A framework for the evaluation of preemptive tax administration policies

Pietro Battiston[1],
Simona Gamba[2],
Alessandro Santoro[3]

[1] Università di Parma
[2] Università Cattolica del Sacro Cuore
[3] Università di Milano-Bicocca

TARC, December 9th, 2020

From reactive to preventive policies

Traditional ex-post audits:

- are focused primarily on sanctioning past evasion;
- may have a deterrence effect but also tend to disrupt the relationship between TA & TP (disputes).

Need to complement them with preemptive or preventive tax administration policy: a set of interventions (audits and others) having the aim to prevent a non-fully compliant behaviour in the future. Examples:

- letters (Slemrod, 2001, and many others) and special audits to promote compliance;
- interventions to prevent late or missing debt payments (OECD, 2019).

Targeting quality

Preemptive policies are socially costly because

- are based on privacy-sensitive data and may reduce individual's freedom (Kerr and Earle, 2013);
- imply prediction errors of Type I (false positives) and Type II (false negatives).

Need to distinguish between

- *targeting quality*, i.e. ability to correctly identify taxpayers that, in the absence of the policy, would display the targeted behaviour;
- *intervention quality*, i.e. ability to prevent the targeted behaviour.

The evaluation problem

The distinction is impossible after the policy has been implemented. Denote:

- by P the occurrence of the behaviour and by NP its absence;
- by T the sets of targeted taxpayers and by NT the set of non-targeted TPs.

Then the subset TNP contains both:

- type I errors (false positives);
- good targets and good interventions.

. The problem is solved by the following design:

- observe behaviour until time $t-1$;
- choose an algorithm to predict behaviour at time t;
- evaluate the targeting quality.

After having done this, the policy can be implemented to prevent behaviour at time $t+1$.

Motivation of the paper

Providing tax administrations with a framework that *before* the policy is implemented allows:

- the evaluation of prediction errors ;
- the inclusion of the social costs of both types of prediction errors;
- the use of advanced data mining techniques.

In the paper we do two things:

1. define the framework;
2. apply it to Italy with the aim to predict which taxpayers are going to manipulate the data to reduce their probability to be audited in the future.

Framework/1

We assume that:

- every intervention entails:
  a. an average private cost for individuals (opportunity cost of using private resources, i.e. the time devoted to deal with the policy) equal to δ;
  b. an average administrative cost (e.g. wages paid to officers involved in planning and implementing the policy) equal to γ;
- every intervention which effectively prevents the targeted behaviour increases the social utility by β.

Framework/2

We define:

- λ as the shadow cost of raising a dollar of budget (i.e. the unitary cost of distortionary taxation);
- $b = \delta + (1 + \lambda)\gamma$ as the average social cost of every intervention;
- $a = \beta - b > 0$ as the social welfare generated by a completely error-free intervention;

so that the social welfare resulting from a given policy is:

$$W = \underline{W} + a \cdot P \cdot TPR - b \cdot (N - P) \cdot FPR$$

where $\underline{W}$ is the social welfare if no policy is adopted; the true positive rate (TPR) is the share of *positive* individuals correctly predicted as such (and the complement to 1 of the FNR, the false negative rate), while the false positive rate (FPR) denotes the share of *negative* individuals wrongly predicted as *positive*.

Framework/3

In the ideal, error-free, policy I, TPR=1 (there are no type-II errors) and FPR=0 (there are no type-I errors), leading to

$$W^I = \underline{W} + aP$$

so that we can evaluate any policy by the difference between W^I and its own W, i.e. by

$$Ł = a \cdot FNR \cdot P + b \cdot FPR \cdot (N - P)$$

i.e. the *loss* in welfare with respect to W^I, assuming 100% intervention efficiency.

Framework/4

Then we:

- use an analytical tool, known as ROC curve, which is common in big data analysis and is graphed in the $(TPR(.); FPR(.))$ space for different thresholds;
- reinterpret geometrically the minimized value of Ł as the point closest (tangency point) to $(1; 0)$, which represents the ideal policy;
- find a way to make this minimization operational for any predicton method, and its results comparable across different prediction methods.

The tangency condition

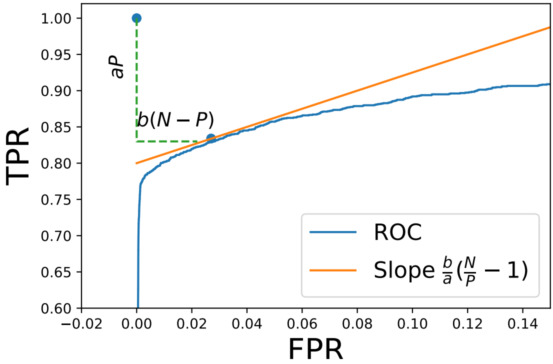

Figure: Minimum Ł as a tangency point

Italian SDS

Within SDS,

- small self-employed workers and sole proprietorships know the value of revenues (i.e. the total value of sales) the tax authority presumes they should report;
- this presumptive value is a weighted sum of input quantities reported by the taxpayers, where weights are given by input productivities estimated by the Revenue Agency on a population of similar and reliable taxpayers;
- they also know that there is a lower probability to be audited if the report a value of revenues not lower than the presumptive one.

Thus, there is a strong incentive to manipulate the data in order to reduce the probability to be audited. *Bunching* at the presumptive value is a strong indication of this data manipulation.

Data [1]

A perfectly balanced panel of 662 241 TPs (self-employeds or sole proprietorships) observed between 2007 and 2011, for a total of 3 311 205 observations. For each observation we know:

- demographic characteristics, such as age, gender, city and province of residence, and number of open VAT positions;
- detailed content of the tax reports, including main revenues, costs, tax bases, and the amount of tax due for three taxes – personal income tax (IRPEF), value-added local tax (IRAP), and VAT;
- presumptive and reported revenues.

[1] Provided by the Italian Revenue Agency on the basis of a Research Agreement with DEMS

Bunching evidence

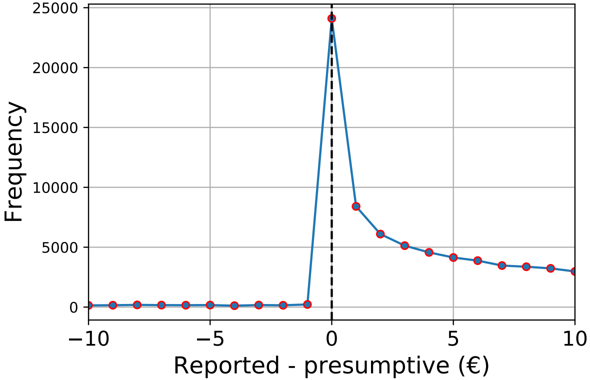

Figure: Exact bunchers, btwn 0.5 and 1% of the observations

Prediction models

We apply different prediction methods:

- OLS penalized models (Ridge and Lasso);
- decision trees;
- random forests;
- neural networks.

Each method generates a model when a given specification of its (hyper)parameters is selected. We ran 20 out-of-sample iterations of predictions for each combination of method, hyperparameters and year. In each iteration, we split the data randomly in two samples of equal size. In total, 6000 models were hence trained and tested.

Summary of main prediction results/1

We set $a = 1000$ and $b = 500$ and note that:

- non linear models perform better (i.e have much lower values of Ł) than linear ones;
- the best model (1st in 3 out of 4 years, and 2nd in the other) is a quite complex random forest with 18/20 levels;
- despite its complexity, this model can be interpreted using methods to evaluate the variable importance to profile bunchers. It emerges that
  - bunchers bunch repeatedly over time;
  - bunchers belong to specific sectors;
  - bunchers tend to underreport input and costs, because this allows them to reduce presumptive revenues.

Summary of main prediction results/2

- The best model has a *TPR* which ranges btwn 70 and 80% and a *FPR* btwn 0.1 and 0.3%;
- the best model has a threshold around 30% and it targets a share of taxpayers close to that of actual bunchers;
- the best model entails a loss of welfare which is remarkably lower than that generated by random intervention: the best model to predict 2011 bunching generates $Ł = 1,900k$ euros whilst that associated to a random intervention policy would be twice larger.

Concluding remarks

The intervention on predicted bunchers can consist of:

- letters to warn taxpayers that a monitoring effort on input data manipulation will be exerted;
- field/desk audits on the internal consistency of input data before they are reported.

Same framework can be used also to improve the targeting efficiency of conventional ex-post audits.