XV

Volume XV, Number 2, November 2012

Journal of Applied Economics

Giovanni Cerulli
Bianca Potì

Evaluating the robustness of the effect of public subsidies on firms' R&D: an application to Italy

UCEMA

Edited by the Universidad del CEMA
Print ISSN 1514-0326
Online ISSN 1667-6726

Journal of Applied Economics. Vol XV, No. 2 (November 2012), 287-320

EVALUATING THE ROBUSTNESS OF THE EFFECT OF PUBLIC SUBSIDIES ON FIRMS' R&D: AN APPLICATION TO ITALY

Giovanni Cerulli and Bianca Potì*
CNR-Ceris (National Research Council of Italy, Institute for Economic Research on Firms and Growth)

Submitted December 2009; accepted October 2011

This paper applies different econometric methods to evaluate the effect of public subsidies on firms' R&D activity. For the sake of robustness, results from the Heckman selection model (Heckit), Control-function regression, Difference-in-differences, and various Matching methods are compared by using the third and fourth wave of the Italian Community Innovation Survey (CIS3, years 1998-2000 and CIS4, years 2002-2004). We predict the absence of a full crowding-out of private R&D performance, both for the whole sample and for some subsets of firms. Nevertheless, we conclude that while for variables expressed as ratio (R&D intensity and R&D per employee) the difference in results is negligible, R&D expenditure presents a strong variability among the approaches, even for those relying on similar identification assumptions. Given the utmost importance of this target-variable, future works should go beyond the use of single methods, especially when they are thought of to steer future policymaking.

JEL classification codes: O32, C52, O38
Key words: business R&D, public incentives, econometric evaluation, robustness

I. Introduction

The paper applies and compares some recent econometric methods used for evaluating the effect of public subsidies on firms' R&D performance. More specifically, results from the Control-function regression, Heckman selection model, Difference-in-differences and various Matching methods are proposed and compared, using as datasets the third and fourth wave of the Italian Community Innovation Survey

* Giovanni Cerulli (corresponding author): CNR-Ceris (National Research Council of Italy, Institute for Economic Research on Firms and Growth), Via dei Taurini 19, 00185 Roma, Italy; email: g.cerulli@ceris.cnr.it. Bianca Potì: CNR-Ceris (National Research Council of Italy, Institute for Economic Research on Firms and Growth), Via dei Taurini 19, 00185 Roma, Italy; email: b.poti@ceris.cnr.it.

(CIS), collecting data for supported and non-supported firms over 1998-2000 (CIS3) and 2002-2004 (CIS4).[1]

An important aspect emerging when looking at the wide literature on the subject is the common practise of providing average results from a single econometric method (using, sometimes, also one single target-variable), without giving sufficient justification for this choice and without providing results from comparative approaches. The aim of our paper is to challenge this attitude, showing to which extent it could generate misleading answers. Our results are: (i) sensitive to the specific econometric method employed, (ii) strongly affected by the heterogeneous character of subsamples (according to, for instance, sectoral patterns, size, geographical location and so on). In particular, since the recent literature on R&D policy evaluation has shown a decisive preference toward the use of Matching methods –notably, the nearest-neighbour version– we devote special attention to the robustness of this class of techniques both on target-variables expressed in level (such as the total R&D expenditure) and in ratio (such as R&D intensity and R&D per employee).

What can we learn from our exercise? Our analysis concludes that while for variables expressed as ratio the difference in results is negligible, R&D expenditure presents a very strong variability among the approaches. It means that, as the objective of R&D public policies is primarily the increase of national R&D outlay, our conclusion turns to be quite worrying both from an evaluation and policymaking perspective. Therefore, we suggest that a comparison of various methods is an essential step for assuring major robustness and fairness of the results and we propose to take this practice more into account for future works. Nevertheless, our results show also that while the choice of the method heavily affects the magnitude of the (estimated) effect of the policy, it does not seem to influence the sign and statistical significance of results. The latter are congruent among the various methods.

The paper also provides useful insights on the impact of Italian R&D and innovation policies on firm R&D performance since we have access to a pool of public R&D incentives managed at national, regional and European level. While we cannot identify the effect of single measures characterized by specific mechanisms of application, we can avoid the potential confounding presence of further (and hidden) R&D incentives when comparing supported and non-supported units (our

[1] More recent CIS data for Italy are not available at firm level yet. Generally, in fact, ISTAT (the Italian Office of Statistics) firstly publishes results at macro and sectoral level, and only after a careful validation, micro-data are provided to external researchers.

sample has 2,352 supported and 3,371 non-supported firms).[2] It goes without saying that we work in a counterfactual setting with a binary treatment variable (taking the form of supported vs. non-supported status), where the term of comparison for the group of supported firms are those innovating firms receiving no public incentives in the considered time span.

The evaluation setting we work with presents some limits. Two of them need to be mentioned: first, by using CIS data we cannot know the level of the subsidy, so that we can control only for the presence of a full crowding-out (rather than for additionality in the proper sense) of the policy considered; second, we can check only the short-run effect of the supporting policy, although an increase in the private R&D effort is expected to occur more likely in the medium term. We believe these two aspects to be important, although they should not affect too much the significance and scope of our achievements.

Our analysis excludes on average a total crowding-out of R&D policies when we look at the outcome of the pooled group of subsidized firms. As said before, it is a statistically robust achievement as confirmed by all the different methods applied. In short, we observe that: (i) in the pooled sample, by an average over seven Matching methods, we get 885 additional thousand euros of R&D expenditure with a ratio of supported to non-supported firms' performance equal to 4.62: it means that, when a generic control-unit performs 1 thousand euros of additional R&D expenditure, a matched treated unit performs 4.62 thousand euros. In the population[3] (when using sampling weights), we find that R&D investment's additionality drops to 242 thousand euros; (ii) the additionality for the R&D intensity is, according to an average over all methods, of about 0.014 in the pooled sample (meaning that supported units do a 1.4 % additional R&D on turnover, with a ratio of 2.67); in the population it drops to 0.010 (that is, a 1 % of additional R&D intensity compared to non-supported companies); (iii) finally, when looking at more disaggregated subsamples of firms (at dimensional, sectoral and geographical level), some cases of total crowding-out appear for low-knowledge-intensive-services (LKIS), very small firms (10-19 employees) and the auto-vehicle industry, while the other subgroups seem to have reached a satisfactory result in term of additionality (although with differential strength).

---

[2] As we will show below, missing values on covariates and data cleaning reduce considerably these sample sizes.

[3] The term "population" refers to whole Italian companies, as CIS3 and CIS4 samples are unbalanced towards larger firms, sectors and specific locations. The use of sampling weights in the estimation phase returns results that are representative of the actual Italian industrial structure.

The paper is organized as follows: Section II presents a short overview of the applied econometric literature on R&D policy evaluation, by setting out a taxonomy of the methods, and devoting special attention to the issue of subsidy's endogeneity; Section III provides a concise technical exposition of the methods applied in this paper, notably: the Heckman selection model (Heckit), and the Control-function, Matching, and Difference-in-differences (DID) methods; Section IV presents the datasets (CIS3 and CIS4[4]), and the target and control-variables with their difference-in-mean t-tests; Section V sets out the econometric results on the effect of R&D public support on target-variables, both for the whole sample and for various subgroups of firms; Section VI concludes the paper suggesting some improvements for future works.

II. A brief overview of the econometric literature

The literature on R&D policy evaluation is wide and rapidly growing. A leading review on the main results achieved by several different works at the micro and macro level is that by David, Hall and Toole (2000), presenting also a general demand/supply model for explaining the effect of an R&D policy incentive on firm R&D investment. The work of Klette, Møen and Griliches (2000) also contains useful insights and rich theoretical discussions on this issue. Finally, very recently, Cerulli (2010) presents a wide review of the main econometric models used so far for modelling and measuring the effect of public R&D financing on firm R&D and innovativeness.

The core issue raised by this literature concerns the endogenous nature of the subsidy within a non-experimental setting. Indeed, assuming the policy variable (subsidy) as strictly exogenous could be seriously misleading, since the R&D funding allocation could depend critically on: (i) firms, deciding (at least to some extent) on their participation status (*self-selection*) and, (ii) the government, choosing to finance particular subjects according to a specific objective function (by adopting, for instance, the principle of "picking-the-winner" or that of sustaining lagging developing areas). Such a non-random assignment of public funds generally embeds R&D policies in a non-experimental setting.

---

[4] CIS4 data are used longitudinally along with CIS3 for getting DID estimation results. The other estimators are assessed in a cross-section setting where only data from CIS3 are employed. The choice of using CIS3 instead of CIS4 in the cross-section analysis draws upon the fact that CIS4 was not available at the outset of this work. Nevertheless, given also the methodological nature of this study, this does not seem to represent a strong limitation.

In econometric terms this means that the treatment variable d (assuming value 1 for treated and 0 for untreated units) and the outcome variable y (assuming value y_1 for treated and y_0 for untreated units, when the latter are used as treated units' counterfactual status) are *stochastically dependent*. In this case, we cannot rely on the classical inference approach, i.e., the simple difference between the mean of treated and untreated units, to infer the effect of a public intervention on firm R&D performance. Indeed, by defining the Average Treatment Effect (ATE) as:

$$\text{ATE} = \text{E}(y_1 - y_0), \tag{1}$$

and the Average Treatment Effect on Treated (ATET) as:

$$\text{ATET} = \text{E}(y_1 - y_0 \mid d = 1), \tag{2}$$

it is easy to show that, when y and d are supposed to be (mean) independent, ATE and ATET coincide with the difference-in-mean estimator of basic statistics (i.e., the average of y for treated minus the average of y for non-treated individuals). This estimator, as is well known, is unbiased, consistent and asymptotically normal (see Wooldridge 2002: 606). Nevertheless, when this (mean) independence hypothesis does not hold, ATE and ATET generally differ and, most importantly, the difference-in-mean estimator becomes inconsistent.

To overcome this estimation problem, econometricians have suggested a plethora of approaches under specific hypotheses, showing their comparative advantages and drawbacks, depending on the underlying process generating the data. Good reviews on the topic are: Heckman (2001), Cobb-Clark and Crossley (2003), and Blundell and Costa Dias (2002). Table 1 sets out a taxonomy of some representative studies found in the R&D policy evaluation literature according to the model employed (structural or based on a reduced-form), dataset (cross-sectional or longitudinal) and type of supporting variable (binary or continuous). The column on methods shows the econometric approaches used in these studies.

Structural models - which we can roughly identify with Selection and Instrumental-variable (IV) models – should in principle better explain the rationale of the subsidy effects, since they explicitly model the interaction between firm's and public agency's behaviour within a system of simultaneous equations. In particular, Selection models make use of a system of two linked relations explaining: (i) the firm R&D investment decision and/or self-selection (into program) on one hand, and (ii) the public agency selection rule on the other. This class of models, moreover, can be consistent with the common analytical frame of a firm and an agency maximising an objective function (Klette and Møen 1998, David, Hall and Toole 2000). IV and Selection models (of the type proposed, for instance, by Heckman 1978 generally known as the "Heckit" model), are well suited for taking into account potential endogeneity caused by observable as well as unobservable factors (i.e., in the so-called case of "selection on observables" and "selection on unobservables"). This is an advantage of these methods, although they require either substantial additional information (the availability of at least one instrumental variable in IV) or strong distributional hypotheses (such as the bivariate normal distribution of unobservables in the Heckit) to be applied.

Nonetheless, the most part of applied works in the field of R&D policy evaluation makes use of less structural approaches, namely: Control-function regression, and Matching and Difference-in-differences (DID) methods. Control-function relies on a usual regression analysis augmented for those covariates determining both the agency selection process and the firm R&D behaviour and self-selection. Matching is a non-parametric estimation procedure which reduces the group of non subsidized firms to a sub-sample of units with characteristics more homogeneous to the subsidized ones. These more "empirical" methods avoid to specify a structural system, as they are based on a reduced-form equation in which theory enters only through the choice of variables aimed at explaining the non-random assignment of funds (Control-function), or at homogenizing subsidized and non-subsidized units (Matching).[5] The main

Table 1. R&D policy evaluation studies according to specification, dataset and policy variable

Method	Model		Dataset		Policy variable		Representative studies
	Structural	Reduced-form	Cross-section	Longitudinal	Binary	Level	
Instrumental variables	X		X			X	Wallsten (2000)
Selection model (Heckit)	X		X		X		Busom (2000)
Control function		X	X			X	Lichtenberg (1987)
Matching		X	X		X		Almus and Czarnitzki (2003)
Difference-in-differences (DID)		X		X	X		Lach (2000)

[5] Studies using Matching in an R&D policy evaluation context are, for instance: Almus and Czanitzki (2003), Duguet (2004), Aerts and Czanitzki (2004), Kaiser (2004), Lööf and Heshmati (2005), and Bérubé and Mohnen (2007). An interesting paper combining Matching with Difference-in-differences (DID) is that by Görg and Strobl (2007).

drawback of Control-function and Matching is that they implicitly assume that agency's unobservable criteria of selection and firm R&D behaviour/self-selection are uncorrelated. This identification assumption can bring about severe biases especially when too few observables are at disposal of the researcher. Difference-in-differences, finally, is generally used when a longitudinal dataset is available. When idiosyncratic and time fixed-effects are incorporated in the model, DID consistently estimates ATET also under selection on unobservables, without any use of instrumental variables or distributional hypotheses. In this sense DID is a powerful and easy-to-apply method in many evaluation contexts.

Starting from this sketched overview, it seems interesting to check to which extent these methods generate similar (or dissimilar) results once applied to the same sample of data. We concentrate our attention on Heckit, Control-function, Matching, and Difference-in-differences methods, and omit IV estimations because CIS data do not provide a sound and reliable set of instrumental variables.

III. Heckit, Control-function, Matching, and Difference-in-differences: a concise exposition

This section provides a brief technical introduction to the approaches compared in this paper, notably: Heckman selection model (Heckit), Control-function regression, Matching, and Difference-in-differences (DID).

Heckit

As said above, the Heckit model is suitable also under selection on unobservables. It is composed of a system of two (correlated) equations, one for the R&D outcome and one for the agency's selection equation:[6]

$$\begin{cases} y_i = \mu + \gamma \mathbf{Q}_i + \alpha d_i + u_i, \\ d_i^* = \eta + \beta \mathbf{Z}_i + v_i, \\ d_i = \begin{cases} 1 & if \quad d_i^* \geq 0, \\ 0 & if \quad d_i^* < 0, \end{cases} \\ Cov(u_i; v_i) = \rho \neq 0, \end{cases} \tag{3}$$

[6] See Busom (2000) for an application to an R&D supporting program.

where **Q** and **Z** are covariates and u and v are unobservable components (error terms) with zero unconditional mean, but assumed to be correlated. Under this assumption $E(d_i \cdot u_i) \neq 0$ so that the OLS estimate of the outcome equation is inconsistent. Indeed, rewriting the first equation of (3) in the two different regimes, we get:

$$d_i = 1: \quad y_i = \mu + \gamma \mathbf{Q}_i + \alpha + u_i,$$
$$d_i = 0: \quad y_i = \mu + \gamma \mathbf{Q}_i + u_i.$$

It would seem possible to run two OLS regressions on them, obtaining α, i.e., the ATET, as the difference between the two (estimated) intercepts. Unfortunately, the problem of this procedure is that under both the regimes the error term does not have a zero unconditional mean. In fact:

$$\mathrm{E}(u_i \mid v_i \geq -\eta - \beta \mathbf{Z}_i) \neq \mathrm{E}(u_i) = 0,$$
$$\mathrm{E}(u_i \mid v_i < -\eta - \beta \mathbf{Z}_i) \neq \mathrm{E}(u_i) = 0.$$

This is a typical case of omitted variable specification error that can be solved by adding the non-zero means into the equations, obtaining:

$$\begin{aligned} d_i = 1: \quad & y_i = \mu + \gamma \mathbf{Q}_i + \alpha + \mathrm{E}(u_i \mid v_i \geq -\eta - \beta \mathbf{Z}_i) + [u_i - \mathrm{E}(u_i \mid v_i \geq -\eta - \beta \mathbf{Z}_i)], \\ d_i = 0: \quad & y_i = \mu + \gamma \mathbf{Q}_i + \mathrm{E}(u_i \mid v_i < -\eta - \beta \mathbf{Z}_i) + [u_i - \mathrm{E}(u_i \mid v_i < -\eta - \beta \mathbf{Z}_i)]. \end{aligned} \tag{4}$$

Now, the errors terms in the squared brackets have zero mean**.** The problem is that we cannot observe $\mathrm{E}(u_i \mid v_i \geq -\eta - \beta \mathbf{Z}_i)$ and $\mathrm{E}(u_i \mid v_i < -\eta - \beta \mathbf{Z}_i)$. Nevertheless, we can estimate them by using the participation equation and the joint normality of u and v. From the joint normality it can be proved that:

$$\mathrm{E}(u_i \mid v_i \geq -\eta - \beta \mathbf{Z}_i) = -\lambda_1 M_{1i},$$
$$\mathrm{E}(u_i \mid v_i < -\eta - \beta \mathbf{Z}_i) = -\lambda_0 M_{0i},$$

where $M_{1i} = \phi(-\eta - \beta \mathbf{Z}_i) / [1 - \Phi(-\eta - \beta \mathbf{Z}_i)]$ and $M_{0i} = \phi(-\eta - \beta \mathbf{Z}_i) / [\Phi(-\eta - \beta \mathbf{Z}_i)]$ are known as *Mill's ratios* (with ϕ and Φ being the normal density function and its cumulative function respectively), while $\lambda_1 = \sigma_u \cdot \sigma_{u,v}$ and $\lambda_0 = -\sigma_u \cdot \sigma_{u,v}$.

We can estimate equations (4) by a two-step procedure or via maximum likelihood (Maddala 1983). In the two-step procedure we first estimate M_{1i} and M_{0i} (once obtained a consistent estimation of η and β from a Probit regression of the participation equation); secondly, with these estimations at hand, we can estimate λ_1 and λ_0 by simple OLS (taking standard errors corrected for generated regressors). We might then estimate also the coefficient of correlation ρ between u and v (as $\rho = \lambda_1 / \sigma_u^2$). Since, under joint normality of (u, v) this method becomes fully parametric, a partial maximum likelihood approach can be used to estimate consistently all parameters. Observe that the sign of ρ shows whether non-observables in the participation and non-observables in the outcome are positively or negatively correlated. Finally, to make the Heckit results as much as comparable with the other methods, in our application we hold **Q=Z=X**, where **X** are the covariates used in the Control-function, Matching and DID estimators, so that we try to reduce any arbitrariness in choosing different sets of **Q** or **Z** in the two Heckit equations.[7]

Control function

The Control-function is based on the standard multiple regression analysis and is aimed at estimating the simple switching equation:

$$y_i = \mu + \gamma \mathbf{X}_i + \alpha d_i + \varepsilon_i \, , \tag{5}$$

where the treatment variable (named as d) is inserted in the right-hand-side along with other covariates (**X**) that the researcher believes to describe sufficiently well the non-random allocation of R&D supports. The coefficient for d can be consistently estimated by OLS (or GLS in case of firm heteroskedasticity), under the additional hypothesis that unobservable confounders are not at work (or have negligible impact). See Lichtenberg (1987) for an application in an R&D policy context.

Matching

Matching is consistent under the same hypothesis as the Control-function (i.e., selection on observables). Nonetheless, it is generally preferred to the Control-function for at least three reasons. Firstly, it is a non-parametric estimation procedure, so it does not need to specify a particular parametric relation between the dependent

[7] An interesting extension of the Heckit two-step approach in a semi-parametric environment has been performed by Hussinger (2008). The author measures R&D subsidy additionality for a sample of German firms by comparing results from standard OLS and parametric Heckit with five semi-parametric estimators. While statistical significance is concordant, remarkable variability of results appear across methods.

variable and the regressors. Secondly, Matching considers only treated and non treated units in the *common support* (by dropping all the controls the value of whose variables is higher or smaller than that of the treated). Thirdly and more importantly, Matching reduces the number of non-treated to a sub-sample (the *selected controls*) with characteristics more homogeneous to the treated units (Cameron and Trivedi 2005: 871-878). The idea behind Matching is to estimate the unobservable quantity $E(y_0 | d=1)$, that is specifically what the average outcome for treated units would have been if they had not been treated, using non-treated units that are "similar" to the treated ones. This similarity can be checked in relation to several firm characteristics such as size, cost and financial variables, sector and so on. When for each treated unit one (or more, depending on the type of Matching) similar non-treated unit(s) has been selected from among all the potential non-treated units, a comparable sub-sample is produced and it can be proved that ATET is consistently estimated. In other words, Matching estimates $E(y_0 | d = 1)$ with those non-treated firms that are like "twins" of the treated ones. More precisely, we have:

$$E(y_0 | d=1, \mathbf{X}=\mathbf{x}) = E(y_0 | d=0, \mathbf{X}=\mathbf{x}). \tag{6}$$

Relation (6) is valid only under the *conditional independence assumption* (Rubin 1977, Rosenbaum and Rubin 1983): conditional on some pre-treatment observables (the variables **X**), we assume y and d to be stochastically independent. In this case, the estimation of the ATET conditional on **X** becomes exactly:

$$\text{ATET}(\mathbf{x}) = E(y_1 | d=1, \mathbf{X}=\mathbf{x}) - E(y_0 | d=0, \mathbf{X}=\mathbf{x}),$$

where an estimate of the unconditional ATET is obtained by averaging the previous equation over the support of **X**.

When **X** is highly dimensional or in a continuous form, exact matching is not possible. To avoid this drawback (known as *dimensionality problem*), Rosenbaum and Rubin (1983) proposed to match units according to a single variable: the *propensity score*, defined as the probability of becoming treated conditional on **X**. The score is obtained through a Probit regression of d on variables contained in **X**. Units with a close propensity score are matched and a counterfactual sample from the untreated group can be generated on this basis and compared with the treated one.

Nevertheless, before making the comparison, the so-called "balancing property" needs to be satisfied. It requires that each regressor contained in **X** (and also the propensity score) is "balanced" within strata formed by building intervals of propensity scores (the so-called "blocks"). Balancing entails that in each stratum (or block) and for each regressor, the mean of treated units equals that of control units. Once this property is satisfied to some acceptable statistical significance, the researcher can apply reliably the Matching estimation. Different types of propensity score Matching have been proposed in the literature: Stratification, One-to-one nearest-neighbour, Multiple-nearest-neighbours, Kernel, and Radius (with various callipers) are among the most used (Caliendo and Kopeinig 2008, Dehejia and Wahba 2002, Heckman, Ichimura, and Todd 1998). A general sample formula to estimate the ATET by Matching is:

$$\mathrm{AT\hat{E}T}^{\mathrm{M}} = \frac{1}{N^T}\sum_{i\in T}\left[y_{1i} - \sum_{j\in C(i)} w_{ij} y_{0i}\right],$$

where: i is the treated unit with characteristics $\mathbf{X}_i$ and N^T the number of treated firms; j are those untreated units belonging to the set $C(i) = \left\{j \mid p_j(\mathbf{X}) \in \mathrm{D}_i(p_j(\mathbf{X}))\right\}$ where $p_j(\mathbf{X})$ is the propensity score and $\mathrm{D}_i(p_j(\mathbf{X}))$ the "characteristics neighborhood" of $p_j(\mathbf{X})$; and, finally, w_{ij} are the weights given to each $j \in C(i)$. Different Matching estimators are generated by varying the choice of $C(i)$ and w_{ij}. Table 2 shows those Matching methods we choose to compare in our application (where N_i^C is the number of control units associated to the treated unit i).

The stratification matching, instead, assumes a bit more complicated formula:[8]

$$\mathrm{AT\hat{E}T}^{\mathrm{S}} = \sum_{b=1}^{B} \mathrm{AT\hat{E}T}_b^{\mathrm{S}} \cdot \left[\frac{\sum_{i\in I(b)} d_i}{\sum_i d_i}\right] \text{ with: } \mathrm{AT\hat{E}T}_b^{\mathrm{S}} = \frac{1}{N_b^T}\sum_{i\in I(b)} y_{1i} - \frac{1}{N_b^C}\sum_{j\in I(b)} y_{0j},$$

Table 2. Different Matching methods according to the specification of $C(i)$ and w_{ij}

Matching method	$C(i)$	w_{ij}
1. One-to-one nearest-neighbour	*Singleton* $j \mid \min_j \lVert P_i - P_j \rVert$	1
2. Three-nearest-neighbours	*First three* $j \mid \min_j \lVert P_i - P_j \rVert$	1/3
3. Kernel	All control units (C)	$K(P_j - P_i/h) / \sum_{j\in C} K(P_j - P_i/h)$
4. Radius	$j \mid \lVert P_i - P_j \rVert < r$	$w_{ij} = 1/N_i^C$

[8] Standard errors for all these estimators can be obtained analytically or via bootstrapping.

where $I(b)$ is the set of units present in block b, N_b^T is the number of treated units in block b, N_b^C is the number of control units in block b. The number of blocks, B, are those obtained when the balancing property is satisfied.

Few studies have compared the performance of these different kinds of Matching estimators. Dehejia and Wahba (2002: 158) found that "The choice among matching methods becomes important when there is minimal overlap between the treatment and comparison groups". They conclude that, either in presence of greater or smaller overlap, the nearest-neighbour Matching performs quite well compared to the others. Indeed, when the true ATET coming from the benchmark (notably, a previous real experimental setting) is about \$ 1,794, the nearest-neighbour's ATET is equal to about \$ 1,360 in the case of greater overlap and \$ 1,890 in the case of smaller overlap. Starting from the same database of Dehejia and Wahba (2002), Cameron and Trivedi (2005: p. 893-896) have shown, on the contrary, that the nearest-neighbour Matching performs worse than other Matching methods when slight modifications in the controls' selection criteria are implemented (such as the "common support" restriction). They obtain a nearest-neighbour's ATET of about \$ 2,385 that overestimates the true value of \$ 1,794 using the same Dehejia and Wahba (2002) propensity score specification. Zhao (2004: 100) compared various Matching models in a Monte Carlo experiment, concluding that "Monte Carlo experiments show that the different methods do not dominate each other in terms of performance". Generally speaking, methods perform very differently according to: (i) the availability of good controls, (ii) their number, and (iii) the specification of the propensity score equation. In the context of innovation policy evaluation, finally, the paper by Arvanitis et al. (2010) compares the performance of four Matching methods on six measures of firm innovation performance. They find very concordant results for all these methods, with opposite sign for firms with low and high levels of subsidization.

Starting from this premise, what we do in our exercise it to implement the previous Matching formulas by the following procedure: First, we estimate $p(\mathbf{X})$ –the propensity scores– by a Probit regression of d on $\mathbf{X}$ for the all sample. Second, we test the balancing property. And, third, if the balancing property is satisfied, we estimate the ATET using the previous formulas. If not, we modify the model by adopting another specification of the Probit regression until the balancing property is satisfied.

Difference-in-differences (DID)

In a longitudinal data setting, the DID model takes the following form:

$$\begin{cases} y_{it} = \mu + \gamma \mathbf{X}_{it} + \delta \mathbf{Z}_i + \alpha d_{it} + \theta_i + \eta_t + \varepsilon_{it}, \\ d_{it-1} = 0, \end{cases} \tag{7}$$

where μ is a constant term, $\mathbf{X}_{it}$ are time-variant covariates, $\mathbf{Z}_i$ time-invariant regressors, θ_i the firm-specific fixed effect, η_t the time-specific fixed effect and ε_{it} an i.i.d. error term uncorrelated with d_{it}, the treatment variable, once controlled for $\mathbf{X}_{it}$ and $\mathbf{Z}_i$. The condition $d_{it-1} = 0$ suggests to restrict the estimation of the ATET (that is, α) to the subsample of all firms that were untreated in *t-1*, that hence becomes a *ceteris paribus* starting condition. As θ_i and η_t can be considered as proxies of firm unobservable heterogeneity and since in the previous regression both can be thought of as freely correlated with d_{it}, $\mathbf{X}_{it}$ and $\mathbf{Z}_i$, the DID proves to be a consistent estimation of the α under the hypothesis of selection on unobservables. This is what makes the DID particularly appealing in the context of econometric policy evaluation. As in our setting we only consider two time periods corresponding to CIS3 (1998-2000) and CIS4 (2002-2004), that we indicate with the subscript 0 and 1 respectively, we can rewrite (7) in this form:

$$\begin{cases} y_{i0} = \mu + \gamma \mathbf{X}_{i0} + \delta \mathbf{Z}_i + \alpha d_{i0} + \theta_i + \eta_0 + \varepsilon_{i0}, \\ y_{i1} = \mu + \gamma \mathbf{X}_{i1} + \delta \mathbf{Z}_i + \alpha d_{i1} + \theta_i + \eta_1 + \varepsilon_{i1}, \\ d_{i0} = 0. \end{cases} \tag{8}$$

By taking the first difference of the two relations in (8) we get:

$$\begin{cases} \Delta y_{i1} = \gamma \Delta \mathbf{X}_{i1} + \alpha d_{i1} + \Delta \eta_1 + \Delta \varepsilon_{i1}, \\ d_{i0} = 0, \end{cases}$$

that simplifies substantially the regression as time-invariant components drop out after the difference. Observe that, since $d_{i0} = 0$ then $\Delta d_{i1} = d_{i1} - d_{i0} = d_{i1}$.

The DID estimator of α is equal to the OLS estimation of the previous regression once conditioned on d_{i0}=0. Furthermore, observe that $\Delta\eta_1$ is the regression constant term so that its standard t-test of significance is meant to test whether or not time has had some importance in explaining the temporal variation of the considered target-variable (in our case, it measures to which extent passing from CIS3 to CIS4 has produced some structural effect). See Lach (2002) for an application of DID to an R&D program.

IV. Datasets, variables and difference-in-mean t-tests

CIS3 refers to the years 1998-2000 and collects 149 variables for 15,512 manufacturing and service firms. CIS4 refers to the years 2002-2004 and collects a comparable number of variables (with just slight modifications in the questionnaire) for 16,537 manufacturing and service companies. We use the CIS3 cross-section setting for applying the Control-function, Matching and Heckit, while the longitudinal dataset obtained by merging CIS3 and CIS4 is used for DID estimation. The merging with CIS3 provides a sample of 5,923 firms observed in both periods.

Both datasets, CIS3 and CIS4, are merged with firm balance sheet variables coming from the Italian Chamber of Commerce civil accounts, containing information on firm accounting variables and on the statement of assets and liabilities. While firm R&D expenditure (as well as the other target-variables) refers to 2000 and 2004 respectively, we define a treated (or supported) firm as one answering at least one "yes" to the questions regarding R&D funding from central and local government and from the EU (and in particular from the EU Framework Programs[9]) within the period 1998-2000 (CIS3) and 2002-2004 (CIS4) respectively. Since only innovating firms answer to the CIS funding questions, the total sample accounts for 2,352 supported units 3,371 non-supported ones (41% and 59% of the total, respectively).

When analyzing CIS3 for the distribution of public subsidies according to the type of financing source (see Table 3), firms receiving a "Government" fund represent about 22% of our sample (a value becoming 16% for the industrial population, once sample weights are considered). Firms receiving "Local" funds are about 20% in the sample, a value that doesn't change too much when reported to the entire population (22%); firms supported by "EU" funds are about 8% in the sample and 6% in the population; firms getting a "Framework program" fund, finally, are about 4% in the sample and about 2% in the population.[10]

---

[9] Questions 9.1a, 9.1b, 9.1c and 9.2 of the CIS3 questionnaire.

[10] Note that receiving or not a certain type of fund does not exclude a firm to benefit from other funds. More interestingly, when considering only national funds ("Government" and/or "Local") the number of supported firms (compared to the non-supported units) becomes about the 35% against the 41% as showed above for the whole subsidization. It means that considering the overall rather than the national

Table 3. Distribution of public subsidies according to different sources of financing

	Government funds				Local funds		
	Freq.	Share	Share in the population		Freq.	Share	Share in the Population
Receiving	1,275	22.28	16.4	Receiving	1,145	20.01	22.29
Non-receiving	4,448	77.72	83.6	Non-receiving	4,578	79.99	77.71
Total	5,723	100	100	Total	5,723	100	100
	EU funds				FP funds		
	Freq.	Share	Share in the population		Freq.	Share	Share in the population
Receiving	460	8.04	6.31	Receiving	230	4.02	2.38
Non-receiving	5,263	91.96	93.69	Non-receiving	5493	95.98	97.62
Total	5,723	100	100	Total	5723	100	100

Source CIS3 data.

We choose to test the effect of the R&D and innovation policy on three target-variables. As for CIS3 cross-section analysis they are:[11] intra-muros R&D in thousand euros in 2000 (*R&D EXPENDITURE 00*); the ratio between R&D expenditure and firm turnover in 2000 (*R&D INTENSITY 00*); and the ratio between R&D expenditure and firm number of employees in 2000 (*R&D PER EMPLOYEE 00*).[12]

---

subsidization should not change too much the sample and the results. As a simple proof, we also performed a simple OLS estimation of the ATET when only "national" subsidization is considered. We found results very similar to the case of the overall subsidization both in terms of magnitude and statistical significance. It means that we should accept with some degree of confidence that it is the national part of the R&D and innovation support that drives on average the results achieved in terms of additionality.

[11] For CIS4 (CIS3) we consider target-variables in 2004 (2000) and pre-treatment covariates, when possible, in 2002 (1998).

[12] In our estimations we also considered as target-variable the "share of the innovative turnover on total turnover in 2000". Results showed that the impact of R&D subsidies on this variable, meant as approximating firm innovative capacity, is generally statistically insignificant. Nevertheless, given the downstream nature of the innovative activity, the effect of public R&D subsidies should be in this case less direct than in that of the R&D-based ones. Furthermore, since innovative turnover is subjectively estimated by the entrepreneurs, the reliability of this indicator, as proxy of firm innovativeness, has to be taken with some caution. For this reason we preferred not to report results on this variable in this paper.

As for the control-variables, we considered a set of covariates upon which there exists a wide consensus in the current literature as those driving the non-random assignment of public funds.[13] When possible, we also take them at 1998 in order to avoid simultaneity with the target-variables, and give them a pre-treatment status. Therefore, our control-variables are: (1) The number of firm employees in 1998 (*EMP 98*). Size is commonly recognized as a leading variable in explaining firm ability to attract financing. Scale economies and a richer set of perceived opportunities generally increase with size; (2) The share of employees with a degree or university diploma on total employees in 2000 (*EMPSKILL 00*). A higher human capital should positively affect the probability of attracting financing. More skilled workers should enhance the capacity of writing projects, promoting fund rising strategies and improving knowledge of opportunities; (3) The share of turnover stemming from exportations on total firm turnover in 1998 (*EXPINT 98*). Supposedly, more internationalised firms operate under a more competitive pressure leading to search for diversified portfolio strategies to attract innovative capacity, such as applications for public funds; (4) The capital stock (from balance sheet) per employee in 1998 (*CAPINT 98*). The higher the capital intensity of a firm, the more it should have an incentive to search for a lifelong technological upgrading by exploiting, among various possibilities, also public subsidies; (5) the cash-flow per employee in 1998 (*CASHINT 98*). A large cash-flow identifies a necessary condition for augmenting firm self-financing: the greater its level, the lower the need to depend on external resources; (6) the share of firm total stock of debt on total liabilities in 1998 (*DEBTINT 98*). A higher debt represents a financial constraint for a firm that can find increasing difficulties to finance its activity by either further indebtedness or equity. In this case, firms can try to attract non-market funds such as public subsidies; (7) the value of intellectual property rights (such as patents) and capitalized R&D expenditures per employee in 1998 (*KNOWLEDGE 98*). Past innovative performance (experience) should matter in attracting current subsidies especially when government implements a policy aimed at awarding previous winners; (8) A dummy variable indicating if the firm belongs to a foreign group or not (*FOREIGN*). The nationality of the mother-firm could be determinant in providing incentives for applying for public subsidies; (9) A dummy variable assuming value one whether the firm was set up between 1998 and 2000 (*AGE*). Along its life-cycle,

[13] We share with the majority of studies in this field a lack of information on the quality of firm R&D proposed projects.

Table 4. Differences in the control and target-variables for R&D supported and non-supported firms

Group	Obs.	Mean	Std. error
Control variables			
EMP 98			
Supported	2369	254	38.15
Difference		10	79.56
EMPSKILL 00			
Supported	1754	0.11	0.004
Difference		0.005	0.005
EXPINT 98			
Supported	2369	0.215	0.006
Difference		-0.076***	0.007
CAPINT 98			
Supported	1959	49.61	3.57
Difference		13.52	15.06
CASHINT 98			
Supported	1959	3.72	0.343
Difference		1.94*	1.003
DEBTINT 98			
Supported	1959	0.642	0.004
Difference		0.011*	0.006
KNOWLEDGE 98			
Supported	1541	1.48E+09	8.11E+08
Difference		-1.29E+09*	7.37E+08
FOREIGN			
Supported	2369	0.936	0.005
Difference		-0.052***	0.008
AGE			
Supported	2369	0.026	0.003
Difference		-0.012**	0.005
Target variables			
R&D EXPENDITURE 00			
Supported	2369	969.6	214.05
Difference		-840.5***	179.43
R&D INTENSITY 00			
Supported	2369	0.02	0.001
Difference		-0.014***	0.001
R&D PER EMPLOYEE 00			
Supported	2369	2.715	0.144
Difference		-1.868***	0.138

Source CIS3 data. "Difference" is the difference between the mean of non-supported and supported firms. The t-tests significant at the 1% level are marked ***; at the 5% level, **; at the 10% level, *.

the firm maturity can be an important feature for attracting subsidy opportunities; (10) A dummy variable taking value one if the firm belongs to a group of firms (*GROUP*). A firm belonging to a group can be more able than others in receiving information on possible financing possibilities; (11) A geographic stratification variable splitting the sample into 10 Italian macro regions (*GEO*); (12) A sectoral stratification variable according to the two-digit Nace Rev. 1 classification[14] (*SECTOR*); (13) A dimensional stratification variable (*SIZE*) splitting the sample into four dimensional groups: small (10-19 employees), medium-small (20-49), medium-large (50-249) and large (more than 250).[15]

Table 4 displays standard t-tests on the difference-in-mean of both control and target-variables for CIS3.[16] The number of employees of non-supported firms is similar to that of the funded ones (around 250 employees) and the difference is not significant; the average percentage of skilled employees is again similar between the two groups (around 11%); export intensity, on the contrary, is different in the two groups (21% in the financially supported and about 14% in the non-treated units); capital intensity is not significantly different, while the cash-flow per employee is significantly different (according to our predictions non-supported firms present a larger cash-flow intensity, about 6 thousand euros per employee against 4 thousand euros in the others); the debt intensity is significantly different, but the distance between the two groups is not too sharp; the immaterial assets (knowledge) identifies a significant difference between the two groups of firms, with supported firms showing a greater level of knowledge accumulation; finally *AGE*, *FOREIGN* and *GROUP* all set out significant differences, with supported firms generally younger, owned by a foreign company and not belonging to a group.[17]

As for the target-variables (the endogenous variables of our application), Table 4 puts in evidence some interesting aspects. R&D expenditures in 2000 presents a strong difference between supported (969 thousands euros) and non-supported firms (129 thousand euros). The R&D intensity also reveals a strong diversity between the two groups (with a value of 0.6% for non-supported and a value of 2% for

[14] Nace is the official statistical classification of economic activities in the European Community.

[15] *GEO*, *SECTOR* and *SIZE* are the stratification variables adopted for sampling firms in the Italian CIS (see "Nota metodologica" in ISTAT 2004: 45-46).

[16] This analysis is dropped for CIS4 as Matching is applied only using CIS3.

[17] The variable *GROUP* is not shown in Table 4 since it had problems satisfying the "balancing property". Nevertheless, we found that the difference of this variable in the two groups is significant, with non-supported firms, on average, more typically belonging to a group of firms.

supported units). The R&D per employee is again very diverse in the two groups (about 2.7 thousand euros of R&D per employee in treated units and 0.8 thousand euros in non-supported firms).

V. Econometric results

According to the procedure depicted in section III, we start our analysis by calculating the propensity scores by a Probit regression using sampling weights (so that we can get parameters valid for the entire population). Results are reported in Table 5. The number of sample observations drops to 2,574 because of the great number of missing values in balance sheet variables. The regression fits quite well since the Chi-square test is highly significant for the overall regression and the pseudo R-square is about 12%.

Table 5. Probit regression of treatment on exogenous variables to identify propensity scores

d = treatment	Coefficient	Std. error	*ey/ex*	Std. error
EMP 98	9.28e-08	8.94e-06	0.00001	0.0015
EMPSKILL 00 ***	0.82***	0.082	0.072***	0.0072
EXPINT 98 ***	0.22***	0.045	0.043***	0.0089
CAPINT 98 ***	-0.00009***	0.00003	-0.006***	0.0019
CASHINT 98 *	-0.001***	0.00056	-0.0044**	0.0023
DEBTINT 98	0.0003	0.05889	0.0002	0.0353
KNOWLEDGE 98 ***	1.10e-10***	1.99e-11	0.027***	0.0049
FOREIGN	0.60***	0.045	-	-
AGE	-1.035***	0.1108	-	-
CONS	-6.5078***	0.7242	-	-
SECTOR ***				
GEO ***				
SIZE ***				
Number of observations	2574			
LR chi2	2472.43			
Prob > chi2	0.000			
Pseudo R2	0.118			
Log likelihood	-9211.25			

Notes: for the sake of brevity single coefficients for *SECTOR*, *GEO*, and *SIZE* are not reported and the related p-value is for testing the hypothesis of global significance of these variables. *ey/ex*=marginal effects. * = significant at 10 %; ** = significant at 5 %; *** = significant at 1 %. The regression takes into account sampling weights.

Following Table 5 covariate by covariate, we briefly comment on the "elasticity value" (*ey/ex*), calculated holding all variables equal to their sample mean: *EMP* is not significant, with an elasticity around zero; *EMPSKILL* is highly significant with a positive sign and an elasticity around 7%: it means that if the *EMPSKILL* doubles, then the probability to become treated increases about 7%; *EXPINT* is highly significant, positive and with an elasticity of about 4%; *CAPINT* is significant with a negative and low elasticity; according to our prediction *CASHINT* is negative and significant too; *DEBTINT* is not significantly different from zero; *KNOWLEDGE* is significant and positive with an elasticity of about 3%; *FOREIGN* is significant with a positive sign; *AGE* is significant with a negative sign; finally, the CIS stratification variables, *SECTOR*, *GEO* and *SIZE* are all highly significant.[18]

Let us now comment on results on ATET. Table 6 shows the estimation of the Average Treatment Effect on Treated (ATET) according to seven different Matching procedures[19], a weighted OLS (using sample weights), an un-weighted OLS, the Heckit model and the Difference-in-differences (DID) for a total of ten different methods applied. We indicate with the symbol Δ the ratio of the value of the target-variable calculated on treated to that calculated on control units, with ρ the correlation between the unobservables in the outcome and selection equations within the Heckit and with $\Delta\eta$ the time fixed-effect difference between CIS3 and CIS4. We also distinguish between the ATET calculated within the sample and ATET calculated for the whole population, obtained using a proportional rule based on sample weights.[20]

It is easy to observe that the number of observations by type of Matching decreases according to the increase in the selectivity of the methods applied. Matching 7 (Radius with a calliper of 0.00001) is the most selective and the number of treated drops to 36 units. It means that there exists in this case a sort of *trade-off* between

---

[18] For the sake of brevity, Table 5 sets out for these three variables only the p-value of the F-test for the null hypothesis of joint overall significance of parameters.

[19] An example of a test for the balancing property, as well as a graphical representation of the propensity scores distribution of treated and untreated firms before and after the One-to-one nearest-neighbour Matching, are reported in Appendix A.

[20] While for the Control-function and DID we use correctly weighted regressions, for Matching and Heckit –because of some computational problems– we employ a proportional rule based on OLS population results to get population parameters. For the endogenous variable Y (such as, for example, R&D intensity), the ATET in the population ($ATET_P$) is then obtained in this case according to: Y_P: $Y_S = ATET_P$: $ATET_S$. Since it is a heuristic rule, population values have to be taken purely as an indication in those cases. Therefore, comments and conclusions will be drawn basically on sample results.

Table 6. Estimation of the Average Treatment Effect on Treated (ATET)

Evaluation procedure	Endogenous variable	Number of treated	Number of controls	ATET (Sample)	ATET (Population)	S.E.	Δ	t-value
Matching								
1. Stratification	R&D intensity ***	1218	3879	0.016	0.012	0.002	-	6.69
	R&D expenditure ***	1218	3879	1245.3	340.9	392.7	-	3.16
	R&D per employee ***	1218	3879	2.172	1.514	0.302	-	7.20
2. One-to-one nearest neighbour	R&D intensity ***	1207	650	0.015	0.011	0.002	2.55	5.65
	R&D expenditure **	1207	650	1038.9	284.4	382.2	5.49	2.72
	R&D per employee ***	1207	650	1.941	1.353	0.034	2.29	5.66
3. Three-nearest neighbour	R&D intensity ***	1207	1053	0.016	0.012	0.002	2.88	6.97
	R&D expenditure **	1207	1053	1001.7	274.2	381.1	4.73	2.63
	R&D per employee ***	1207	1053	2.066	1.440	0.290	2.50	7.10
4. Kernel	R&D intensity ***	1123	1281	0.013	0.010	0.001	2.71	6.76
	R&D expenditure ***	1123	1281	340.3	93.16	55.55	2.98	6.12
	R&D per employee ***	1123	1281	1.825	1.272	0.262	2.46	6.95
5. Radius 1	R&D intensity ***	1113	1133	0.013	0.010	0.002	2.62	6.49
(r = 0.001)	R&D expenditure **	1113	1133	693.4	189.8	302.8	3.91	2.29
	R&D per employee ***	1113	1133	1.932	1.347	0.283	2.56	6.81
6. Radius 2	R&D intensity ***	336	337	0.010	0.008	0.002	2.42	3.85
(r = 0.0001)	R&D expenditure	336	337	1351.7	370.0	981.4	6.17	1.38
	R&D per employee ***	336	337	1.514	1.055	0.474	2.17	3.19
7. Radius 3	R&D intensity *	36	38	0.014	0.011	0.007	2.85	1.78
(r = 0.00001)	R&D expenditure	36	38	525.2	143.8	442.6	4.43	1.19
	R&D per employee ***	36	38	2.097	1.462	0.814	2.93	2.58

Table 6. (continued) Estimation of the Average Treatment Effect on Treated (ATET)

Evaluation procedure	Endogenous variable	Number of treated	Number of controls	ATET (Sample)	ATET (Population)	S.E.	Δ	t-value
OLS								
8. OLS regression	R&D intensity ***	1308	1319	0.014	0.011	0.0017	-	8.00
(unweighted)	R&D expenditure ***	1308	1319	1127.6	308.7	397.8	-	2.83
	R&D per employee ***	1308	1319	1.897	1.322	0.231	-	8.19
9. OLS regression	R&D intensity ***	1308	1319	0.010	0.010	0.0005	-	18.18
(weighted)	R&D expenditure ***	1308	1319	538.7	538.7	108.2	-	4.98
	R&D per employee ***	1308	1319	1.29	1.29	0.064	-	20.01
Selection model								
10. Heckit	R&D intensity	1308	1319	0.014	0.011	0.010	ρ = 0.002	1.30
(ML)	R&D expenditure ***	1308	1319	2618.1	716.7	755.1	ρ = -0.1*	3.47
	R&D per employee ***	1308	1319	2.338	1.630	0.658	ρ = -0.05*	3.55
DID								
11. Difference-in-differences	R&D intensity ***	113	248	0.013	0.010	0.003	$\Delta\eta$ = 0.003**	4.35
	R&D expenditure	113	248	229	85	177	$\Delta\eta$ = 135	1.30
	R&D per employee ***	113	248	1.55	1.46	0.65	$\Delta\eta$ = 0.45	2.39
Average on Matching								
Average	R&D intensity ***	891	1196	0.014	0.011	0.002	2.67	5.45
(over 1-7)	R&D expenditure ***	891	1196	885.1	242.3	419.7	4.62	2.78
	R&D per employee ***	891	1196	1.935	1.349	0.351	2.49	5.64

Notes: Δ = Ratio of the value of the endogenous variable calculated on treated to that calculated on control units; ρ = correlation between the outcome and the selection equations; $\Delta\eta$ = inter-temporal fixed-effect coefficient; * = significant at 10 %; ** = significant at 5 %; *** = significant at 1 %. The ATET for the population is calculated by a proportional rule based on sampling weights.

the number of treated (sample size) and the counterfactual precision: if we want a better counterfactual sample ("good twins") we have to renounce to a larger sample size. This opens up the related problem of singling out the *best* Matching approach to apply and we will come back to this point in the concluding section.

We limit our attention to the sample results. Within the Matching methods, the difference in the ATET for the R&D expenditure ranges from a minimum of 340 in the Kernel method to a maximum of 1,351 thousand euros in the Radius 2. The greatest value is anyhow reached by the Heckit (2,618 thousand euros), with a very high level compared with all the other approaches, while the lowest one is reached by DID with a significant 229 thousand euros. On average from Matching methods we get 885 thousand euros with a Δ ratio equal to 4.62: it means that if a generic control unit does 1 thousand euros of R&D expenditure a matched treated does 4.62 thousand euros. For each target-variable, Table 7 shows the ranking of methods according to their distance from the average level of the ATET over the ten methods applied. For the R&D expenditure, the average additionality over the ten methods is 1,017 thousand euros, so that: DID, Kernel and Radius 3 are the most pessimistic estimators of the ATET for this variable, while Heckit and Stratification the most optimistic ones. Quite surprisingly, Control-function is fairly in line with the overall average (only a 10% of divergence) although the most precise approach seems to be in this sense the 3-nearest neighbour Matching (with only 1.5% of divergence in absolute terms). Nevertheless, the most striking result comes from the coefficient of variation of Table 7, calculated for the ATET distribution of each target-variable over the ten methods: this indicator reaches for the R&D expenditure the highest level of 0.67, that is about six times that reached by the coefficient of variation calculated on the R&D intensity and on the R&D per employee (0.13). This is a remarkable point as it undermines quite seriously the robustness of results on the R&D expenditure additionality, especially when just one single method is used. We deem it a significant finding of our application.

It is also of worth to briefly comment the results on the other target-variables. The difference in ATET for R&D intensity among the seven Matching procedures is negligible since they give very similar results. Only for Radius with a calliper of 0.0001 there is a little lower value (0.010) compared to the Matching mean (0.014) where, according to the Δ ratio, the R&D intensity level of treated firms is 2.67 times that of the control units.

The ATET on the R&D per employee produces results similar to the R&D intensity. Indeed, no significant differences can be found among the various methods: the average of Matching methods is around 1.9, with a Δ ratio of 2.49, while the

Table 7. Ranking of methods according to their distance from the average level of the ATET

	R&D expenditure		R&D intensity		R&D per employee	
Method	ATET	Percentage over the average	ATET	Percentage over the average	ATET	Percentage over the average
1. Stratification	1245	22.46	0.016	15.942	2.17	12.39
2. One-to-One NN	1038	2.10	0.015	8.696	1.94	0.48
3. 3-NN	1001	-1.54	0.016	15.942	2.06	6.70
4. Kernel	340	-66.56	0.013	-5.797	1.82	-5.73
5. Radius 1	693	-31.84	0.013	-5.797	1.93	-0.04
6. Radius 2	1351	32.88	0.01	-27.536	1.51	-21.79
7. Radius 3	525	-48.36	0.014	1.449	2.10	8.61
8. Control-function	1127	10.85	0.014	1.449	1.89	-2.11
9. Heckit	2618	157.50	0.014	1.449	2.34	21.20
10. DID	229	-77.48	0.013	-5.797	1.55	-19.72
Average	1017		0.014		1.93	
St. error	680.35		0.002		0.26	
Coef. of variation	0.67		0.13		0.13	

Note: Coefficient of variation = standard error / average.

Heckit has a value of 2.33, and DID of 1.55. Results from Control-function (based on an OLS regression) are quite identical to those of the average Matching methods: the ATET for R&D intensity, for example, is equal to 0.014 as before. It seems that the OLS bias due to the linearity and lack of similarity in the control group is quite negligible in our application (a result shared by other work using CIS data and in particular, Aerts and Czarnitzki 2004). The Heckit model conveys similar results on R&D intensity while, as said before, both OLS regressions and Heckit show substantial differences, when compared with Matching, on R&D expenditure. For the Heckman selection model the ρ is quite low for all target-variables: generally speaking, it means that the unobservable factors influencing the selection-into-program equation are little correlated with the unobservable factors influencing the firm R&D decision/self-selection.[21] As for DID, a method able to deal with selection

[21] It is noteworthy to observe that in this paper we perform a little spurious use of the Heckit as we assume, for the sake of comparison with other methods, that the covariates determining the selection into program are the same as those feeding into the firm R&D behavior (see section III for details). In a linear setting this assumption would generate no identification of parameters as no exclusion restrictions (based on theoretical statements) are defined. But the Heckit provides estimates also in this case thanks

on unobservables, we get - as sketched before - substantial differences compared to other methods only in the ATET for the R&D expenditure that reaches in this case its lowest magnitude (229). If we assume DID to be the most consistent estimation approach, this leads to the conclusion the other methods generally overestimate additionality for R&D expenditure, an aspect due essentially to the fact that they heavily overlook both idiosyncratic heterogeneity (for example, innate ability) and time fixed-effect.

Table 8, finally, presents results according to different sub-groups of firms. We only use the R&D intensity as outcome variable and only OLS, Stratification and One-to-one nearest-neighbour matching as estimation methods. This choice reflects the idea of comparing the three matching approaches most commonly applied in the literature. As results on all these methods are strongly concordant, we concentrate the attention on figures from the One-to-one nearest neighbour matching. We start with groups that show a total crowding-out effect: low knowledge intensive services (LKIS), and very small firms (10-19 employees).[22] Very small firms can be considered an example of asset constrained firms and, since they are generally not engaged in formal R&D activities, they could have used R&D incentives as substitutes for other type of investments.

We now look at groups showing no crowding-out effect. We found that South-Italy and Centre-Italy are significant, but only at 10 or 5% depending on the methods; all the other groups present a value of the ATET significant at 1%.[23] South-Italy shows the greatest level of the Δ ratio (3.90) among the geographical groups, even if with a lower significance, given probably the lower number of observations (only 88 matched treated units). Looking at the Δ ratio among the more significant groups, the Medium-Small seized firms (20-49 employees) have the greatest value of 4.57; in the second position we find Large firms (>250 employees) with a value of 4.25 and, in the third position, the knowledge intensive services (KIS) with a value of 4.15.

---

to its non-linear structure. Nevertheless, by avoiding exclusion restrictions the Heckit is used a little spuriously as both coefficients magnitudes and standard errors are affected by this assumption (see Cameron-Trivedi 2009: 543-547). It is for this reason that we prefer not to stress too much results from Heckit.

[22] See Appendix B for the definition of macro-sectors. At the level of specific sectors, there is full crowding out in motor vehicles, which could reflect the Fiat Group crisis in the early 2000s, derived by a previous reduction on R&D investments (results available from authors).

[23] The effects in specific sectors are significant at 5/10% in Chemicals, and at 1% in Mechanics (results available from the authors).

Table 8. Estimation of the Average Treatment Effect on Treated (ATET) for different geographical, sectoral and dimensional patterns

Group	Matching Procedure	Number of treated	Number of controls	ATET (Sample)	ATET (Population)	Δ	S.E.	t-value
Geographical pattern								
North	Stratification ***	1059	3650	0.012	0.009	-	0.002	7.59
	One-to-one N-N ***	1061	596	0.0118	0.009	3.22	0.0016	7.34
	OLS ***	433	615	0.0111	0.013	-	0.002	5.84
Centre	Stratification **	229	775	0.018	0.014	-	0.008	2.22
	One-to-one N-N *	229	116	0.0169	0.013	2.01	0.009	1.72
	OLS **	122	153	0.023	0.029	-	0.010	2.19
South	Stratification **	88	460	0.028	0.021	-	0.014	2.08
	One-to-one N-N **	89	46	0.029	0.022	3.90	0.014	1.99
	OLS	48	45	0.025	0.024	-	0.023	1.08
Sectoral pattern								
H-T	Stratification ***	155	177	0.03	0.023	-	0.008	3.92
	One-to-one N-N ***	157	60	0.031	0.023	2.82	0.007	4.06
	OLS ***	93	69	0.041	0.031	-	0.013	3.05
M-H-T	Stratification ***	350	555	0.008	0.006	-	0.001	6.31
	One-to-one N-N ***	350	150	0.009	0.007	2.80	0.001	6.27
	OLS ***	200	187	0.007	0.007		0.001	4.16
M-L-T	Stratification ***	327	954	0.007	0.005	-	0.001	5.23
	One-to-one N-N ***	328	181	0.0078	0.006	3.33	0.001	5.91
	OLS ***	122	132	0.006	0.004	-	0.001	4.58
L-T	Stratification ***	264	1021	0.007	0.005	-	0.002	3.24
	One-to-one N-N ***	264	158	0.006	0.005	2.5	0.002	2.85
	OLS ***	93	156	0.011	0.024	-	0.003	2.88

Table 8. (continued) Estimation of the Average Treatment Effect on Treated (ATET) for different geographical, sectoral and dimensional patterns

Group	Matching Procedure	Number of treated	Number of controls	ATET (Sample)	ATET (Population)	Δ	S.E.	t-value
Sectoral pattern								
KIS	Stratification ***	95	405	0.085	0.064	-	0.02	4.25
	One-to-one N-N ***	96	53	0.06	0.045	4.15	0.02	2.94
	OLS **	59	98	0.043	0.035		0.02	2.13
LKIS	Stratification	91	983	0.001	0.001	-	0.00	1.16
	One-to-one N-N	91	72	0.0003	0.000	1.38	0.001	0.34
	OLS	36	171	0.0006	0.001		0.001	0.46
Dimensional pattern								
Small	Stratification	103	598	0.016	0.012	-	0.016	0.94
	One-to-one N-N	103	61	0.005	0.004	1.18	0.014	0.40
	OLS	17	47	0.011	0.035	-	0.035	0.32
Medium-Small	Stratification ***	247	1075	0.025	0.019	-	0.006	4.15
	One-to-one N-N ***	247	143	0.025	0.019	4.57	0.006	4.00
	OLS ***	55	119	0.041	0.012		0.014	2.79
Medium-Large	Stratification ***	594	1674	0.012	0.009	-	0.003	3.94
	One-to-one N-N ***	594	295	0.011	0.008	2.48	0.003	3.28
	OLS ***	243	375	0.013	0.009	-	0.004	3.27
Large	Stratification ***	268	456	0.014	0.011	-	0.003	5.28
	One-to-one N-N ***	270	116	0.013	0.010	4.25	0.002	5.02
	OLS ***	288	272	0.013	0.013	-	0.002	4.64

Notes: in the OLS the sample size reduces as we use always the entire set of covariates so that more missing values are computed.
Note: Δ = Ratio of the value of the endogenous variable calculated on treated to that calculated on control units; * = significant at 10 %; ** = significant at 5 %,;*** = significant at 1 %.

As for sectors, the most significant Δ ratio, after KIS, is for Medium-Low-Tech manufacturing firms with a value of 3.33, followed by the High-Tech and Medium-High-Tech manufacturing with the same Δ ratio of around 2.8; finally, Medium-Large firms (50-249 employees) have a lower value of 2.48.

In summary, although the statistical significance seems concordant, great differences emerge among subgroups of firms in terms of the magnitude of the ATET. It means that the effect of R&D and innovation policies is strongly heterogonous and that, even in this case, great non-linearity could be at work. Again, relying only on an average result could be limiting and quite misleading for evaluation purposes.

VI. Conclusions

This paper acknowledges the need of giving greater attention to the robustness of R&D policy evaluation results. Our exercise shows that R&D expenditure seems to be particularly affected by the choice of the method, more than variables expressed as ratios (R&D intensity, R&D per employee). The level of R&D expenditure is central in research and innovation policy programs: it can be taken as a strategic target for the policymakers, as their main objective is to enlarge the national level of industrial R&D and to increase social welfare. Our evaluation exercise states that, when the target-variable is expressed as a ratio, the results are not sensitive to the econometric methods and are quite close in sign, significance and magnitude. In the case of R&D expenditure, conversely, differences appear also within approaches, such as Matching, based on the same identification assumptions. This finding opens the problem of how to choose the more appropriate result for R&D expenditure and understand what are the underlying factors leading to these heterogeneous outcomes.

As for the last point, we know that –in using various methods– differences in results depend on two related aspects: (i) each method makes use of a different sample; (2) each method applies a different formula (depending on the method's *identification* assumptions). It is fairly intuitive that, when a target-variable is particularly volatile across observations, it could be more likely that it is also more volatile in getting results across methods, provided that their application requires diverse samples as in our case:[24] we calculated (not reported) that the coefficient of variation across observations of R&D expenditure is about five times that of R&D intensity, while Table 7 shows that this variability across methods (i.e., refereed

[24] We are in debt with one of the referees for raising this point that we had previously overlooked.

to the ATET) is about six times: it proves a strong concordance between these two findings, that cannot be neglected.[25]

As for which result should be reported, we have to notice that while previous discussion can explain –at least to a certain extent– why R&D expenditure leads to more volatile results, we cannot say nothing about which should be the best result to take and thus the best method to apply. Consider the case of the choice among Matching methods: it is not possible to determine *ex ante* which is the best Matching because of the *trade-off* between sample size on one hand and counterfactual precision (i.e., availability of good twins) on the other. This is quite well evident in the case of Radius Matching where one would be tempted to say that the best is to use a Matching procedure leaving her/him with the highest sample size; nevertheless, this can be obtained only by assuming a large calliper, that is, by considering in the counterfactual comparison also non-treated subjects that could be (potentially) strongly different from the treated ones. On the contrary, one could say that the best is to use Matching procedure leaving her/him with very similar (non-treated) twins, but it requires assuming a very short calliper, thus reducing sensibly the sample size. Actually, since estimation *efficiency* depends both on sample size and counterfactual precision, then –without knowing the true Data Generating Process (DGP) as is always the case with real datasets– only some rule of thumb and/or common sense can guide the choice of the "right" method. We suggest, for instance, to take an average, but one could also consider the method better balancing the two dimensions (again, sample size and counterfactual precision) when possible. This is, of course, case-specific and depends –at least partly– on some analyst's choice.[26]

Relevant differences arise also in subsamples. It means that going beyond the average result of the ATET can produce a substantial rethinking of the way in which the policy considered acted. It confirms, furthermore, that firm heterogeneity severely

[25] For R&D expenditure, we performed also a comparison of Matching, options 2 and 3, with OLS and Heckit, by keeping the *same* sample. Apart from results from Heckit - proving to be quite large as already seen in Table 6 - results from OLS and Matching were quite close, thus showing in this case the relative importance of the choice of sample for our results.

[26] Observe in fact that, whereas all Matching methods are *consistent* estimators under "selection on observables", they could have –according to the specific data characteristics– a different *efficiency* (see Zaho 2004). Efficiency, in turn, depends on sample size and counterfactual goodness: for the same sample size, the Matching procedure producing the best set of twins will be that closer to the true value of ATET. The problem emerges in practice since one cannot obtain the same counterfactual precision along with the same sample size.

affects the policy impact at various levels of disaggregation of treated units. Therefore, relying on a comparison of various methods (rather than on a single value from a single econometric technique and target-variable) could be a more fair and correct practise.

A potential methodological improvement of this work could be that of including, within the comparison of methods, also the IV approach. Of course, going into this direction requires having access to at least one variable correlated with the R&D policy, but (directly) uncorrelated with firm R&D and self-selection decisions. It is well known that finding such a variable in an R&D context is sometimes difficult and that –more importantly– in a (usual) just-identified setting it is not possible to test for its exogeneity.[27] We think that a Monte Carlo experiment could be more suited to assess the robustness of IV, and we plan to go into this research direction in our next paper on this subject.

As for the results from our sample of Italian firms, our analysis predicts the absence of a full crowding-out of the private R&D effort, both on average and on the majority of firm subgroups. Nevertheless, we can conclude nothing certain about the actual additionality reached. Indeed, since CIS data do not provide the level of the subsidy, but only a binary indicator of it, we cannot know if a positive increment of firm R&D has been greater, equal or lower than the actual subsidy received, an aspect shared by all the evaluation exercises using CIS or a binary treatment variable. Anyway, we deem our results to provide a good indication - at least in terms of direction - of the actual causal effect in question.

Appendix

A. One-to-one nearest-neighbour Matching

In this Appendix we report: (i) a representative test of the *balancing property* for Matching estimation using the Becker and Ichino (2002) algorithm; (ii) a graphical representations of the propensity scores distribution of treated and untreated firms *before* and *after* the One-to-one nearest-neighbour Matching.

As for point (i), the balancing property is satisfied by splitting the sample into six blocks. Only representatively, Table A1 shows the difference-in-mean t-test for the control variables and for the propensity score in block 1.

[27] In his IV application to the SBIR program in the U.S., Lichtenberg (1988) uses the "value of competitive contracts that were *potentially* awardable to each firm" as instrumental variable for the R&D subsidy (see also Wallsten 2000). We do not have access to such information within CIS data.

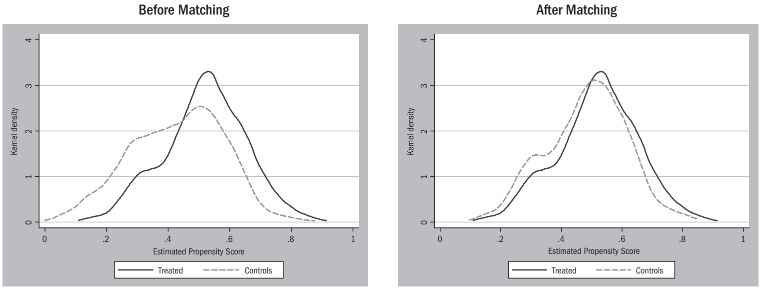
As for point (ii), Figure A1 shows the improvement reached by the One-to-one nearest-neighbour Matching by comparing the distribution of the propensity score before and after this procedure. After matching, treated and non- treated units seem to be drawn from the same data generating process.

Table A1. Test of the balancing property for the model adopted

Block 1	Group	Number	Mean	Difference	P-value
Propensity score	0	187	0.132	-0.004	0.612
	1	39	0.137	-	-
EMP 98	0	187	171.59	-119.9962	0.203
	1	39	291.59		
EMPSKILL 00	0	187	0.125	-0.033724	0.263
	1	39	0.159		
EXPINT 98	0	187	0.083	-0.0813	0.036
	1	39	0.16		
CAPINT 98	0	187	152.06	73.37	0.752
	1	39	78.68		
CASHINT 98	0	187	18.04	7.25	0.620
	1	39	10.79		
DEBTINT 98	0	187	0.66	0.055	0.141
	1	39	0.6		
KNOWLEDGE 98	0	187	8.65E+07	-1.84E+08	0.015
	1	39	2.70E+08		
AGE	0	187	0.06	-0.043	0.320
	1	39	0.10		
FOREIGN	0	187	0.598	-0.195	0.020
	1	39	0.794		

Note: The algorithm used rejects equality in the means of the variables just at a level of significance smaller than 1% (see Becker and Ichino 2002).

Figure A1. Propensity scores distribution of treated and untreated firms before and after the One-to-one nearest neighbour Matching



B . Definition of macro-sectors

Table A2. Definition of the macro-sectors employed in the subgroups' analysis

High-technology manufacturing	Medium-high-technology manufacturing	Medium-low-technology manufacturing
Pharmaceuticals	Chemicals	Coke, refined petroleum products and nuclear fuel
Office machinery and computers	Chemicals, excluding pharmaceuticals	Rubber and plastic products
Radio, TV and communication equipment	Machinery and equipment	Other non-metallic mineral products
Instrument engineering	Electrical machinery	Basic metals
Manufacture of aircraft and spacecraft	Motor vehicles	Fabricated metal products
	Other transport equipment, excluding ships and aerospace	Building and repairing of ships and boats
	Other transport equipment	
Low-technology manufacturing	**Knowledge-intensive services**	**Less knowledge-intensive services**
Food and beverages	Water transport	Motor trade
Tobacco products	Air transport	Wholesale trade
Textiles	Post and telecommunications	Retail trade
Clothing	Computer and related activities	Hotels and restaurants
Leather products	Other business activities	Land transport
Wood products		Auxiliary transport activities
Pulp and paper products		
Publishing and printing		
Manufacturing n.e.c.		
Recycling		

References

Aerts, Kris, and Dirk Czanitzki (2004), Using innovation survey data to evaluate R&D policy: The case of Belgium, Discussion Paper 04-55, ZEW.

Almus, Matthias, and Dirk Czarnitzki (2003), The effects of public R&D subsidies on firms' innovation activities: The case of Eastern Germany, *Journal of Business and Economic Statistics* **21**: 226-236.

Arvanitis, Spyros, Laurent Donzé, and Nora Sydow (2010), Impact of Swiss technology policy on firm innovation performance: An evaluation based on a matching approach, *Science and Public Policy* **37**: 63-78.

Becker, Sascha O., and Andrea Ichino (2002), Estimation of average treatment effects based on propensity scores, *Stata Journal* **2**: 358–377.

Bérubé, Charles, and Pierre Mohnen (2007), Are firms that received R&D subsidies more innovative?, Working Paper Series 015, UNU-MERIT.

Blundell, Richard, and Monica Costa Dias (2002), Alternative approaches to evaluation in empirical microeconomics, *Portuguese Economic Journal* **1**: 91-115.

Busom, Isabel (2000), An empirical evaluation of the effects of R&D subsidies, *Economics of Innovation and New Technology* **9**. 111–148.

Caliendo, Marco, and Sabine Kopeinig (2008), Some practical guidance for the implementation of propensity score matching, *Journal of Economic Surveys* **22**: 31-72.

Cameron, A. Colin, and Pravin K. Trivedi (2005), *Microeconometrics: Methods and applications*, New York, NY, Cambridge University Press.

Cameron, A. Colin, and Pravin K. Trivedi (2009), *Microeconometrics using Stata*, College Station, TX, Stata Press.

Cerulli, Giovanni (2010), Modelling and measuring the effect of public subsidies on business R&D: a critical review of the econometric literature, *Economic Record* **86**: 421-449.

Cobb-Clark, Deborah A., and Thomas Crossley (2003), Econometrics for evaluations: An introduction to recent developments, *Economic Record* **79**: 491 - 511.

David, Paul A., Bronwyn H. Hall, and Andrew A. Toole (2000), Is public R&D a complement or substitute for private R&D? A review of the econometric evidence, *Research Policy*: **29**: 497–529.

Dehejia, Rajeev H., and Sadek Wahba (2002), Propensity score-matching methods for nonexperimental causal studies, *The Review of Economics and Statistics* **84**: 151-161.

Duguet, Emmanuel (2004), Are R&D subsidies a substitute or a complement to privately funded R&D? Evidence from France using propensity score methods for non-experimental data, Public Economics 0411007, EconWPA.

Görg, Holger, and Eric A. Strobl (2007), The effect of R&D subsidies on private R&D, *Economica* **74**: 215-234.

Heckman, James J. (1978), Dummy endogenous variables in a simultaneous equation system, *Econometrica* **46**: 931-59.

Heckman, James J. (2001), Micro data, heterogeneity, and the evaluation of public policy: Nobel Lecture, *Journal of Political Economy* **109**: 673-748.

Heckman, James J., Hidehiko Ichimura, and Petra Todd (1998), Matching as an econometric evaluation estimator, *Review of Economic Studies* **65**: 261-94.

Hussinger, Katrin (2008), R&D and subsidies at the firm level: An application of parametric and semi-parametric two-step selection models, *Journal of Applied Econometrics* **23**: 729-747.

ISTAT (2004), *Statistiche sull'innovazione nelle imprese. Anni 1998-2000*, Collana Informazioni.

Kaiser, Ulrich (2004), Private R&D and public R&D subsidies: Microeconometric evidence from Denmark, Discussion paper 2004-19, CEBR.

Klette, Tor J., and Møen Jarle (1998), R&D investment responses to R&D subsidies: A theoretical analysis and a microeconometric study, Mimeo, Presented at the NBER Summer Institute.

Klette, Tor J., Jarle Møen, and Zvi Griliches (2000), Do subsidies to commercial R&D reduce market failures? Microeconometric Evaluation Studies, *Research Policy* **29**: 471–495.

Lach, Saul (2002), Do R&D subsidies stimulate or displace private R&D? Evidence from Israel, *Journal of Industrial Economics* **50**: 369-90.

Lichtenberg, Frank R. (1987), The effect of government funding on private industrial research and development: a re-assessment, *The Journal of Industrial Economics* **36**: 97–104.

Lichtenberg, Frank R. (1988), The private R&D investment response to federal design and technical competitions, *American Economic Review* **78**: 550–559.

Lööf, Hans, and Almas Heshmati (2005), The impact of public funds on private R&D investment: New evidence from a firm level innovation study, in A. Heshmati, Y-B. Sohn and Y-R. Kim, eds., *Commercialization and Transfer of Technology: Major Country Case Studies*, New York, NY, Nova Science Publishers.

Maddala G S. (1983), Limited dependent and qualitative variables in econometrics. Cambridge, England: Cambridge University Press.

Rosenbaum, Paul R., and Donald B. Rubin (1983), The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**: 41–55.

Rubin, Donald B. (1977), Assignment to treatment group on the basis of a covariate, *Journal of Educational Statistics* **2**: 1-26.

Wallsten, Scott J. (2000), The effects of government–industry R&D programs on private R&D: The case of the small business innovation research program, *The RAND Journal of Economics* **31**: 82-100.

Wooldridge, Jeffrey M. (2002), *Econometric Analysis of cross section and panel data*, Cambridge, MA, MIT Press.

Zhao, Zhong (2004), Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence, *Review of Economics and Statistics* **86**: 91-107.