variables for location and type. The final dataset had 97 inde-

pendent variables observed on 14.651 building and construction companies. The variable to be predicted is the last inspection result and it takes two possible values: presence or absence (in the tables are coded as Yes and No respectively) of irregular workers.

3. The Estimation of moonlighting probability

The aim is to associate to any firm a measure of irregularity risk, i.e. the probability of using undeclared workers. This probability can be estimated through statistical models that relate the inspection result (presence/absence of irregular workers) to firm's facets. We used both parametric (logistic) and nonparametric (CART) models. After the information system was built, we wanted three tasks to be achieved before fitting any model: 1) investigate the contribute of each firm's facet on moonlighting, 2) verify the predicting power of the information system 3) reduce the number of independent variables still maintaining as much predicting ability as possible. Goal one was achieved performing a principal component analysis on every single firm's facet considered and then building a tree using only the main factors extracted. Goal two was achieved building a tree with all 97 independent variables. The misclassification rate was used as a measure of the predictive power of the dataset. Variables importance (in terms of impurity reduction capability) and some logical prerequisites, were used to select a subset of independent variables and therefore to achieve goal three. On the selected variables subset, both logistic and CART models were fitted. On behalf of CART, the key points we observed on all trees grown are the following: a) territorial variables (region and province) are the most important to predict moonlighting; b) Legal structure is also a crucial variable; in particular sole proprietorships have a higher moonlighting probability; c) Labor cost systematically have an impact on moonlighting probability; d) Variables importance ranking appear to be stable over all trees grown. That means that main risk factors can be identified; e) It's much easier to classify correctly firms which don't use irregular workers rather than companies that do. To reduce misclassification rate on the latter, we had to increase prior probability on the presence of irregular work value. Out of more than fifty trees grown, we selected the one with a satisfactory balance between parsimoniousness (measured in terms of final nodes) and predictive ability (measures as correct classification rate on the test sample). Tables 2 and 3 show model characteristics and predictive power respectively.

Table 2: *Characteristics and dimension of final nodes (Training sample)*

<i>N</i>	<i>Final node number</i>	<i>Predicted value</i>	<i>Probability associated to predicted value</i>	<i>Number of firms in the final node</i>
1	5	No	0.900	54
2	7	Yes	0.333	521
3	8	No	0.795	110
4	4	Yes	0.592	326
5	2	No	0.833	3.366

As mentioned, for logistic regression we started from the same independent variables subset used for CART. Backward selection procedure was used to identify candidate models.

Table 3: *Model predictive ability: the classification matrix*

Sample	Observed	Predicted		Correct Percentage
		No	Yes	
Training	No	6927	1102	86,3%
	Yes	1382	863	38,4%
	Overall percentage	80.9%	19.1%	75.8%
Test	No	2936	484	85.8%
	Yes	594	363	37.9%
	Overall percentage	80.6%	19.4%	75.4%

The main results obtained are pretty much in agreement with those found using CART and could be summarized in the following points: a) territorial variables (region and province) are the most important to predict irregularities; b) companies organization variables and contracts awards mode play a key role in moonlighting prediction; c) it's much easier to classify correctly companies which don't have irregular employee. We had to reduce to 0.3 the cut off point. In other word we forced the model to classify as irregular all firms which had an estimated probability of using irregular workers greater than or equal to 0.3 d) all models have a predictive performance greater than 76%

Backward selection procedure gave six candidate models; we chose the one with the best predictive power and goodness of fit. Residuals analysis showed no critical elements for the chosen model. Scatterplot of predicted probabilities vs Cook's influential statistics showed some points that have to be investigated more in depth because they could be influential. Due to lack of space we cannot report neither parameters estimates nor residual graphs but only classification matrix.

Table 4: *Model predictive ability: the classification matrix*

Observed	Predicted		Correct percentage
	No	Yes	
No	7330	1359	84.40%
Yes	1171	1128	49.10%
Overall percentage			77.00%

References

- Ahmed E., Rosser J. B., Rosser M. V (2007) *Income Inequality, Corruption, and the Non-Observed Economy: A Global Perspective*, Springer.
- Breiman L., Friedman J., Olshen R. and Stone C. (1984) *Classification and Regression Trees*, Wadsworth.
- Hosmer D., Lemeshow S. (1989) *Applied Logistic Regression*, Wiley.