Italy: VAT gap estimation via bottom up approach

1. Context of analysis and objectives

In Italy, the Value Added Tax (VAT) represents the main source of revenue among indirect taxes, and provides the 25% of the total tax revenue. For this reason, fighting VAT evasion has been put at the top of the agenda by tax administration and other national institutions. In this context, there is the need to have reliable measures of VAT gap.

Most commonly, the VAT gap is estimated via "top down" methods. These methods generally combine aggregated data coming from different data sources (mainly National Account and tax administration data) provide an exhaustive picture of the evasion. Unfortunately, top down methodology can provide gap estimates at very high level of aggregation and do not allow identification of the main risky areas, in terms of different taxpayers characteristics such as sector of activity or economic dimensions.

In order to "disentangle" the VAT gap estimation according to different analysis "dimensions", Italian Revenue Agency has been recently implementing a novel methodology that, using data from tax audits performed by the tax Authority, combines traditional parametric inference methods, modern machine learning techniques and nearest neighbor imputation procedures. The final outcome of the proposed methodology, that could be defined "machine learning assisted", is a set of individual values of the target variable, that, at least in principle, can be used to obtain estimates of the VAT gap at whatever level of disaggregation.

Differently from many estimation methods proposed in literature, the proposed strategy does not adopt the classical Heckman approach to remove possible selection bias originated by *non-random* selection mechanism of assessed taxpayers. In fact, this would imply strong distributional assumptions on the model generating the data. Instead, the estimation is based on a two-stage procedure relying on the "conditional independence assumption".

The final goal is to build an integrated approach which combines advantages of top down and bottom up methodologies, providing estimates with high level of exhaustiveness and granularity at the same time.

These estimates can be used for different requests from national and international institutions, such as provision of ex-ante or ex-post assessment of fiscal effects of reforms, identification of risky areas according to the different economic and/or fiscal dimension of the phenomenon, evaluation of the main channels through which VAT evasion takes place, etc.

2. Choice of the target variable

In the bottom up approach, the estimates of interest on different domains are obtained by summing up predictions of the target variable Y at individual level. Typically, Y is supposed to be known on a subset S of size n of the target population U of size N. Correspondingly, a set of values $y_1, .. y_n$, viewed as n independent realizations of Y, can be used, together with a set of input variable X, to estimate the predictive model.

The first issue to deal with, in the context of VAT gap estimation, is the choice of the response variable Y. This is not obvious because Y should be defined in terms of quantities actually measurable on (at least) a subset of the target population. Ideally, in order to meet the general definition of tax gap, the difference between tax potentially collectible and tax actually collected should be known for each taxpayer of *S*. However, this quantity is not likely to be measured exactly by the real assessment processes performed by the tax administration.

To address this problem, the concept of tax gap at micro level is approximated by the "*additional VAT assessed*" (AVA) by the tax administration, for each taxpayer subject to an operational audit.

3. Overall estimation strategy and treatment of selection bias

The bottom up methodology for VAT gap estimation is essentially based on tax assessments carried out by the fiscal administration on audited taxpayers. Thus, if *n* taxpayers are checked, a set of corresponding values $y_1, .. y_n$ of the target variable is available to train models through supervised procedures.

From a theoretical point of view, different inferential frameworks are possible depending on the selection mechanism adopted to assess the taxpayers. In particular, in case of probabilistic sampling under some specific sample design, randomization-based inference is often the best choice, since it does not rely on the validity of distributional assumptions on the model supposed to generate data. Unfortunately, in many countries (among them Italy) random audits are not implemented, so that there are no sampling weights associated to audited taxpayers. Instead, selection of taxpayers is based on complex risk analysis criteria, generally unknown to the analyst. This implies the need to think of observed data as realizations from some underlying probability distribution. In particular, model specification should account for both the selection mechanism and the distribution of the target variable over the analyzed population.

Formally, for $i = 1, .., N$, let I_i be the selection indicator variable, taking value 1 if the *i*-th unit is selected (audited) and 0 otherwise. Then, what is to be modeled, conditionally on the available information, is the joint distribution of (I_i, Y_i). In practice, given sets of variables Z and X, possibly overlapping, this is done by modeling the probability distributions $P(I_i|Z_i)$ and $P(Y_i|I_i, X_i)$, corresponding, respectively, to the selection model and the "substantial" model. However, since the target variable Y_i is not observed if $I_i = 0$, i.e. if the *i*-th taxpayer is not selected, $P(Y_i|I_i = 0, X_i)$ cannot be estimated. Different approaches are adopted to overcome this problem. One of the most popular is the two stage Heckman methodology (Heckman, 1976), where the unconditional (with respect to the selection variable) distribution of the target variable is supposed to be Gaussian, as well the probability distribution of the latent variable governing the selection mechanism. The resulting model is a type II Tobit model (Amemiya, 1985), which implies that observed values of *Y* are distributed according to a censored Gaussian distribution.

An alternative approach is based on the assumption that, conditionally on a suitable set of covariates Z_i, the target variable and the selection variable are independent: $Y_i \perp I_i \,|Z_i$[1]. This assumption is generally referred to as Conditional Independence Assumption (CIA), and implies that, since $P(Y_i|I_i = 0, Z_i) = P(Y_i|I_i = 1, Z_i)$, the selection mechanism is *ignorable* so that valid inferences and predictive analyses can be performed based (only) on the observed data.

In the present context, it is assumed that conditional independence (and hence ignorability of the selection process) holds, although it is not testable by its very nature. This choice is motivated by the

[1] More precisely, it is assumed that the conditional independence property is still valid if the conditioning variables are extended to all covariates that are used to make statistical prediction.

need for having high degree of flexibility in modeling the data. In fact, in Heckman methodology, data are assumed to be normally distributed.

The conditional independence assumption discussed above motivates the two stage estimation procedure illustrated in the following paragraphs.

3.1 First stage: stratification based on selection probabilities

In the first phase, aimed at reducing the effects of selection bias, the target population *U*, as well as the set *S* of taxpayers subject to an assessment, is partitioned into classes of "approximately constant selection probability". Specifically, we use a logistic model to estimate, for each taxpayer $u_i(i = 1,..,N)$ of the target population, the probability $p_i \equiv p(z_i)$ of being assessed, given the set of covariates Z_i. Then, the population is stratified according to the quintiles of the distribution of the p_i's. It is assumed that, within each stratum Q_j $(j = 1,..,5)$, the variability of the p_i's is low enough to justify the approximation of "constant selection probability". This allows us to look at taxpayers in $Q_j \cap S$ as a simple random sample from the set of taxpayers in $Q_j \cap U$, so that within each stratum audited taxpayers can be considered as "representative" of all taxpayers.

In the second phase of the estimation procedure, within each subset of the target population defined *a priori* in terms of the structural characteristics of the taxpayers, separate predictive models are estimated for each class Q_j. In the following sections, details on prediction strategy within strata are provided.

3.2 Second stage: Prediction via Bagging of regression trees

The goal of the second stage is to build a statistical model for prediction of the target variable *Y*. To this aim, a nonparametric approach is adopted where, within each stratum defined as illustrated in the previous section, predictions of the AVA are obtained via bagging of regression trees (RT). Bagging is a general machine learning technique whose objective is to reduce variability of predictions by averaging on several replications of the predictive procedure. Specifically, in the present context, where the algorithm is trained on the set *S* of the assessed taxpayers, the method is composed of the following steps:

```
$for\ b = 1\ to\ B$
{
i) draw a bootstrap sample $S^b$ of size n from $S$[2]
ii) train a regression tree $T^b$ on each bootstrap sample $S^b$
iii) $for\ i = 1\ to\ N$, use the tree $T^b$ to obtain a prediction $\hat{y}_i^b$ of the target variable Y for the i-th unit
}
$for\ i = 1\ to\ N$ compute the bagging prediction $\hat{y}_i^{bag}$ by averaging on the B predictions $\hat{y}_i^b$, as in following equation:
```

[2] Bootstrapping is a *resampling* method based on simple random sampling with replacement from the original data sample. Introduced by Efron (1979) in the context of jacknife methodology, it is commonly used, in both parametric and nonparametric version, to estimate the variance in cases where analytic approach is not feasible.

$$1) \quad \hat{y}_i^{bag} = \frac{1}{B}\sum_{b=1}^{B} \hat{y}_i^b$$

Note that, as a result of the above algorithm, predictions are constant on the subsets of the input space of the form:

$$2) \quad \left\{P_{j^1}^1 \cap \cdot\cdot P_{j^b}^b \cap \cdot\cdot P_{j^B}^B\right\};\ j^b = 1,..,n^b,$$

where $P_{j^b}^b$ $(b = 1,..B)$ is the j^bth set of the partition of the input space corresponding to the tree trained on the b-*th* bootstrap sample, and $n^b$is the corresponding number of final nodes.

In other words, the sets in 2) are obtained by taking intersections of all the partitions corresponding to the *B* regression trees.

3.3 Predictive Mean Matching

Bagging of regression trees is a flexible technique to obtain "model free" predictions of the target variable. In fact, given the well-known optimality property (in terms of quadratic loss function) of the expected value, the prediction accuracy depends on the capability of the predictive technique to approximate the expectation of the target variable (conditional on the available information).

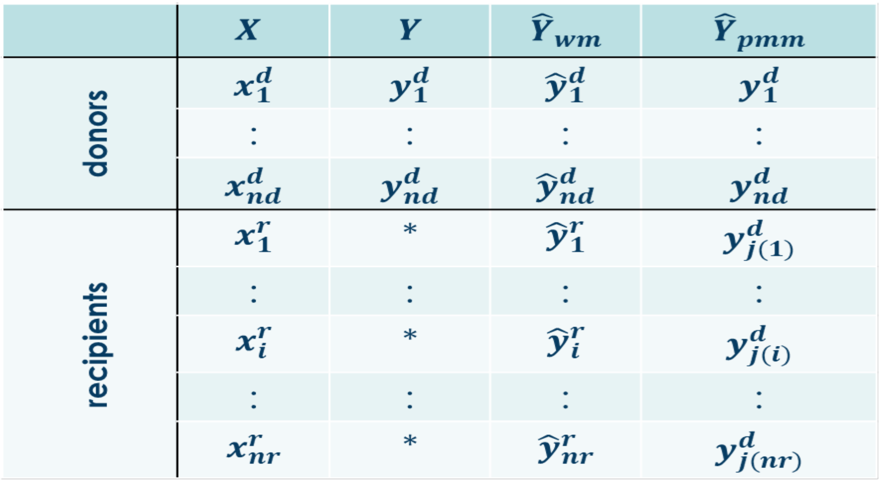
However, predictive accuracy is not the only parameter to consider when the objective is to build a *multipurpose estimation tool*. Specifically, in some circumstances, it can be useful to build a "synthetic" set of values that can be thought of as random draws from the distribution of the target variable *Y*, conditional on the set of available covariates, rather than expectations of that distribution. This is the case when one is interested in several distributional characteristics besides means, such as, for instance, higher order moments or quantiles. In the present application, where VAT evasion is the subject under analysis, a methodology able to guarantee preservation of distributional features could be preferable to one that, although having high predictive performances, does not allow valid inferences on non linear functions of data.

In non parametric setting, as in the present one, a possible option is to adopt a Nearest Neighbor Donor (NND) approach. Basically, for each record where *Y* is not observed (*recipient*), one searches for the record with known value of *Y* (*donor*), which is the *nearest* in terms of some predefined metrics in the input space. Then, the *Y* value of the donor is *imputed* to the recipient. Common choices for the metrics are Euclidean, Mahalanobis and Manhattan (or block distance). In case of univariate problems (only one input), if ignorability of the selection process can be assumed, it can be proved (Chen and Shao, 2000) that, under mild technical conditions, NND imputation is asymptotically unbiased, in the sense that it tends to be equivalent to random drawing from the target distribution, as the number of sample units becomes large. However, with finite samples, NND imputation can result in poor predictive performance if the dimension of the input space is large (see, for instance, Hastie et al. 2019).

An interesting method, which can be a possible alternative to a genuine NND approach, is the Predictive Mean Matching (PMM). PMM can be considered as a *model assisted* technique, in the sense that it uses a "working model" (for instance a simple linear regression model) to obtain rough predictions of the response variable *Y*, for both sampled and non sampled units. Then, distances with respect to these predictions are used to match each recipient r_i with the "nearest" donor $d_{j(i)}$ where $i = 1,..,n_r,\ j\ \in$

$(1, .., n_d)$, and n_r, n_d are the number of recipients and donors respectively. Finally, the observed value of *Y* observed on $d_{j(i)}$ is imputed to r_i (see Fig. 1).

Figure 1. Illustration of PMM.



	X	Y	$\hat{Y}_{wm}$	$\hat{Y}_{pmm}$
donors	x^d_1	y^d_1	$\hat{y}^d_1$	y^d_1
	:	:	:	:
	x^d_{nd}	y^d_{nd}	$\hat{y}^d_{nd}$	y^d_{nd}
recipients	x^r_1	*	$\hat{y}^r_1$	$y^d_{j(1)}$
	:	:	:	:
	x^r_i	*	$\hat{y}^r_i$	$y^d_{j(i)}$
	:	:	:	:
	x^r_{nr}	*	$\hat{y}^r_{nr}$	$y^d_{j(nr)}$

Note: $\hat{Y}_{wm}$ and $\hat{Y}_{pmm}$ represent, respectively, predictions from working model and PMM. The *i*th recipient is imputed with the value $y^d_{j(i)}$ such that $j(i) = \underset{j}{\text{argmin}}\, |\hat{y}^r_i - \hat{y}^d_j|$.

In general, PMM does not have the same asymptotic properties as a genuine NND technique based on a proper distance function in the input space. In fact, PMM performances rely to some extent on the choice of the working model. In particular, if the model fit is poor, predictive means may be very far from the real expected values.

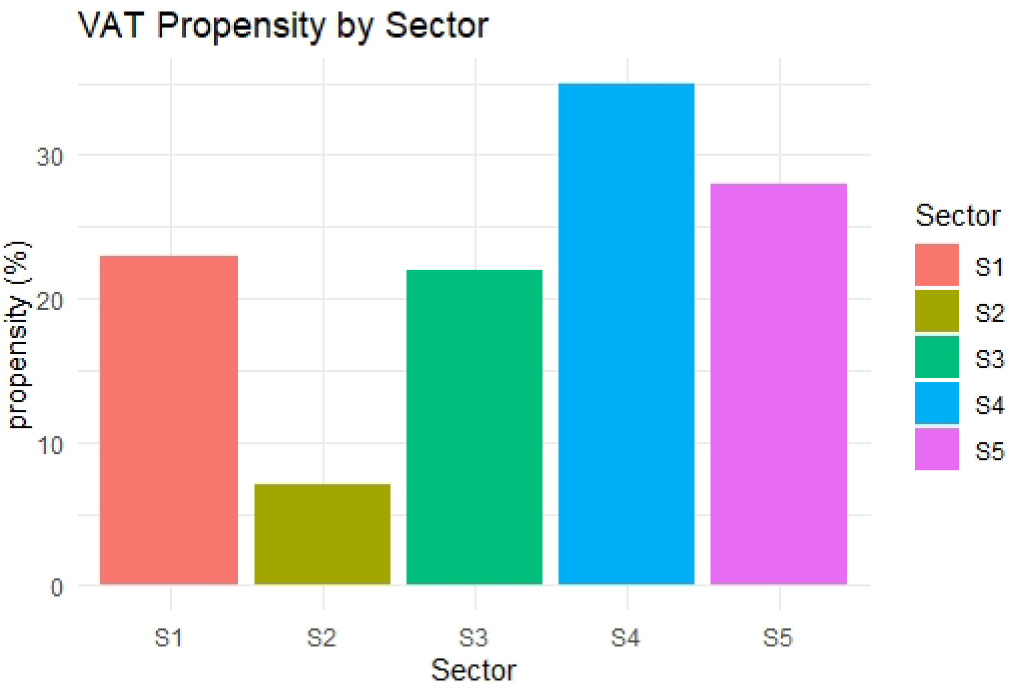
In the present application, predictions are obtained in a nonparametric manner by averaging outputs from many regression trees (bagging). As already noticed, this results in taking means of *Y* within subsets of the input space, obtained by overlapping the partitions corresponding to the different trees. Thus, in the present context, PMM is very much the same as random drawing from these subsets. In this sense, PMM can be viewed as a within strata random imputation, based on an efficient way of defining strata.

4. Final output

As noticed in the previous sections, the main result provided by the bottom up approach is the multidimensional picture of the VAT gap. The methodology, based on the information derived from the assessment activity of the fiscal administration, allows for the disaggregation of the phenomenon across different domains of interest (sectoral, provincial, type of taxpayer, different level of turnover, etc.), providing useful inputs for the risk analysis activities carried out by Tax Authorities. Moreover, since, ideally, the bottom up approach is able to produce estimate of the potential VAT gap at micro level, it represents a useful tool to bread down the VAT gap components defined in terms of taxpayer behavior.

An example of how results from bottom up methodology can be used for VAT gap estimation on specific domains, is provided below (Fig.2). In the figure evasion propensity is reported separately for 5 economic sectors (year 2018).

Figure 2. Vat evasion propensity in 5 macro economic sectors.



References

Amemiya, T. (1985). Advanced Econometrics. Harvard University Press: Cambridge, MA.

Chen J., Shao J. (2000) Nearest Neighbour Imputation for Survey Data, Journal of Official Statistics, 16, 113-131.

Heckman, James (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. Annals of Economic and Social Measurement, 5 (4): 475-492.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics. 7 (1): 1–26.

Hastie T., Thibshirani R., and Friedman J. (2009), The Elements of Statistical Learning. Springer Series in Statistics.

Little, Roderick J. A. (1988). Missing-Data Adjustments in Large Surveys. Journal of Business & Economic Statistics. 6 (3): 287–296.

Rosenbaum P.R., Rubin D.B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika. 70, 1, pp. 41-55.