VAT Tax Gap prediction: a 2-steps Gradient Boosting approach

Keywords: *Tax Gap, Selection Bias, Gradient Boosting*

1 INTRODUCTION

Tax evasion represents one of the main problems in modern economies because it results in a loss of State revenue. The aim of this project is to provide an estimate of the Vat Tax Gap through a *bottom-up* approach based on compliance controls.

The aim of this work is to produce an estimate of the Italian VAT Tax Gap for the year 2011 via *machine learning* techniques. The observed data have been taken from two sources: the register of Irpef[1], VAT and Irap[2] declarations (available on all units, actual tax revenue due unknown) and the compliance control papers, performed only on a non-random sample of units (assessed units, actual tax revenue due known). One of the main problems of this analysis is related to the non-randomness of the compliance controls, that induce a *selection bias* on the observed sample. The final target of the analysis is to get trustful estimates for the undeclared tax base of the unassessed units. However, our model will focus on the estimation of the potential tax base (BIT) and the undeclared part will be derived as a difference: BIND = BIT − PAYED.
We propose a non-parametric 2-steps approach based on machine learning techniques in place of the standard methodology based on the Heckman model. The necessity of two steps is addressed by the necessity to handle the selection bias in order to provide more accurate estimates of the undeclared tax base. The advantages of this kind of approach are different. Machine learning techniques are (usually) *distribution free* and therefore are more flexible and able to adapt to a number of different contexts. Moreover, if adequate computational power is available, they can be succesfully applied to very large sets of data.

2 METHODS

The 2-steps approach consists of the subsequent application of two predictive algorithms. The first one is trained to the whole sample and targets the binary variable *assessed* and *not assessed*. It tries to find some regularity in the compliance control system, in order to produce an estimate of the probability to be selected for a tax assessment given the entire set of explicative variables

$$\hat{\pi}_i = P\ (\ i \in \mathcal{S} \mid \mathrm{X}\ ), \qquad i = 1, ..., n$$

These can then be used as weights in order to correct for the selection bias in the second step. Indeed, the second learner is trained only on the assessed units but each input observation has been weighted in proportion to the inverse of the probabilities obtained in the previous step $\nu_i \propto \frac{1}{\hat{\pi}_i}$. That is because units with high probability should already be over-represented in the sample, while ones with low probability are under-represented.

[1] Personal Income Tax
[2] Produced Activities Income Tax

The *Gradient Boosting* has been chosen among a set of algorithms for both steps because of its better performances in either tasks. This is a very powerful tool to get the costruction of predictive models for both regression and classification problems.
It has been used to directly produce point estimates and also to provide interval estimates by a *bootstrap* technique.
Point and interval estimates of the 2-steps Gradient Boosting have then been compared with the point estimates given by the Heckman model, highlighting some sort of (weak) concordance.

3 RESULTS

The following results are not referred to the data concerning the entire population of the Individual Firms in 2011. Indeed, the analysis has been carried out on a stratified sample of the unassessed units (2%) and all the controlled units (Table 1). The extracted sample consists of 64′207 tax-payers, whose 18′718 controlled.
The validation of the *first* Gradient Boosting has been performed via *out-of-sample validation*: the whole sample has been divided in a train-set (70% of the units, 44′900) and a test set (30% of the units, 19′307). The optimal tuning parameters have been chosen according to the AUC index. The best value obtained for the AUC is 0.79.
The same procedure has been adopted also for the validation of the *second* Gradient Boosting, which instead has involved only the 18′718 assessed units. Also in this case the sample has been split in a train-set (70% of the units, 13′098) and a test-set (30% of the units, 5′620). The optimal parameters have been chosen according to the R^2 index. The best value obtained for the R^2 is 0.83.
The predictions on the test set have been then compared to the ones produced by the standard Heckman model. While the aggregate estimate of the total BIND resulted very close to each other, it is possible to notice differences in term of individual estimation. In particular, the R^2 obtained by the estimates from the Heckman model is equal to 0.65, sensibly lower than the one achieved by our new approach ($R^2 = 0.83$).
Finally, the two models have been used to produce predictions for all the units whose actual BIT is unknown (unassessed units). These predictions allowed the computation of a synthetic measure of the propensity to Vat non-compliance, that we named *VAT evasion propensity*. This has been obtained as the ratio of the undeclared tax base (BIND) and the potential tax base (BIT). A low value of this ratio amounts to a compliant behaviour and viceversa.
The VAT evasion propensity for the entire sample of the taxpayers has been estimated to be of the 29.77% with Heckman's model and of the 30.40% for the 2-steps Gradient Boosting.
Propensities has been computed with both models seperately for different classes of individuals according to some observed variables.
Both models highlighted a greater propensity to evade VAT for female taxpayers with respect to the male ones (this difference is slightly smaller in the Heckman case).

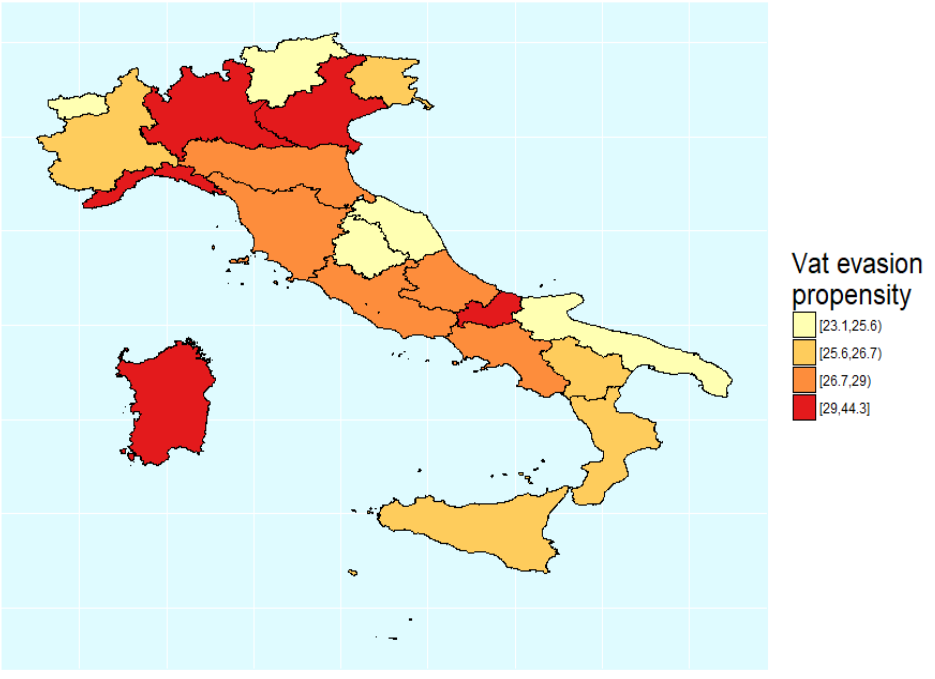
	Total Population		**Sample**	
Compliance Control	**Frequency**	**Percentage**	**Frequency**	**Percentage**
Unassessed	2′275′219	99.18%	45′489	70.85%
Assessed	18′718	0.82%	18′718	29.15%
	2′293′937	100%	64′207	100%

Table 1: Total and sampled population of individual firms.

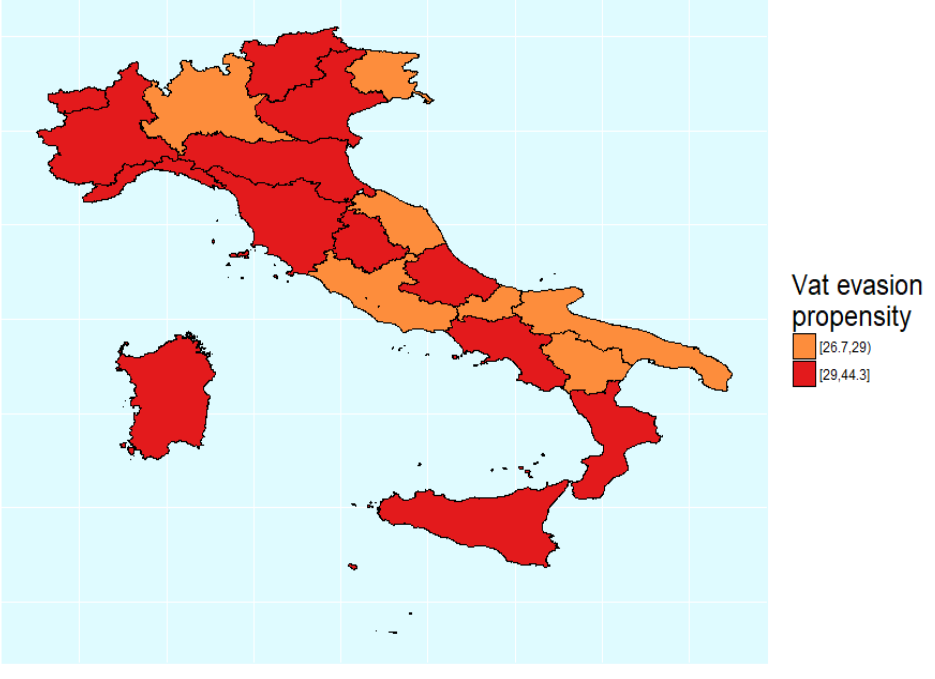
Analyzing this compliance index by age it is interesting to notice how the propensity decreases as the age increases.
Looking at the regional differences, we can notice some spatial variability (Figure 1). In particular, the propensities obtained by Gradient Boosting are more variable than the Heckman ones.

4 CONCLUSIONS

In conclusion, the machine learning approach is preferable because of its *distribution free-ness*. It is not necessary to apply any kind of transformations to the variables and it is not sensible to multicollinearity issues. Furthermore, machine learning based models usually provide good performances also in high dimensional settings, and allow to exploit all the information contained in large datasets. The total undeclared tax base estimate is similar for both tecniques. However the Gradient Boosting based model produced sensibly more accurate estimates of the single undeclared tax bases, cathching the individual variability associated to observed variables as it was desired. The Heckman model, on the contrary, flattens out individual differences.
The possible further development of this kind of approach are various and promising. For instance, the analysis exposed in this work has been performed only on a small subset of all the avalaible observations because of hardware limitations.
The results may be improved extending the analysis to the entire population with a more proper computational power. Moreover, this increased computation power would allow to apply more expensive and efficient techniques such as the *Extreme Gradient Boosting* and *Neural Networks*.



(a) *Gradient Boosting*



(b) *Heckman Model*

Figure 1: Regional italian map of tax evasion propensity.