UNIVERSITÀ POLITECNICA DELLE MARCHE

DIPARTIMENTO DI SCIENZE ECONOMICHE E SOCIALI

To Match, Not To Match, How To Match: Estimating the Farm-Level Impact of the CAP-First Pillar Reform

(Or: How to Apply Treatment-Effect Econometrics when the Real World is a Mess)

ROBERTO ESPOSTI

QUADERNO DI RICERCA n. 403
ISSN: 2279-9575

June 2014

Collana curata da:
Massimo Tamberi

To match, not to match, how to match: Estimating the farm-level impact of the CAP-first pillar reform

(or: How to Apply Treatment-Effect Econometrics when the Real World is a Mess)

Roberto Esposti♦

Department of Economics and Social Sciences – Università Politecnica delle Marche
Piazzale Martelli 8 – 60121 Ancona (Italy)
e-mail: r.esposti at univpm.it; phone: +39 071 2207119

Abstract

This paper aims at evaluating the impact of the 2003/2005 CAP reform on farm production choices. The outcome of "market orientation" is measured by considering both the short-term production choices and the long-term investment decisions. The Treatment Effect (TE) is estimated through alternative approaches due to the difficulties encountered in finding appropriate counterfactuals. Different versions of the Propensity Score Matching (PSM) estimators, the Difference-In-Difference (DID) estimate, and alternative multiple/continuous TEs estimates, based on the Generalized Propensity Score (GPS), are performed, their statistical robustness assessed and results compared. Results show that the 2003/2005 reform of the first pillar of the CAP actually had an impact more in (re)orienting short-term farm production choices than investment decisions and this effect is significantly more evident for farms with a limited contribution of the CAP on their own Gross Production Value.

***Keywords**: treatment effects, common agricultural policy, farm production choices, matching*

♦ Department of Economics and Social Sciences – Università Politecnica delle Marche
Piazzale Martelli 8 – 60121 Ancona (Italy)
e-mail: r.esposti at univpm.it; phone: +39 071 2207119

CONTENTS

To match, not to match, how to match: Estimating the farm-level impact of the CAP-first pillar reform

(or: How to Apply Treatment-Effect Econometrics when the Real World is a Mess)

1. INTRODUCTION: OBJECTIVES OF THE PAPER

In the last decade the empirical research on policy evaluation has paid increasing attention to the so-called Treatment-Effects (TE) literature (Imbens and Wooldridge, 2009). A rich toolkit allowing the estimation of these TEs under different and complex circumstances has progressively taken form. Nonetheless, though equipped with this powerful toolkit, practitioners often encounters serious problems in adapting it to the peculiar real-world circumstances into which policy measures have been actually implemented. As a consequence, the recent empirical literature has also focused on assessing the robustness of the estimated policy impacts to identify those that are strongly dependent on the specific limitations of the adopted methods with respect to the real context under study (Chabé-Ferret, 2010).

In the case of the quantitative evaluation of Common Agricultural Policy (CAP) measures and reforms, these two apparently contrasting tendencies of an increasingly powerful toolkit and the need of robustness of results clearly emerge. On the one hand, the growing interest in such approaches is evident for the second pillar of the CAP, whose measures seem particularly suitable for these empirical methodologies (European Commission, 2006; EENRD, 2010; Pufahl and Weiss, 2009; Lukesch and Schuh, 2010; Salvioni and Sciulli, 2011; Michalek, 2012; Chabé-Ferret and Subervie, 2013). On the other hand, it is widely agreed that the same does not hold true for the first pillar of the CAP (still the largest part of CAP budget) as the way it is designed and delivered makes these methods not particularly helpful, or simply useless, to achieve a proper impact evaluation (Esposti, 2011a).

The objective of this paper is to critically re-consider the apparent infeasibility of the TE econometrics' toolkit in the case of the First Pillar of the CAP by pursuing the evaluation of the impact of its 2005 Reform (FPR henceforth).[1] The first strategy consists in adopting the largely prevalent matching methodology in the literature (the Propensity Score Matching, PSM) enriched by a set of testing procedures and robustness checks (Becker and Ichino, 2002; Abedie et al., 2004; Becker and Caliendo, 2007; Nannicini, 2007; Nichols, 2007) to assess to what extent the approach is suitable for the case under study. The second strategy is not based, in its simplest form, on matching because the counterfactuals simply are the beneficiary farms themselves observed before the policy intervention. It is the Differences-In-Differences (DID) approach that has been already proposed for

[1] Though approved in 2003, the FPR implementation actually started in 2005. In addition, for some productions (cotton, hop, olive oil, tobacco, sugar, fruits and vegetables, wine) the decoupling of the support formerly delivered through the individual OCMs was actually approved and accomplished in the following years (2004-2008) (OECD, 2011, pp.64-65). For the sake of simplicity, we refer here to all these reform steps as the "2005 First Pillar Reform" or FPR (see also sections 3.2.1 and 4).

the evaluation of single measures of the second pillar of the CAP and is often combined with the PSM (Salvioni and Sciulli, 2011; Chabé-Ferret and Subervie, 2013).
The third strategy does not look for non-treated units (or counterfactuals) but takes advantage of the multivalued nature of the first pillar support. It is based on the concept of Generalized Propensity Score Matching (GPSM) where matching is performed among farms with different treatment intensity. This seems particularly appropriate for the first pillar of the CAP where most farms receive the treatment (the payment) but its intensity largely varies. In recent years, several alternative estimators have been developed for the case of multivalued or continuous treatments[2] (Hirano and Imbens, 2004; Bia and Mattei, 2008; Cattaneo, 2010; Cattaneo et al., 2013). They will be implemented here to compare results and assess their robustness.
This paper applies these three empirical strategies in sequence to compare their results and assess whether some robust evidence emerges from this comparison. They are applied to a balanced panel of 6542 Italian FADN farms observed over years 2003-2007. Therefore, these farms are observed before and after the FPR, whose implementation, in Italy, actually started in 2005, and the period under analysis is symmetric with respect to the treatment. Over these years, about 80% of these farms received payments from the first pillar; therefore, only 20% can be considered as non-treated units. This panel sample seems appropriate to apply all the three abovementioned approaches by demonstrating, at the same time, their strengths and limitations.

2. THE CONCEPTUAL BACKGROUND: WHY EVALUATING THE IMPACT OF THE FPR IS DIFFICULT?

In 2003 the first pillar of the CAP underwent what can be considered the most radical reform of its half-century history; the so-called Fischler Reform (Sorrentino et al., 2011). With decoupling the FPR substantially changed the way support is delivered to farms. One of the purposes of decoupling was (re)orientation to market, that is, to eliminate the distorting effect on production decisions of the precedent coupled support and leave the farmers free to produce what they consider more profitable in the market. This is explicitly mentioned in many EU Commission documents and, therefore, it can be considered as the main declared objective of the reform itself: “the next movement towards market orientation for the European agricultural sector came in 2003, when a major overhaul of the CAP was undertaken. […] The current decoupled direct payment […] ensures that farmers respond to market signals while providing income support” (European Commission, 2011, p. 6; see also OECD, 2011, pp. 140, 184).

[2] Some texts refer to this case as “multilevel treatment” while with “multivalued treatment” it is designated the case of more than one treatment simultaneously administered to the same units (StataCorp, 2013). Here, following the prevailing terminology in the TE literature (Imbens and Wooldridge, 2009; Cattaneo, 2010) the former case is called “multivalued treatment”, the latter “multiple treatment”.

Therefore, an *ex post* evaluation of the effectiveness of the FPR should assess whether and to what extent the decoupling of first pillar's support really oriented farmers to market. Despite the wide empirical literature produced on the impact of the FPR at both the farm and several aggregate levels (OECD, 2011; Sorrentino et al., 2011), such evaluation has never been performed using the rich toolkit of TE econometrics, mainly because it is considered unsuited, or of too difficult adaptation, to this case. Four major issues emerge in adopting this methodological approach to the FPR.

The first possible issue of the FPR when considered as a "treatment" is that it should be associated to a clearly recognized *ex-ante* target, i.e. an outcome variable with respect to which the treatment/policy impact can be evaluated. Although we may agree that the main objective of the FPR was to reorient farms' choices towards the market, it remains questionable whether and how this objective can be univocally expressed by a single variable, a target/outcome variable, that eventually expresses the effect of the reform. Such outcome variable should measure to what extent farms have changed their production orientation (e.g., their production mix) after the FPR, but computing such variable isn't easy for production units that naturally are multi-output and, therefore, systematically change their production mix on a yearly basis. In these cases, it could be preferable to identify and estimate the TE acknowledging the presence of *multiple outcomes*. In fact, the impact evaluation of the FPR is often carried out through methods and measures that are only aggregate or indirect or partial effects of the treatment (OECD, 2011).[3]

A second issue is that the treatment under study must be clearly identified and observable. Apparently, the FPR itself has substantially increased the clarity and the identifiability of the treatment associated to the first pillar of the CAP. Many and heterogeneous coupled payments (associated to about 30 Common Market Organizations, CMOs) have been transformed into a unique and decoupled Single Farm Payment (SFP) directly delivered to the farmer (European Commission, 2011; OECD, 2011). This makes the FPR treatment easily measurable as the amount of money that, at the farm level, have been transformed from several coupled payments into a decoupled SFP. However, still a complication arises in considering the FPR as a "treatment" because, unlike other typical policy program evaluations (e.g., job training programs), the treatment associated to the FPR is not a binary treatment, (i.e., expressed as a dichotomous variable: 1=treated, 0=not treated). The FPR is a continuous treatment (*multivalued treatment*; Cattaneo, 2010) where only few units (farms) are not treated, while most units are treated but with different intensity. Therefore, in the case of the FPR the TE reasonably depends on the treatment intensity and not only on being treated or not.[4] Nonetheless, the multivalued nature of the treatment, rather than being a limitation, actually represents the key opportunity to identify and estimate the TE of the FPR.

A third issue is that farms receiving the SFP may also undergo other treatments that can affect, directly or indirectly, production choices and market orientation, thus confusing the impact of the FPR. This is

[3] See section 5.2 for more details in this respect.
[4] See section 5.1 for more details.

the case of several second pillar's measures; for instance, those supporting physical investments and product quality and innovation. Actually, the second pillar support was never coupled to particular commodities and it apparently has nothing to do with the aim of 'market orientation' that drove the FPR. Still, though these two policies are currently implemented and delivered almost independently, following different selection procedures and aiming at different objectives, we can not exclude, in principle, that the outcome of one treatment depends on the presence and intensity of the other, therefore making the identification and estimation of the single TE more difficult (the problem of *multiple treatments*) (Frölich, 2004).

The final and more important issue, however, really represents the key limitation that apparently prevents the application of TE methods to the evaluation of the FPR. To apply the TE logic counterfactual observations must exist; that is, observations where the outcome variable(s) is (are) observed without the treatment (*control group* or *counterfactuals*).[5] In the case of the FPR, however, finding a proper strategy to identify counterfactuals and compare them to treated units represents an often unsolved research challenge. On the one hand, the non-treated units (that is, farms receiving no support under the first pillar of the CAP and, therefore, not involved by the FPR) are rare. On the other hand, they are not treated just because of their peculiar production and managerial choices. Even though a non-treated sample can be observed, it can be hardly considered a proper counterfactual sample because of this peculiarity, that is, those unobserved characteristics that affect, at the same time, the outcome and the treatment assignment. In the TE jargon, the specificity of the FPR as a treatment makes almost impossible to get rid of the *selection-on-unobservables bias* (Esposti, 2011a).

When dealing with the FPR, all these issues simultaneously arise. In such uncomfortable and unconventional condition (multiple outcomes, multiple and multivalued treatments, no natural counterfactuals), it should not surprise that the TE logic and econometrics has been considered unsuited for the evaluation of the first pillar of the CAP and of the FPR, in particular.

3. THE METHODOLOGICAL APPROACH: THREE ALTERNATIVE SOLUTIONS TO IDENTIFY AND ESTIMATE THE TE

From the discussion above it emerges that the selection of appropriate counterfactual observations depends on the farm sample under observation. In general term, a large farm panel is needed both in the cross-sectional dimension, to include enough not treated farms, and in the time dimension, to include enough pre and post-treatment observations. This panel sample must also include all the relevant variables to indentify and estimate the ATT: univocal outcome variables (the target of the treatment) (Y); the variable expressing the presence (if binary) or the intensity (if multivalued or continuous) of the treatment (T); the set of all other socio-economic, structural, production variables

[5] Assessing treatment effects within experimental or quasi-experimental situations presents substantially different methodological issues and solutions (Duflo et al., 2006).

that affect the outcome variable and its dynamics beyond and regardless the treatment (**X**). If these conditions are met,[6] the TE of the FPR can be identified and estimated following three alternative strategies, each with own pros and cons with respect to the four abovementioned major issues.

3.1. The Propensity Score Matching (PSM)

The first, and currently very popular, strategy consists in matching the most similar treated and non-treated units and then compare their outcome, *Y*. Similarity is assessed by looking at the set of exogenous variable (covariates), **X**, that may affect the outcome beside the treatment. The use of matching in program and policy evaluation has increased in the last due decades mostly because of a strongly simplifying methodology (Rosenbaum and Rubin, 1983). It consists in performing the matching not on the basis of the multivariate set **X** but only on the basis of a scalar variable, the *Propensity Score* (PS), representing the estimated probability of a given unit to be assigned the treatment conditional on **X**. Matching based on the propensity score (the Propensity Score Matching, PSM) strongly facilitates the identification and estimation of the TE.

In fact, the proper identification and estimation of the TE with this matching strategy requires precise conditions. Whether the application of matching to the case of interest here (the FPR) really satisfies these conditions can be questioned. In particular, one problem is the lack of balance between the two groups of units to be matched and compared. Not only the number of counterfactuals is low (i.e., the non-treated units are few compared to the treated units); they are also peculiar and, maybe, unsuitable for the comparison with the treated units.

3.1.1. Identification issues: Unconfoundedness and ATT

Consider a sample of *N* observations (farms). Let Y_i indicate the outcome variable observed in the generic i-th farm (unit), $i = 1,..., N$ (where *N* is the sample size) and $T=0,1$ the binary policy treatment (T=0 if not treated, T=1 if treated). Let's assume that the attribution of a given treatment to the i-th farm does not affect the TE on the j-th farm, $\forall j \neq i = 0,...,N$. This assumption is called *stable-unit-treatment-value assumption (SUTVA)*. It seems plausible whenever micro data are used and the treatment assignment to single units may hardly have aggregate (or macro) effects (e.g., on partial and general equilibrium market adjustments) (OECD, 2011). In the present case, however, the SUTVA also implies the absence of diffusion or spillover effects like, for instance, imitation.[7]

[6] As will be detailed in section 4, the FADN/RICA database now allows large-enough farm panels in this respect (Cagliero et al., 2010).

[7] As the FPR concerns farms' market orientation, excluding such spillover effects of the treatment may seem a relevant assumption. Nonetheless, it is still hardly testable. Chabé-Ferret and Subervie (2013) actually suggest that a test on this assumption's validity can be still attempted by looking at neighboring farmers' outcome variable dynamics before and after the treatment. Such kind of statistical test, however, would imply making the spatial dimension explicit within the adopted panel sample, that is, to introduce spatial econometrics techniques. This solution seems computationally demanding when micro data are used (6542 units in the present case) and is here ignored. However, it may represent an interesting direction for future research.

By *Average Treatment Effect* (ATE) we intend the following expected value:[8]

(1) $ATE_i = E(Y_{i1} - Y_{i0})$

ATE actually expresses the difference that would be observed in the outcome in a purely experimental (or randomization) situation, that is, as the same i-th farm were observed, in sequence, under treatment and non-treatment. In practice, with observational (or non-experimental) data, we really observe only the outcome under one of the possible states. The outcome in all other cases is, in fact, hypothetical or potential (Rubin, 1974; Imbens and Wooldridge, 2009).[9] With observational data, the actual research question is: which is the impact of the treatment on the outcome observed in treated units? The answer to this question is provided by the *Average Treatment effect on the Treated* (ATT):

(2) $ATT = E(Y_1 - Y_0 | T = 1)$

where the answer only concerns the units that were actually treated and does not apply to units that were not treated.

Frölich (2004), Nichols (2007) and Imbens and Wooldridge (2009), just to mention a few, provide a clear explanation on why in non-experimental settings $ATE \neq ATT$. As we can write $ATE = E(Y_1 - Y_0) = E(Y_1) - E(Y_0)$ and $ATT = E(Y_1 - Y_0 | T = 1) = E(Y_1 | T = 1) - E(Y_0 | T = 1)$, ATE and ATT are equal only if $E(Y_0) = E(Y_0 | T = 1)$ and $E(Y_1) = E(Y_1 | T = 1)$, that is, only if the expected value of a given treatment is independent on the subsample (treated units or counterfactuals) on which we are measuring it. In non-experimental settings, this condition is granted only under specific assumptions. Even if we just focus on $ATT = E(Y_1 - Y_0 | T = 1) = E(Y_1 | T = 1) - E(Y_0 | T = 1)$, unless these strong assumptions are made, in non-experimental settings the *TE* remains unidentified.

The key identification problem comes from the fact that, in observational data, we can not observe the counterfactual or potential outcome $E(Y_0 | T = 1)$, that is, the outcome that would be observed if the treated units were not treated. What we really observe is only $E(Y_0 | T = 0)$. In practice, with observational data we can only compute the difference $E(Y_1 | T = 1) - E(Y_0 | T = 0)$ but this difference does not necessarily correspond (i.e, does not identify) the *ATT* as it is:

(3)

$$E(Y_1 | T = 1) - E(Y_0 | T = 0) = [E(Y_1 | T = 1) - E(Y_0 | T = 1)] - [E(Y_0 | T = 0) - E(Y_0 | T = 1)] = ATT - [E(Y_0 | T = 0) - E(Y_0 | T = 1)]$$

[8] If i indexes a randomly drawn unit in the population we can also write $ATE = ATE_i$, where ATE is also called *Population Average Treatment Effect* (PATE). In the sample, ATE is calculated averaging ATE_i across the sample units and it is also called *Sample Average Treatment Effect* (SATE) (Abadie et al., 2004). Henceforth, in the present study, whenever the ATE has to be intended as SATE, ,we drop the *i* index in the notation for simplicity.

[9] This is also known as the *fundamental problem of causal inference* (Holland, 1986). From this perspective, estimating the parameters of the potential outcome distribution is a missing data problem because we can see only one outcome per individual.

The latter term of the right-hand side is the so-called *selection bias* as it corresponds to the difference between what we can observe, $E(Y_1 | T=1) - E(Y_0 | T=0)$, and what we want to estimate (*ATT*). It is a "selection" bias because it occurs whenever a difference in the outcome between the treated and the control units would be observed regardless the treatment itself. So, the difference does not depend on the treatment but on how these units have been selected within the two groups: there is some unobserved difference between the two groups that, at the same, conditions the participation to the treatment and the outcome regardless of the treatment. As the selection bias depends on the presence of some unobserved characteristics, it is also called *selection-on-unobservable bias* but it is, in fact, an *omitted variable bias*. Getting rid of this bias, consists, in practice, in finding ways to make the term $[E(Y_0 | T=1) - E(Y_0 | T=0)] = 0$ and, consequently, $E(Y_1 | T=1) - E(Y_0 | T=0) = ATT$.

A methodological solution to this identification problem directly tackles the issue of selection-on-unobservables. The idea is that, though a selection bias may be observed, the analyst is in the condition to detect and observe all the pre-treatment variables or characteristics **X** that generate it. This approach is generally referred to as *selection-on-observables* approach. In practice, the identification of the ATT is achieved by assuming that:

(4) $ATT = E(Y_1 | \mathbf{X}, T=1) - E(Y_0 | \mathbf{X}, T=0)$ since $[E(Y_0 | \mathbf{X}, T=1) - E(Y_0 | \mathbf{X}, T=0)] = 0$

This is the so-called *Conditional Independence Assumption* (CIA) or *Unconfoundedness Assumption* as we are assuming that, once we control for all relevant pre-treatment covariates **X**, the selection bias disappears. In particular, we recreate the condition of a randomized experiment and the ATT can be estimated by directly computing the difference between the observed outcome of the treated and the control units. Vector **X** is expected to contain all those pre-treatment variables that are, at the same time, correlated to the treatment assignment and to the outcome variable: once we control for them the difference in the outcome can be exclusively attributed to the treatment. This identification assumption can be hardly tested, if not *ex post*,[10] and still remains the critical point of this approach as we can not definitely exclude that a further unobserved confounding variable (i.e., correlated with both the treatment assignment and the outcome variable) still exists.

Nonetheless, under the CIA, the identification of the ATT is granted and the problem becomes how to estimate the conditional expected values in (4), that is, $E(Y_1 | \mathbf{X}, T=1)$ and $E(Y_0 | \mathbf{X}, T=0)$. An easy way could be to estimate $E(Y_1 | \mathbf{X}, T=1)$ and $E(Y_0 | \mathbf{X}, T=0)$ by running a parametric regression of Y on **X** and *T* (a dummy variable). Nonetheless, this parametric approach finds two major drawbacks. First of all, a parametric linear specification of the relation between Y_1 on **X** must be assumed, whereas this relation might be more complex and vary over the sample. Secondly, such approach estimates the ATT by using all the observed variation of covariates **X** while, in fact, it may be the case

[10] See Chabé-Ferret and Subervie (2013) for more details.

that only a portion of this range of variation is common to both treated and control units. This common portion is also called *common support* and it is intuitively more appropriate to limit the estimation of the conditional expectations in (4), thus of the *ATT*, to this common support (Imbens and Wooldridge, 2009).

For these two reasons, the recent empirical TE literature tends to prefer nonparametric approaches to the estimation of (4) and, in particular, it adopts matching estimators. Among this PSM estimator has become the most popular solution.

3.1.2. Alternative Estimators and Robustness Check

Generally speaking, matching is a statistical procedure that aims at pairing each treated observation with one or more control (non-treated) units showing the closest (i.e. statistically equal) observed covariates **X** in such a way we can assume that the treatment assignment to these pairs is random (i.e., units are equal for all relevant and observable characteristics except for the treatment). In practice, matching raises two serious empirical issues (Nichols, 2007). The first problem consists in finding a metric (a scalar variable) measuring the distance among observations across the elements of vector **X**. Once this metric has been established, the second problem consists in finding appropriate rules to match treated and non-treated units (or groups) on the basis of this metric. Both aspects are computationally more demanding the greater the dimension of **X**. At the same time, however, a larger **X** guarantees about the validity of the CIA. Therefore, the key empirical problem in matching (also know as the *curse of dimensionality*) is the trade-off between the validity of the identifying assumption and the often unaffordable computational burden to achieve a proper matching.

In their seminal contribution, Rosenbaum and Rubin (1983) proposed to use a scalar variable, the propensity score, as the metric to perform the matching. The Propensity Score (PS), $p(\mathbf{X})$, is the probability of a given unit of being treated conditional on covariates **X**:

(5) $p(\mathbf{X}) \equiv \Pr(T|\mathbf{X}) = E(T|\mathbf{X})$

The key contribution of Rosenbaum and Rubin (1983) consists in demonstrating that, under the CIA, if treatment assignment is random conditional on **X** then it is also random conditional on $p(\mathbf{X})$. As matching based on $p(\mathbf{X})$ (a scalar) is empirically much more affordable than matching based on **X** (a vector), the ATT can be thus more easily estimated (PSM estimation) as:[11]

(6) $ATT = E(Y_1 \,|p(\mathbf{X}), T = 1) - E(Y_0 \,|p(\mathbf{X}), T = 0)$

The identification of the ATT on the basis of the PS still depends on the CIA assumption. The difference is that the assumption holds on $p(\mathbf{X})$. In other words, under the CIA, the PSM implies that $E(Y_1 \,|p(\mathbf{X}))$ is independent on **X** as the PS already contains all the information about how **X** conditions

[11] For more details and discussion on the asymptotic properties of these PSM estimators see Hahn (1998) and Imbens and Wooldridge (2009).

the treatment assignment. When refereed to $p(\mathbf{X})$, the CIA is also called the *Balancing Hypothesis* (Becker and Ichino, 2002) because the validity of this hypothesis can be tested within the sample by checking the *balancing condition*: observations showing a very close PS also show a statistically equal distribution of **X** independently on the treatment status. An empirical practice that contributes to the balancing condition and improves the quality of matching consists in imposing the common support in $p(\mathbf{X})$. It implies that those treated (control) units whose $p(\mathbf{X})$ does not find a corresponding value in at least one control (non-treated) unit are excluded from the analysis.

PSM identifies and estimates the ATT following a three-step procedure. Firstly, a parametric binary choice model (also called the *Propensity Score Equation*, usually taking the form of a conventional binomial probit or logit model) is estimated to obtain estimates of $p(\mathbf{X})$. Secondly, usually imposing common support and once the balancing condition has been validated,[12] the matching of units is performed on the basis of $p(\mathbf{X})$ and pair-wise (or group-wise) ATTs are computed as in (6). Finally, the average ATT is computed over the whole sample (or the common support) as the weighted average (where weights depend on the number of treated units) of the pair or group-wise ATTs.

The second step of this procedure (the matching) is critical because it can be achieved following different strategies. *Strictu sensu* matching implies that for any treated unit (or for blocks of units) we look for the closest control unit(s) in terms of $p(\mathbf{X})$ (*best matches*). This case can take three alternative forms: *Nearest Neighbour Matching*, where matching is made one-by-one;[13] *Stratification Matching*, where matching is made on groups or block of units; *Radius Matching*, where any treated unit is matched with all control units falling within a predetermined distance, or radius *r*, from its own $p(\mathbf{X})$. An alternative strategy adopts a weighting procedure. Any treated unit is matched with all control units (within the common support) but each of them is weighted by the inverse of its distance from the $p(\mathbf{X})$. This case is known as *Kernel Matching*.

There is no clear-cut and univocal indication on which of these matching approaches should be preferred, though under the CIA they are asymptotically equivalent. In practice, in finite samples the performance of Stratification Matching is usually poorer compared to the other solutions while Kernel Matching is preferable. At the same time, given the specific conditions on which matching is performed a trade-off between bias and variance (*accuracy*) is often observed (Abadie and Imbens, 2002; Abadie et al., 2004). For these reasons, presenting results of all these matching procedures may serve as sensitivity analysis to assess the robustness of the ATT estimates.

In fact, limiting matching estimation to the common support and testing for the balancing condition only ensures against the *selection-on-observables bias* but, evidently, can do nothing against the *selection-on-unobservables bias*. Evidently, these variables being unobservable, the presence of the

[12] See Becker and Ichino (2002) for more details on this first stage of the PSM estimation.
[13] As the number of treated and control units may be different, replacement is allowed, that is, the same control unit can be the best match for more than one treated units.

bias they generate can not be tested: it is an hidden bias. At the same time, however, if there are unobserved variables that affect assignment to treatment and the outcome variable simultaneously, the respective hidden bias would make matching estimators not robust. As a consequence, checking the robustness and sensitivity of the matching estimates has become an increasingly important topic in the applied evaluation literature as a sort of indirect evidence of the presence of this hidden bias.

To check whether there is some evidence of this selection-on-unobservables bias, Rosenbaum (1987; 2002) has proposed a bounding approach assessing how strongly an hypothetical unmeasured disturbing variable must influence the selection process to undermine the results of matching. This bounding approach does not test the unconfoundedness assumption itself.[14] Instead, Rosenbaum bounds provide evidence on how much the statistical significance of a matching result depends on this untestable assumption. If the results turn out to be significantly sensitive to this disturbing factor, the matching estimation of the ATT can not be considered robust enough with respect to a possible unobserved heterogeneity between treatment and control cases.

DiPrete and Gangl (2004) implements the Rosenbaum bounds approach by calculating the Wilcoxon signrank test, a nonparametric test providing upper and lower bound estimates of significance levels at given values of the hidden bias, the Γ parameter. Γ is a measure of the degree of departure from the CIA. $\Gamma=1$ whenever matched individuals have the same probability of participating to the treatment; otherwise, the greater Γ (>1), the more units that appear to be similar (in terms of observed covariates) can actually differ in their odds of receiving the treatment. Under the assumption of additive treatment effects, the DiPrete and Gangl (2004) approach also provide Hodges-Lehmann point estimates and confidence intervals for the ATT.

An alternative implementation of the Rosenbaum approach is put forward by Nannicini (2007) and Ichino et al. (2008). Their application simulates a potential confounder in order to assess the robustness of the matching estimation of the ATT if the CIA is violated. The analysis is based on a simple idea. Under the hypothesis that the CIA is not satisfied given the observables but would be satisfied if one could observe an additional binary variable, a potential confounder is simulated in the data and used as an additional covariate in performing the matching estimation. The comparison of the estimates obtained with and without matching on the simulated confounder shows to what extent the original results are robust to specific sources of failure of the CIA, since the distribution of the simulated variable is aimed at capturing different hypotheses on the nature of potential confounding factors (Nannicini, 2007).

In the present empirical application, both applications of the Rosebaum bounds approach will be used.

[14] This would mean testing that no (unobserved) variable influences the selection into treatment.

3.2. *Differences-In-Differences (DID)*

3.2.1. *Identification issues*

Beside any helpful sensitivity analysis, it remains true that the major drawback of the matching method consists in the possible and not testable presence of hidden biases caused by unobservable covariates. In particular, in the case of the FPR the treatment is applied to a large majority of farms, therefore it is a non-selective generalized policy. Farms not involved by this policy are very peculiar and, therefore, represent a self-selected exception, thus apparently not suitable counterfactuals in matching estimation. It seems more natural to consider, as suitable counterfactuals, the same treated units observed before the treatment. After all, any treated unit can be compared with itself observed before the treatment (the FPR). Such approach is called *Difference-In-Difference (DID)* estimation of the ATT (Smith and Todd, 2005).

The DID estimation of the ATT is thus the following:

(7) $$ATT = \left[E\left(Y_{1,t+h} \middle| T = 1\right) - E\left(Y_{t-k} \middle| T = 1\right)\right] - \left[E\left(Y_{0,t+h} \middle| T = 0\right) - E\left(Y_{t-k} \middle| T = 0\right)\right]$$

Where the treatment occurs at time t and units are observed before (*t-k*) and after (*t+h*) the treatment.[15]

If *k=h*, (7) is also called the *symmetric DID estimator* (Chabé-Ferret and Subervie, 2013).

The ATT is thus identified and estimated as the different variation observed in the outcome variable between the treated ant the non-treated units under the assumption that this difference can be entirely attributed to the treatment itself. It is worth noticing that in such case the counterfactual observation is the treated unit itself before the treatment. Therefore, in principle, it is not necessary to condition the comparison on a set of pre-treatment covariates **X** that may affect, at the same time, outcome and treatment assignment. Consequently, matching itself is not needed and the CIA is not binding. Nonetheless, compared to matching, DID estimation implies three further requirements.

First of all, the DID identification and estimation of the ATT requires that the outcome is repeatedly observed over time in both the treated and the control units. In practice, it requires also a time dimension in the data set while the matching approach can be typically performed only on cross-sectional data. This is even more true in the present case. Here, as will be clarified in sections 5.1, the outcome variable is itself a difference between two years/periods. The implication is that to compare the before and after-treatment outcome variable at least four years must be observed: years *t-h* and *t-h+1* before the treatment, years *t+k-1* and *t+k* after the treatment. Therefore, a large enough (in the time dimension) panel dataset is needed to perform DID estimation of the ATT.

Secondly, the DID estimation is based on the assumption that treatment effects are instantaneous. On the one hand, it means that the treatment is assumed to have an impact only after its application, that

[15] $E\left(Y_{1,t+h} \middle| T = 1\right)$ indicates the expected value of the outcome variable that is observed, *h* times after the treatment, in units assigned to the treatment whenever they actually received the treatment. Conversely, $E\left(Y_{0,t+h} \middle| T = 1\right)$ indicates the expected value of the outcome variable that would be observed, *h* times after the treatment, in units assigned to the treatment under the hypothesis that they did not receive the treatment.

is, *anticipation effects* (i.e., outcome variable's response to the treatment before the treatment) are excluded (Chabé-Ferret and Subervie, 2013). On the other hand, such assumption excludes that the response to the treatment, though starting in a well-identified point in time, then continues for some time (*lagged effects*) thus making the identification of the post-treatment period not univocal.

The third, and more important, requirement for the DID estimation to be valid has to do with the time-varying effects. Evidently, the time dimension may definitely affect the outcome. To get rid of the effects of time (i.e., of specific time-varying factors or variables) that are invariant across the treatment, DID estimation uses the variation over time of the outcome variable in the treated units, $\left[E\left(Y_{1,t+h}|T=1\right)-E\left(Y_{t-k}|T=1\right)\right]$ and compares it to the same variation in the non-treated units, $\left[E\left(Y_{0,t+h}|T=0\right)-E\left(Y_{t-k}|T=0\right)\right]$, since in this latter case the variation can be fully attributed to the action of time. Unfortunately, as we are not in a purely experimental situation, not only time itself may affect the outcome, but this influence may vary across the treatment. Therefore, the identifying assumption underlying DID estimation is that the influence of time on the outcome variable is the same across treated and non-treated units. This assumption, somehow analogous to the CIA in the time dimension, is called *parallel-trend assumption* or *Conditional mean-Independence of Increments Assumption* (CIIA):

(8) $$\left[E\left(Y_{0,t+h}|T=1\right)-E\left(Y_{t-k}|T=1\right)\right]=\left[E\left(Y_{0,t+h}|T=0\right)-E\left(Y_{t-k}|T=0\right)\right]$$

As for the CIA, also the CIIA can not be empirically tested since the left hand side of (8) can not be observed. In the present application, the CIIA seems particularly strong as, in fact, the already mentioned structural (but unobservable) differences between the control and the treated groups might definitely imply different dynamics of the outcome variables over the observed period (for instance due to different specific agricultural market dynamics).

Also the anticipation and lagged effects can not be excluded in the case of the FPR. On the one hand, while the year-by-year production choices may have hardly anticipated the 2005 regime change, investments decisions might definitely have been taken before the FPR entered into force, as farmers were already fully informed, since 2003, about contents and timing of the reform. On the other hand, the FPR took three years (2005-2007) to fully enter into force (the decoupling of support for all products, i.e., all CMOs) (OECD, 2011, pp.64-65). Therefore, though most of the effects concentrate in 2005, some effect can be still observed in 2006 and 2007.[16] Therefore, only working with the extreme years (2003-2007) of the observed period can take both possible anticipation and lagged effects into account.

[16] Actually, the decoupling of support of most agricultural productions relevant for Italian agriculture (therefore, for the farm sample under study here) concentrates in 2005 and 2006. The only exception is wine and fruit&vegetable production whose decoupling of support took place in 2008 (Povellato and Velazquez, 2005; Frascarelli, 2008). As this year is here not included the analysis, those farms that are strongly specialised in these productions are here dropped from the sample. These aspects will be discussed in section 4.

More generally, though appealing, the application of the DID estimation to the present case raises several practical issues that require appropriate empirical strategies. First of all, the reliability of the CIIA assumption can be somehow assesed whenever two pre-treatment periods are observed. The DID applied to these periods should estimate a not significant ATT under the validity of the CIIA. On the contrary, a significant ATT would indicate that the CIIA is not supported by data. For evident reasons, such strategy is also called *placebo testing* (Chabé-Ferret and Subervie, 2013; Di Porto et al., 2014). In the present case, two pre-treatment periods are not available, since the only pre-treatment observation is the 2003-2004 variation. Nonetheless, we have three no-treatment observations, 2003-2004, 2005-2006 and 2006-2007. Applying the DID estimation to these periods may thus represent a placebo test, though, as mentioned, the latter two observations may still incorporate some lagged effects of the FPR. Another empirical strategy that may contribute to the validity of the CIIA consists in performing only the symmetric DID estimation (Chabé-Ferret and Subervie, 2013). In the present case, this implies comparing observations 2004-2005 and 2003-2004 (or 2005-2006) which is in contrast, however, with the abovementioned suggestion of using the extreme years of the interval under study.[17]

3.2.2. Conditional DID (CDID)

It is worth noticing that the DID estimation by itself (i.e., the *unconditional* or *naïve DID*) can not get rid of the selection-on-observables bias. As the method still requires the comparison between differences observed in treated and non-treated units (see (7)), it can not be excluded that such differences depend on a set of pre-treatment observable covariates **X** which, in turn, also affect the treatment assignment. In fact, even in the DID approach the identification and estimation of the ATT can be performed conditional on **X**: this is the Conditional DID (CDID) (Chabé-Ferret, 2010; Villa, 2012).

Heckman et al. (1997; 1998) (see also Heckman, 2005) firstly proposed a CDID estimation by combining a matching approach with a DID estimation. This combination evidently requires a panel dataset and allows controlling for both the observed and the unobserved heterogeneity, provided the CIIA holds true. Therefore, the CDID estimation is an extension of matching estimation robust to selection-on-unobservables and for this reason this estimator has been found to be the closest to the experimental benchmark (Smith and Todd, 2005). As Abadie (2005) points out, in CDID estimation also the CIIA can be reformulated in a weaker version conditional on observed covariates:

(9) $$\left[E\left(Y_{1,t+h}\middle|\mathbf{X},T=1\right)-E\left(Y_{t-k}\middle|\mathbf{X},T=1\right)\right]=\left[E\left(Y_{0,t+h}\middle|\mathbf{X},T=0\right)-E\left(Y_{t-k}\middle|\mathbf{X},T=0\right)\right]$$

This reformulated CIIA implies that, in CDID estimation, units with significantly different observed characteristics may still experience different increments without generating a bias in the ATT estimation. The CDID ATT estimator can be thus written as follows:

[17] See section 6.2.2 for more details with respect to the observations here considered to perform the DID estimation.

(10) $ATT = \left[E\left(Y_{1,t+h} | \mathbf{X}, T = 1\right) - E\left(Y_{t-k} | \mathbf{X}, T = 1\right)\right] - \left[E\left(Y_{0,t+h} | \mathbf{X}, T = 0\right) - E\left(Y_{t-k} | \mathbf{X}, T = 0\right)\right]$

Following the discussion made in section 3.1.2, to reduce the curse of dimensionality due to a large X, the CDID estimation can be evidently combined with the PSM and the ATT is then estimated as follows:

(11) $ATT = \left[E\left(Y_{1,t+h} | p(\mathbf{X}), T = 1\right) - E\left(Y_{t-k} | p(\mathbf{X}), T = 1\right)\right] - \left[E\left(Y_{0,t+h} | p(\mathbf{X}), T = 0\right) - E\left(Y_{t-k} | p(\mathbf{X}), T = 0\right)\right]$

Provided that matching eliminates the selection bias due to observed pre-treatment (or time invariant) covariates, (11) is the correct (unbiased) estimator of the ATT under the SUTVA and the conditional CIIA (Chabé-Ferret and Subervie, 2013).

Introducing the time dimension in ATT estimation, however, may also generate other uncontrolled sources of bias.[18] CDID estimation controls for those sources of bias coming from the common (across matched units) time-varying effects and from the time-unvarying (fixed) unobserved individual effects, while it assumes that the time-varying (non-fixed) and unobservable individual effects are negligible.[19] Time-varying individual effects, however, occur whenever observed covariates are themselves time-varying and dependent (non-separable) on the fixed effects and on the treatment. More generally, admitting this time-varying individual effects means to take into account that the time pattern of the outcome variable and of the covariates matters.[20]

What is worth emphasizing, here, is that the DID approach (conditional or not) still needs a control group of farms, that is a sample of non-treated farms to perform the ATT estimation. Therefore, it does not really solve the basic problem underlying matching estimation in the present case: when the FPR is under study, the use of the non-treated cases is always critical. The peculiarity of these units strongly questions the validity of the CIA in the case of matching, as well as the validity of the CIIA in the (C)DID estimation because this peculiarity may also concern their variables' patterns over time. It would be helpful, rather, to find an alternative estimation strategy, to estimate the TE without making any use of non-treated farms. This can be achieved in the multivalued treatment case.

3.3. Multivalued treatment and the Generalized Propensity Score (GPS)

An alternative approach to indentify and estimate the effect of a treatment can be adopted any time the treatment is not binary ($T = 0,1$), but behaves as a multivalued (either discrete or continuous) variable

[18] The SUTVA itself can be less acceptable. Diffusion effects (i.e., the effect of the treatment on the non-treated's outcome) may occur but they take time (imitation, partial and general equilibrium market adjustments, etc.). Therefore, though it is reasonable to rule them out in cross-sectional matching, they may become relevant over time.

[19] There is an evident analogy, in this respect, with the identifying conditions for conventional fixed-effects panel models (Chabé-Ferret and Subervie, 2013).

[20] For instance, it matters whether the stochastic processes generating these variables show any autocorrelation and are around their stationary long run equilibriums (i.e., they are stationary processes). In such cases, controlling for symmetry in the DID can be even more relevant as symmetric CDID is the least biased estimator under full information. Chabé-Ferret and Subervie (2013) provides an excellent detailed analysis about the CDID validity under such circumstances. These aspects will not be explicitly considered here but should still be carefully taken into account in interpreting the CDID estimates presented in the following sections.

($T \in \mathbf{N}$ or $\mathbf{R}^+$) and the response of the outcome variable (Y) to the treatment is itself continuous. In such circumstance, the intensity of the treatment can be correlated to the magnitude of this response and this allows the identification and estimation of the TE without using the non-treated units. In fact, these latter are no more needed to observe how the Y varies with T|X. Any treated unit can behave as a counterfactual case for units with a different treatment intensity. As in this approach the treatment effect is estimated only looking at the treated units, without involving the non-treated cases, it seems appropriate refer to the estimated treatment effect not as ATT but simply as ATE.

In general terms, a continuous treatment does not alleviate the fundamental problem of identification and estimation of the ATE, i.e. the selection bias due to the presence of observable and non-observable factors that affect, at the same time, the treatment assignment and the outcome. In fact, in the case of a multivalued treatment, a much more naïve approach to the ATE estimation could be to simply estimate, the following regression model with OLS:

(12) $E\left[(Y_i | T_i, \overline{\mathbf{X}}_i\right] = \boldsymbol{\beta}'\overline{\mathbf{X}}_i + \alpha T_i$

where the estimated α or, more generally, the estimated first order derivatives of $E\left[(Y_i | T_i, \overline{\mathbf{X}}_i\right]$ with respect to *T* indicates the ATE. As discussed so far, however, this estimate might incur a selection bias generated by the fact that the treatment assignment itself, T_i, actually depends on $\mathbf{X}_i$ and, even if the unconfoundedness assumption potentially holds true (i.e, $\mathbf{X}_i$ takes all factors influencing both outcome and treatment assignment into account) (12) may still not properly represent the stochastic properties of the assignment of any i-th unit to T_i given $\mathbf{X}_i$. In practice, writing (12) as $Y_i = \boldsymbol{\beta}'\mathbf{X}_i + \alpha T_i + \varepsilon_i$, with ε_i representing the conventional spherical disturbance term, we can not exclude that T_i is endogeneouns, that is, $E(T_i\, \varepsilon_i) \neq 0$. Moreover, the relationship between Y_i and T_i it is actually unknown; unlike (12), it may be non-linear and actually vary across treated individuals.

The case of the FPR seems particularly suitable for a multivalued treatment approach. In Italy the SFP has been established on a historical basis (Povellato and Velazquez, 2005; Frascarelli, 2008), therefore the amount of support shifting from coupled to decoupled payment varies across treated units (farms) and this treatment clearly behaves as a continuous variable. We can here reasonably assume that for a farm with given characteristics (that is, conditional on **X**) the larger the coupled (thus constraining the production choices) support then converted into the SFP, the larger the expected change in the production mix (market reorientation). Nonetheless, whenever we have a multivalued treatment, therefore even in the present application, the critical issue shifts from finding appropriate counterfactuals to properly define the functional relationship between *Y* and *T*|**X**.

3.3.1. The Hirano-Imbens approach: Matching and the Dose-Response Function (DRF)

The approach that follows this intuition has been originally proposed by Hirano and Imbens (2004) and it is based on the concept of Generalized Propensity Score (GPS). In a broad sense, it can be considered a generalization of the conventional matching estimation based on the Propensity Score (PSM). The Hirano-Imbens approach can be described as sequence of three steps.

For any treated unit $i = 1, \ldots, N$, we observe the covariates $\mathbf{X}_i$, the treatment level T_i, the outcome variable Y_i. We define, $\forall i$, a set of *potential outcomes* $\{Y_i(T)\}_{T \in \Xi}$ where Ξ is the set of potential treatment levels and $Y_i(T)$ is a random variable that maps, for the i-th unit, a particular potential treatment, T, to a potential outcome. Evidently, of these potential outcomes only one is observed, that associated with the actual treatment T_i. Hirano and Imbens (2004) refer to $Y_i(T)$ as the *unit-level Dose-Response Function* (*uDRF*). In fact, we are interested in the *average Dose-Response Function* (*aDRF*), $aDRF(T) = E[(Y(T)]$.

The first step of the Hirano-Imbens approach consists in the estimation of the univariate variable GPS. Hirano and Imbens (2004) define the GPS_i as the probability that the i-th unit is assigned the treatment level T given its observed characteristics $\mathbf{X}_i$: $GPS_i = r(T_i, \mathbf{X}_i)$, where $r(T, \mathbf{X})$ is the *propensity function*, i.e., the conditional density of the actual treatment given the observed covariates. For the GPS to be meaningful in the calculation of the ATE, the following condition must be satisfied within the sample: units with statistically equivalent values of **X** are expected to show, around a given interval of *GPS*, both treatment levels lower and higher than a given level *T*. This is the balancing condition in the continuous treatment case. Hirano and Imbens (2004) demonstrates that if this condition is respected, and CIA assumed, the assignment to treatment is unconfounded, given the estimated GPS.[21] Therefore, the different *Y* observed across units showing the same estimated $GPS|\mathbf{X}$ can be fully attributed to the different treatment level *T*. Once the propensity function is estimated and the balancing condition is met, the second methodological step consists in estimating the conditional expectation of the potential outcome as a function of two scalar variables, the estimated *GPS* and *T*: $g(G\hat{P}S, T) = E\left[Y \middle| G\hat{P}S, T\right]$. The third step estimates the *aDRF* as $aDRF(T) = E[(\hat{g}(T)]$, $T \in \Xi$, that is, by averaging the estimated conditional expectation $\hat{g}(G\hat{P}S, T)$ over the *GPS* at any level of the treatment we are interested in.

The estimation steps imply arbitrary specification assumptions whose validity can be only assessed *ex post* by checking for the robustness of results. The complexity of the overall estimation procedure together with this arbitrariness may explain why, though originally proposed some years ago by Hirano and Imbens (2004), this approach has been only recently applied to the evaluation of

[21] In the multivalued treatment case, Hirano and Imbens (2004) actually call the CIA *weak unconfoundedness assumption* since it only requires conditional independence to hold for each value of the treatment, rather than joint independence of all potential outcomes. Such assumption is also called (*weak*) *ignorability* (Cattaneo, 2010).

multivalued program or policies (Bia and Mattei, 2007, 2012; Flores and Mitnik, 2009; Kluve et al., 2012; Magrini et al., 2013) and never to the evaluation of the FPR.

The first arbitrary assumption implied by the method is the specification of the distribution of T_i conditional on $\mathbf{X}_i$ to compute its conditional density. The common practical implementation of the methodology, also followed here, assumes a normal distribution for the treatment given the covariates:

$$(13) \qquad r(T_i, \mathbf{X}_i) = T_i|\mathbf{X}_i \sim N\left(\boldsymbol{\beta}'\overline{\mathbf{X}}_i, \sigma^2\right)$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters and $\mathbf{X}$ is the matrix of covariates. Therefore, the assumption is that the propensity function is linear in unknown parameters that can be thus estimated by OLS. Evidently, it is possible to assume other distributions, to adopt different (even non-parametric) specifications other than the linear regression and to estimate the GPS by other methods such as MLE.[22]

In fact, while the normality assumption can be tested, the empirical specification of (13) remains arbitrary. In particular, it seems questionable here to assume a linear relationship between T and some set of conditioning variables $\mathbf{X}$. Nonetheless, this problem can be prevented by using $\overline{\mathbf{X}}$ instead if $\mathbf{X}$, where $\overline{\mathbf{X}}$ includes transformations (e.g., polynomial terms) of $\mathbf{X}$ and/or interactions terms across variables in $\mathbf{X}$, in such a way that $\overline{\mathbf{X}}$ satisfies both the normality assumption and the balancing condition.

The estimated GPS is thus calculated as:

$$(14) \qquad G\hat{P}S_i = \frac{1}{\sqrt{\left(2\pi\hat{\sigma}^2\right)}} \exp\left\{-\frac{1}{2\hat{\sigma}^2}\left(T_i - \boldsymbol{\beta}'\overline{\mathbf{X}}_i\right)^2\right\}$$

A second and, probably, more critical arbitrary assumption concerns the specification of the *uDRF*, $g(G\hat{P}S, T) = E\left[Y | G\hat{P}S, T\right]$, that is, the conditional expectation of the p

otential outcome with respect to T and the estimated *GPS*. The often adopted specification of the conditional expectation is a fully interacted flexible function (K,H-th order polynomial) of its two arguments providing a good approximation of the underlying unknown relationship:

$$(15) \qquad g\left(G\hat{P}S, T\right) = E\left[Y | G\hat{P}S, T\right] = \alpha_0 + \sum_{k=1}^{K} \alpha_k (T)^k + \sum_{h=1}^{H} \gamma_h \left(G\hat{P}S\right)^h + \sum_{k=1}^{K}\sum_{h=1}^{H} \lambda_{kh} (T)^k \left(G\hat{P}S\right)^h$$

where $\alpha_0, \alpha_k, \gamma_h, \lambda_{kh}$ are unknown parameters to be estimated.[23] For each i-th unit the observed Y_i, T_i and the estimated GPS_i are used to estimate the unknown parameters of (15) by OLS. Our empirical approach (see next sections) start with the general form (15) and then adopts the best fitting specification according to the usual Akaike Information Criterion (AIC).

[22] See next section for an Multinomial Logit specification of (13). In any case, following Bia and Mattei (2008) and also to test for the validity of the normality assumption, a MLE instead of a OLS estimation of parameters $\boldsymbol{\beta}$ is here performed.

[23] Hirano and Imbens (2004) emphasize that there is no direct meaning (i.e., economic interpretation) of the estimated coefficients in (14), except that testing whether all coefficients involving the GPS are equal to zero can be interpreted as a test of whether the covariates introduce any bias.

The final step thus consists in using these estimated parameters to compute the average potential outcome at a given treatment level, *T*. $aDRF(T) = E[(\hat{g}(T)]$ is estimated as:

$$(16) \qquad aDRF(T) = E[(\hat{g}(T)] = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{\alpha}_0 + \sum_{k=1}^{K}\hat{\alpha}_k (T)^k + \sum_{h=1}^{H}\hat{\gamma}_h \left(G\hat{P}S_i\right)^h + \sum_{k=1}^{K}\sum_{h=1}^{H}\hat{\lambda}_{kh}(T)^k \left(G\hat{P}S_i\right)^h\right]$$

The entire *aDRF* can be thus obtained by computing this average potential outcome for each level of the treatment, i.e. $\forall T \in \Xi$. Bootstrap methods can be used to obtain standard errors of the estimated $aD\hat{R}F(T)$ taking into account the estimation of parameters in (13) and (15) (i.e., the entire estimation process is bootstrapped).[24] Eventually, the first order derivative of (16) with respect to *T* represents the ATE of the various treatment levels and as such is here estimated.

In fact, large part of the empirical problem in properly performing policy evaluation within this approach consists in the limited or null knowledge about how the policy effect may vary with different intensity of the treatment. In the specific case of the FPR, it is reasonable to argue that the decoupling of the support may have led to a change in production mix but this response only partially, and not linearly, depends on the amount of decoupled payments. There are economic and physical limitations to the extent of the farm response to the treatment. When this limit is approached and reached a further increase of the treatment intensity (i.e., *T*) may be ineffective. Therefore, in this specific case it seems reasonable to argue that a *DRF* does exist but it is expected to be monotonous and non-linear with a decreasing slope (first order derivative). Accordingly, in the case of the FPR, the ATE is expected to be positive but decreasing in *T*.

3.3.2. The Cattaneo approach: IPW and EIF estimation of the ATE

Evidently, beside the normality assumption in (13), the Hirano-Imbens GPSM approach strongly depends on the arbitrary and non-testable specifications adopted. A feasible empirical strategy to assess their reliability consists in comparing respective results with those obtained using alternative specifications or estimation strategies.

Cattaneo (2010) proposes an alternative approach to the estimation of the ATE under multivalued treatment. Though it shares several points in common with the original Hirano and Imbens (2004) estimation, the method proposed by Cattaneo (2010) significantly differs in the way the treatment (and the treatment variable) enters the analysis, how functional relationships are specified and, eventually, how the ATE is estimated. Therefore, this novel approach seems particularly interesting not just to pursue an allegedly superior estimation but to investigate the robustness of the results, i.e. the estimated ATE, once these applicative variants are admitted.

In fact, both approaches are based on the estimation of the GPS and on the consequent estimation of the potential outcome given the treatment. However, Cattaneo (2010) does not estimate a parametric

[24] As shown in Hirano and Imbens (2004), in principle, it is possible to derive the analytical calculation of the asymptotic standard errors of the estimated DRF. In practice, it may be unaffordable or highly computationally demanding.

DRF but look for a non-parametric identification and estimation of some parameter (e.g., the mean) of the statistical distribution of the potential outcome. While the Hirano and Imbens (2004) approach applies to continuous treatment, Cattaneo (2010) estimation implies multiple but discrete treatment levels. Thus, continuous treatment variables must be converted in advance into a categorical variable. Though both approaches share the parametric GPS estimation, therefore, the different nature of the treatment variable (continuous vs. categorical variables) makes the specification of the propensity function differ: a Multinomial Logit Model (MLM) in Cattaneo (2010) instead of the linear regression as in (13). Moreover, the Cattaneo (2010) approach is a three-step semi-parametric estimation where the first step (the parametric GPS estimation) is followed by two semi-parametric stages estimating the conditional distribution of the potential outcome given the treatment.

Cattaneo (2010) actually proposes two semiparametric-efficient estimation procedures and derives their large-sample properties: the Efficient-Influence-Function (EIF) and the Inverse-Probability Weigthed (IPW) estimators.[25] To shortly describe how these two estimations proceed, let's assume that there are J possible treatment levels $j = 0, \dots , J;$ therefore, it is $T_i = j$. Also define a new indicator variable $d_i(j)$ taking either value 1 if the i-th unit has received the j-th treatment or value 0 otherwise. It is thus possible to define the J potential outcomes $Y_i(j)$, $\forall j \in J$. $F_{Y(j)}(Y)$ is the distribution function of the potential outcome $Y(j)$, $j = 0, 1, \dots , J$, that is, the distribution of the outcome variable that would occur if individuals were administered the treatment level j. The only observed outcome variable, however, is given by $Y_i = d_i(0)Y_i(0) + d_i(1)Y_i(1) + \cdots + d_i(J)Y_i(J)$[26] where $\{Y_i(0), Y_i(1), \dots, Y_i(J)\}'$ is an independent and identically distributed draw from $\{Y(0), Y(1), \dots, Y(J)\}'$ for each i-th individual in the sample. The estimation method proposed by Cattaneo (2010) allows to estimate several parameters of the distribution function $F_{Y(j)}(Y)$. In particular, the interest here is in estimating the $Jx1$ vector of potential outcome means: $\mathbf{H}=(\mu_0, \mu_1, \dots , \mu_J)'$.[27]

Under some regularity conditions (Cattaneo 2010; Cattaneo et al., 2013), $F_{Y(j)}(Y)$ can be identified from observed data and, consequently, also its parameters in $\mathbf{H}$. Assuming weak unconfoundedness, this identification starts from the fact that $F_{Y(j)}(Y) = E\{F_{Y(j)|\mathbf{X}}(Y|\mathbf{X})\} = E\{F_{Y(j)|\mathbf{X}}(Y|\mathbf{X}, T = j)\}$, where the latter term is identifiable from the observed data. The estimation methods proposed by Cattaneo (2010), therefore, estimate $\mathbf{H}$ exploiting the fact that the observed potential outcome distributions have been marginalized over the covariate distributions. Cattaneo (2010) proposes two Z-estimators,[28] one constructed using an inverse-probability weighting scheme and the other constructed using the full

[25] Efficiency has to be intended with respect to the ideal case of randomized treatment: in this case, these estimators are more efficient estimators than the usual parametric estimators. As the weak unconfoundedness assumption and the balancing condition mimic the randomization case, this result can be extended to the non-experimental cases as far as such conditions are valid.

[26] The *fundamental problem of causal inference* in a multivalued treatment context.

[27] Cattaneo (2010) also presents and discusses the quantiles' estimation.

[28] Z stands for "zero" and indicates those estimators based on some conditions imposing a function of the data and of the unknown parameter to be estimated to be = 0 (as (18)).

functional form of the EIF. Both estimators are shown to be consistent, asymptotically normal, and semiparametric efficient under appropriate conditions. Therefore, the two estimators are asymptotically equivalent.

The IPW estimation extends the idea of inverse-probability weighting, widely used in the case of binary treatment, to a multivalued treatment context. The estimator is based on a set of moment conditions implied by the analysis above and by the weak unconfoundedness assumption. In this sense, it can be considered a sort of GMM estimation. These moment conditions are all motivated by the fact that for each treatment level j, the following holds true:

$$(17) \qquad E\left\{\frac{d(j)\,\left(Y-\mu_j\right)}{p_j(\mathbf{X})}\right\} = E\left[\frac{E\{d(j)|\mathbf{X}\}E\{Y(j)-\mu_j|\mathbf{X}\}}{p_j(\mathbf{X})}\right] = E\{Y(j)-\mu_j\} = 0$$

where $p_j(\mathbf{X})$ is the conditional probability function expressing the probability of receiving the j-th treatment conditional on $\mathbf{X}$.

From (17) it is possible to derive the following set of moment conditions based on observed data and from which the GMM estimation of **H** can be derived (those means that make all the moment restrictions hold true):

$$(18) \qquad E\left[\psi_{IPW,j}\{\mathbf{z}_i;\mu_j,p_j(\mathbf{X}_i)\}\right] = 0 \text{ with } \psi_{IPW,j}\{\mathbf{z}_i;\mu_j,p_j(\mathbf{X}_i)\} = \frac{d_i(j)\,\left(Y_i-\mu_j\right)}{p_j(\mathbf{X}_i)}$$

where $\mathbf{z}_i = (Y_i, T_i, \mathbf{X}_i')'$ simply represents the observed data (outcome variable, treatment level, covariates) and μ_j is the unknown element in **H** to be estimated. $p_j(\mathbf{X}_i)$ is not observed but can be estimated specifying a propensity function, as $\mathbf{p}(\mathbf{X}) = \{p_0(\mathbf{X}), p_1(\mathbf{X}), ..., p_J(\mathbf{X})\}$ simply is the GPS|**X**. Once the GPS is estimated, i.e. $\hat{\mathbf{p}}(\mathbf{X})$, and provided that the balancing condition is respected, all elements in **H** can be estimated imposing restrictions in (18):

$$(19) \qquad \hat{\mu}_{IPW,j} \text{ s.th. } \frac{1}{N}\sum_{i=1}^{N}\psi_{IPW,j}\{\mathbf{z}_i;\hat{\mu}_{IPW,j},\hat{p}_j(\mathbf{X}_i)\} = 0$$

From (19), $\hat{\mu}_{IPW,j}$ can be also expressed in closed form as follows:

$$(20) \qquad \hat{\mu}_{IPW,j} = \left\{\sum_{i=1}^{N}\frac{d_i(j)}{\hat{p}_j(\mathbf{X}_i)}\right\}^{-1}\sum_{i=1}^{N}\frac{d_i(j)Y_i}{\hat{p}_j(\mathbf{X}_i)}$$

(20) clearly shows why this estimator is called IPW, as the GPS (the probability of receiving that treatment level) acts as a inverse weighting factor of the observed outcome Y_i.

Compared to IPW, EIF estimation uses more information on the marginal potential-outcome distribution and, therefore, involves further functions to be estimated.[29] In particular, let's define the following function for each treatment level:

(21) $$e_j(\mu_j;\mathbf{X}_i)=E\{Y_j(j)-\mu_j|\mathbf{X}_i\} = E\{Y_j(j)-\mu_j|\mathbf{X}_i, T_i = j\}$$

The EIF estimation is then obtained by imposing the following set of moment conditions:

(22) $$E[\psi_{EIF,j}\{\mathbf{z}_i;\mu_j, p_j(\mathbf{X}_i), e_j(\mu_j;\mathbf{X}_i)\}] = 0$$

with $\psi_{EIF,j}\{\mathbf{z}_i;\mu_j, p_j(\mathbf{X}_i), e_j(\mu_j;\mathbf{X}_i)\}=\dfrac{d_i(j)\,(Y_i-\mu_j)}{p_j(\mathbf{X}_i)}-\dfrac{e_j(\mu_j;\mathbf{X}_i)}{p_j(\mathbf{X}_i)}\{d_i(j)-p_j(\mathbf{X}_i)\}$.

As in the case of the IPW estimator, the EIF estimator uses these moment conditions replacing expectations with sample averages and unknown functions with appropriate (parametric or nonparametric) estimations. In particular, beside the GPS estimation, $p_j(\mathbf{X}_i)$, (22) also requires the estimation of $e_j(\mu_j;\mathbf{X}_i)$. This leads to the following estimates:

(23) $$\hat{\mu}_{EIF,j} \quad \text{s.th.} \quad \frac{1}{N}\sum_{i=1}^{N}\psi_{EIF,j}\{\mathbf{z}_i;\hat{\mu}_{EIF,j}, \hat{p}_j(\mathbf{X}_i), \hat{e}_j(\hat{\mu}_{EIF,j}, \mathbf{X}_i)\}=0$$

that can be expressed in closed form as follows:

(24) $$\hat{\mu}_{EIF,j} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{d_i(j)Y_i}{\hat{p}_j(\mathbf{X}_i)}-\left\{\frac{d_i(j)}{\hat{p}_j(\mathbf{X}_i)}-1\right\}\hat{Y}_i(j)\right]$$

where $\hat{Y}_i(j)$ is the predicted value of the outcome variable obtained regressing Y_i on $\mathbf{X}_i$ for those observations with $d_i(j)=1$.

Though asymptotically equivalent, there may be reasons to prefer one of the two estimators. The EIF is expected to be more robust in finite samples as it enjoys the so-called double-robust property when viewed from a (flexible) parametric implementation perspective, while the IPW estimator does not have this property. On the contrary, however, the IPW estimation could be preferred to the EIF case because of its simplicity though, in fact, if also the $\hat{\mu}_j$ variance-covariance matrix has to be estimated (and this is evidently needed for inference purposes) the IPW estimator requires implementing the same ingredients of the EIF estimator (Cattaneo, 2010; Cattaneo et al., 2013).

Therefore, in practice, the implementation of both estimators is very similar and requires the same three steps. First of all, $p_j(\mathbf{X}_i)$ must be estimated. As mentioned, this corresponds to the GPS estimation of Hirano and Imbens (2004), the only difference being that we have here a categorical dependent variable. Therefore, $p_j(\mathbf{X}_i)$ is here estimated with a nonlinear Multinomial Logit (ML)

[29] In statistics, the *influence function* is a measure of the dependence of the estimator on the value of one sample point. The EIF (Efficient Influence Function) is a concept adopted in the estimation of semiparametric models, that is, models combining a parametric form for some component of the data generating process with weak nonparametric restrictions on the remainder of the model.

estimation approach. Nonlinearity comes from the fact that as independent variables of the ML model we use $\overline{\mathbf{X}}_i$, that is, a flexible fully interacted polynomial of the covariates. This ML model is estimated via MLE. The second step requires the estimation $e_j(\mu_j, \mathbf{X}_i)$. Even in this case, estimation is obtained by specifying a flexible fully interacted polynomial (i.e., a parametric function) then estimated with OLS. In both cases, $p_j(\mathbf{X}_i)$ and $e_j(\mu_j, \mathbf{X}_i)$, the actual adopted specification is selected among all possible forms following the usual AIC.[30]

The final step consists in computing the ATE that can be easily obtained as:

(25) $ATE_{IPW,j} = (\hat{\mu}_{IPW,j} - \hat{\mu}_{IPW,j-1})$ and $ATE_{EIF,j} = (\hat{\mu}_{EIF,j} - \hat{\mu}_{EIF,j-1}), \forall j = 1,...,J$

To perform inference on these estimated ATEs, the estimated standard errors of $\hat{\mu}_{IPW,j}$ and $\hat{\mu}_{EIF,j}$ are obtained through bootstrapping of the whole estimation process.

4. THE SAMPLE

A suitable sample to apply the estimation methods discussed in the previous section must be observed over a period including both pre and post-treatment observations (years). In other words, the sample must be a balanced panel and must contain all the needed information about the outcome variables, the treatments and the confounding variables (covariates $\mathbf{X}$). As the objective here is to assess the impact of the FPR on (Italian) farm's production choices, these conditions are met by extracting a constant sample of Italian Farm Accountancy Data Network (FADN/RICA) farms yearly observed over a period including the pre and post-FPR years. The numerosity of the FADN database allows for a quite large balanced panel.

For the selection of the time period covered by this panel, the choice is here made to avoid years that are progressively far from the moment of the treatment (2005). Moreover, it seems appropriate to select a period of analysis that is symmetric with respect to the treatment (FPR) and contains most, if not all, of its effects while excluding other possibly overlapping effects due to other policy treatments (or other confounding factors) and that could occur before 2003 and after 2007. For instance, adding years 2000-2003 in the pre-treatment period can be troublesome as they may still incorporate some effects of the previous CAP reform (Agenda 2000) (Esposti, 2007). At the same time, the post-2007 years could raise the same kind of problem due the implementation of the so-called Health Check Reform (HC) (Esposti, 2011b). In addition, the considerable price turbulence observed in agricultural markets in years 2008, 2009 and 2010 (Esposti and Listorti, 2013), then accompanied by the negative effects on agriculture of the general economic crisis (De Filippis and Romano, 2010), suggests particular caution in adding these years to the post-treatment period.

[30] It may seem now more clear why the IPW and EIF estimations of the ATE are considered a semiparametric estimation compared to the fully parametric estimation of Hirano-Imbens. While the underlying estimation theory is grounded in nonparametric estimation, in practice, they still imply (though not strictly necessary) some parametric specifications.

For this reason, the balanced sample is here limited to the 2003-2007 interval. 6542 farms are observed over these five years.[31] This balanced panel constitutes the sample on which the present analysis is performed. It is worth noticing that the FADN sample is not fully representative of the whole national agriculture. The reference population from which the FADN sample is ideally drawn, in fact, excludes a significant (at least in terms of numerosity) amount of Italian farms (those with Economic Size<4 ESU, that is, less than 4800 Euro of Standard Goss Margin).[32] In this respect, the FADN sample is only representative of a sub-population of Italian farms, those farms that can be here refereed as *professional or commercial farms* (Cagliero et al., 2010; Sotte, 2006). Nonetheless, these 6542 farms are quite homogeneously distributed across the national territory, and the scattering of farms across the Italian macro-regions (North-West, North-East, Centre, South and Islands) well represents the pretty diverse agricultural conditions and structures of these different parts of the country (Esposti, 2011c).

5. APPLYING THE TE ESTIMATION TO THE FPR

5.1. Binary vs. Multilevel Treatment

The application of the TE estimation toolkit in the present case implies a proper identification of the treatment, therefore of the treated and non-treated units, and of the outcome variable. On the first aspect it is worth reminding that here the treatment (the FPR) the change of first pillar support from coupled to decoupled payments. Evidently, the participation to the treatment is not voluntary. It depends, in Italy, on the history of the individual farm and on the respective support it received in the 2000-2002 period; farms can not decide to remain in the old regime, as this is not admitted. Therefore, the treated units ($T = 1$) are those farms that received the first pillar CAP support in the form of coupled payment before the reform and then in the form of SFP.

According to such interpretation, the treatment consists in the change of the form of support and not in its amount as in Italy the conversion from coupled to decoupled payments has been defined on a purely historical basis. It is still possible to find farms that did not receive any CAP support in the old regime (for whatever reason mostly due to peculiar production and managerial choices). For them, the change in regime did not occur simply because they remain in a no-policy situation both before and after 2005, and no CAP first pillar support has been received over the entire 2003-2007 period. Therefore, they are not treated (control) units ($T = 0$) simply because they experience no change in the form of support.

[31] As mentioned in the previous section, this panel excludes farms that are strongly specialized in crops whose CMOs have been reformed only starting in 2008. Therefore, farms of typologies 1430, 2011, 2013, 3110, 3120, 3130 (see Table A2) with revenue share of vegetable or wine production higher than 75% are excluded from the sample.

[32] According to 2000 Census data, more than 82% of Italian agricultural holdings had an economic size smaller than 8 ESU but they accounted for just 27% of total Italian agricultural area (Sotte, 2006). According to 2010 Census data, about 67% of Italian agricultural holdings has an economic size smaller than 18 ESU but they account for just 17% of total Italian agricultural area (Sotte and Arzeni, 2013).

Moving from the binary treatment ($T = 0,1$) to the multivalued treatment case, the treatment level is here measured in terms of *treatment intensity* (*TI*). To take into account the different economic size of farms, *TI* is computed as the amount of first pillar support received by a given farm divided by its Gross Production Value (GPV). For both values, the yearly average over the whole 2003-2007 period is computed. Here, the multivalued treatment case only considers the treated units while non-treated units are excluded from the analysis under the assumption that, as already discussed in the case of the FPR, the selection-on-unobservables bias can not be ruled out. Therefore, *TI* is here always greater than 0.[33]

The distribution of the *TI* within the sample is displayed in Figure 1. If we exclude the 1124 (17%) non-treated units (for which $T = 0$), the *TI* across the 5430 treated units (for which $T = 1$) tends to concentrate in the [0-10%[interval (52% of treated units); the mode is, in fact, around 5%. For only 8% of treated units *TI* is equal or greater than 30%, while for only 1% *TI* is equal or greater than 50%. The maximum *TI* observed within the sample amounts to about 72%. Therefore, the continuous variable *TI* has a truncated distribution starting at 0, finding a peak at about 5% and then regularly declining up the maximum value. It is clearly a non-normal distribution. To restore normality, therefore, the *TI* variable must be properly transformed.[34] If we consider *TI* as a discrete or categorical variable, *n* intervals of *TI* can be identified. Here, 7 treatment intervals are considered beside the initial level of $TI = 0$ (non-treated units). Following the frequency histogram in Figure 1, these intervals are: $0<TI<5\%$; $5\%\leq TI<10\%$; $10\%\leq TI<15\%$; $15\%\leq TI<20\%$; $20\%\leq TI<25\%$; $25\%\leq TI<30\%$; $30\%\leq TI$.

With this measurement of the treatment intensity we can conclude that, in the case of the FPR and given the respective outcome variables presented in the next section, we expect a positive but decreasing ATE as the treatment intensity increases. This *monotone concave response* can be explained by three different factors. Firstly, the treatment intensity is a right-uncensored variable possibly ranging between 0 and $+\infty$. On the contrary, outcome variables y^1 and y^2 are right-censored variables. Therefore, when their maximum level is approaching the response to an increase of the treatment intensity will inevitably become little or null. Secondly, farms may encounter technical or environmental constraints in responding to the treatment. A marginal change in the output vector associated to a low treatment intensity may occur without costs while, on the contrary, a major change associated to higher treatment levels may imply significant adjustment costs and, consequently, the observed response turns out to be less than proportional. A final motivation for a concave response concerns the relationship between the treatment intensity, and how it is here measured, and the actual increase of freedom in farm choices (i.e, the real degree of decoupling of support). A high intensity, 50% of support on farm's GPV, for instance, does not mean that 50% of the farm activity (e.g. available land) is now free from the production constraint to the coupled support. The farm's

[33] See Bia and Mattei (2008, p. 362) for more details.
[34] When needed, the Box-Cox transformation is here adopted to restore a distribution that passes the statistical test of normality. See Bia and Mattei (2008, p. 362) for more details.

entitlements that give right to support may be actually associated to a very limited amount of farm activity (land). Thus, decoupling of this support frees only a limited part of the farm activity and the response to the treatment intensity will be limited, as well.[35]

5.2. The Problem of Multiple Outcomes

A second practical issue concerns the proper definition of an outcome variable (Y) for this kind of treatment. It is worth noticing that the empirical literature on the evaluation of CAP reforms, and FPR in particular, has focused on several possible indicators to assess the impact. In most cases, such indicators, and the consequent models, actually refer to some aggregate dimension rather than the micro (i.e., farm) level. Typically, changes in land use, labour or other input intensification/extensification, farms' income and welfare, supply volumes, commodity market prices (OECD, 2011). However, though to a different extent, all these outcome variables are just indirect consequences of the micro-level response to the treatment (the policy reform). The objective, here, is to perform TE identification and estimation by directly looking at this level, thus focusing on outcome variables that really represent the first-instance response of the farmer. When looking at this more proper dimension, a lesser amount of micro-level empirical reveals a more complex response. For instance, some works concerning the FPR's impact at the farm level (Renwick and Revoredo-Giha, 2008, for instance) show that, though decoupled, first pillar CAP support still acts as a cross-subsidisation of pre-existent farm activities and, consequently, the farm-level response in terms of production choice turns out to be lower than expected and lower than could emerge by looking at aggregate figures.

At this micro-level, the FPR is expected to affect production decisions by (re)orienting farmers' to market. The hypothesis is that decoupling leaves farmers free to adjust and reorient their production decisions given their individual characteristics and market conditions (i.e., prices).[36] Therefore, a proper outcome variable should express the degree of change in production orientation or mix. Finding a synthetic variable expressing such change in farm production choices, however, is not trivial. Within typically multioutput activities, production decisions are expressed by an output vector rather than by a scalar variable. For any element of this vector, the change in production choices can take a different form: to start producing a new (for the farm) agricultural product but also (in the case of a product that is already part of the farm's supply) to increase or reduce the amount of production of that particular good, to improve or not its quality level and so on. Moreover, whatever this change eventually is, its

[35] Following this argument, one could also conclude that a more predictable relationship between treatment intensity and response would be obtained by a more refined definition of the treatment intensity variable. A better measure should more strictly capture how much the farm activity and choices depend on support by strictly linking these entitlements to farm activity (e.g. eligible land). However, the very heterogeneous condition across farms and the limited available information in this respect make any alternative measure of the treatment intensity variable unfeasible. Eventually, the measure here adopted (total first-pillar support on farm's GPV) remains the more suitable, robust and less distortionary solutions.

[36] In more technical terms, the most significant impact expected from the FPR is to improve farm's allocative efficiency. See Moro and Sckokai (2011) for an exhaustive theoretical background on this aspect.

timing may be different. The introduction of a new perennial crop in the farm output vector (for instance, wine production) implies a long-term horizon; in such case, what we observe in the short-term, is just an investment decision. On the contrary, the introduction (or a larger production) of an annual crop (for instance, durum wheat) operates in a short-term horizon and can be directly observed in terms of higher cultivated area or higher revenue or higher input expenditure related to that specific crop.

To take this multiple nature of the farm-level production response to FPR into account, different outcome variables are considered. We can divide them in two typologies. The first type of outcome variable is a synthetic (scalar) measure of the change in the supply vector (that is, in the shorter-term production decisions) between two years or periods. The second type considers the investment behaviour, that is, production decisions oriented towards a longer-term programming horizon.

In the first typology, the outcome variable expresses the distance between two output vectors. This distance is computed using two different metrics:

$$(26) \qquad y_i^1 = \sqrt{\sum_{k=1}^{K} \left(s_{ik,B} - s_{ik,A}\right)^2}$$

$$(27) \qquad y_i^2 = \sum_{k=1}^{K} \frac{\left|d_{ik,B} - d_{ik,A}\right|}{N}$$

where *k=1,....,K* indexes the k-th product within the vector of potential production activities, s_{ik} expresses the share of the k-th commodity on the total revenue of the i-th farm, d_{ik} is a dummy variable taking value 1(0) if the k-th product is (is not) produced by i-th farm. Finally, indexes A and B express the two points in time when these variables are observed. Typically, A = pre-treatment year, B = post-treatment year. y_i^1 and y_i^2 are just distance variables: the former is an Euclidean distance, the latter is a variant of a conventional similarity index.

$0 \le s_{ik} \le 1$, y_i^1 varies between 0 and 2, with the lower value taken by farms whose revenue distribution across potential products remains the same between the two years/periods. In such case, no change in production decision is observed over time. The maximum value, on the contrary, is taken by those farms that concentrate all revenue in only one product and this unique product changed between the two periods. Therefore, this outcome variable not only accounts for the change in production decisions between the two years/periods but also for the degree of specialization of the given farm.

As d_{ik} is a dummy variable, it is $0 \le y_i^2 \le 1$. Even in this case, the outcome variable increases the more the output vector changes. The 0 value is taken by farms for which all productions observed in A are confirmed in B and no other activity is added. In this case, however, specialization does not tend to increase the value of the outcome variable as, on the contrary, an higher value is observed for those farms that change their production activities over a large range of products. It must be also noticed that this second outcome variable does not take into account the different relevance (share) of a given k-th production in the i-th farm revenue. Therefore, it is not able to take into account changes in production decisions that take the form of an extension (reduction) of an activity over a continuous domain.

Apparently, therefore, y_i^2 is a less accurate measure of the treatment outcome than y_i^1. This latter, however, may encounter a major drawback because revenue shares s_{ik} does not only depend on farmer's production decisions but also on market prices. Prices may not only be independent on the treatment but may be even unpredicted by producers. Under remarkable price volatility, therefore, the former outcome variable may overestimate the response of farmers to treatment by attributing to it an exogenous movement of prices.

Finally, we may also argue that, under market reorientation, the response of a farm's production choices measured with y_i^1 is higher than the response of the same farm measured with y_i^2. First of all, as mentioned, the former can take into account also a change in the production decisions that give more relevance to activities with higher market convenience even if no new activities is really added (therefore, y_i^1 can be >0 even when y_i^2=0). Secondly, the latter simply counts the addition or substitution of production activities. Under market reorientation these new activities are expected to show more market convenience (for instance due to a temporary positive price dynamics) than those they replace or the preexisting ones, therefore this addition/replacement impacts more on y_i^1 than on y_i^2. According to this interpretation, a positive TE of the FPR is expected to be revealed, in treated farms, not only by a response of both y_i^1 and y_i^2, but also by an higher response of y_i^1 than y_i^2.

The second typology of outcome variable consists of a scalar measure expressing the investment decisions taken in response to the FPR. The idea simply is that the treatment may induce extra (more than "business-as-usual") investments allowing the farm to activate (extend) new (existing) activities in the longer-term. Therefore, the outcome measuring such effect is simply the change in investment expenditures of the i-th farm (I_i) between years/periods A and B. This change is here measured in two different ways:

(28) $$y_i^3 = \left(I_{i,B} - I_{i,A}\right)$$

(29) $$y_i^4 = \left(\frac{I_{i,B}}{VA_{i,B}} - \frac{I_{i,A}}{VA_{i,A}}\right)$$

y_i^3 merely is the difference between the yearly total investment expenditure. y_i^4 expresses this difference not in absolute terms but in relative terms, that is, as investment rate given by the ratio between total investment expenditure and the respective farm value added.[37] This latter outcome variable may better capture the real investment effort of the i-th farm and get rid of the wide size

[37] The value added rather the value of production is here considered because the former can be more properly considered a proxy of farm profits, that is, of the capacity to generate surplus from which further investments can be undertaken.

heterogeneity among farms both in physical and economic terms. In doing this, however, y_i^4 may partially sterilize the effect of the treatment on investment decisions over the whole sample. In fact, a real increase in investments concentrated in larger farms may be entirely compensated by a decline in investment rates in smaller farms.

In practice, all these four outcome variables present pros and cons. The first typology only partially captures the farms' production response to the treatment. At the same time, as already mentioned, one possible problem in using investment decisions as outcome variable in the present case is that both anticipation and lagged effects may occur. For instance, a pre-treatment year/period could already contain some anticipated investments response of the farmers to the FPR and this makes the identification and estimation of the TE more complex and, consequently, results less reliable and robust. Therefore, all outcome variables (26)-(29) represent a relevant but incomplete dimension of the production response to the FPR. Actually, they are more complements than alternatives in providing a comprehensive picture of the reorientation to market. For these reasons, all the four outcome variables will be used throughout the present empirical study.

A final issue in the proper definition of the outcome variable, concerns the selection of observations/years A and B. In principle, several couples of years/periods could be compared to compute the outcome variables in (26)-(29). Years 2003 and 2004 unquestionably represent before-reform (thus, before-treatment) years as the implementation of the reform started in 2005 in Italy. At the same time, years 2005, 2006 and 2007 can be considered as after-reform (after-treatment) years. As a consequence, the following pairs of years can be candidate for a before and after-treatment comparison: 2004-2005, 2004-2006, 2004-2007, 2003-2005, 2003-2006 and 2003-2007. However, the choice here is to consider years that are symmetric with respect to the treatment year (2005) and are far enough from it to exclude (or minimize) anticipated or lagged responses. For this reasons, the outcome variables are here computed assuming A=2003 and B=2007.[38]

[Figure 1 here]

5.3. The Problem of Multiple Treatment

A further complication in identifying and estimating the TE of the FPR comes from the fact that, additionally to CAP first pillar measures, support to farms may be delivered through second pillar measures. Not only there are many possible second pillar measures administered to any given farm, but this latter may receive, at the same time, both first and second pillar measures. In other words, the CAP (both before and after the FPR) potentially is a multiple-measure policy. Therefore, an appropriate TE assessment should not disregard this aspect. Figure 2 reports, for the 5 years under consideration, the percentage distribution of the total support within the sample among the most

[38] Evidently, different years will be considered in the DID estimation. See section 6.2.2 for more details.

significant (those with more than 1% on total support) measures.[39] Even if we exclude the negligible measures, the fragmentation of the support remains, especially in the second pillar, even after the introduction of the simplifying SFP scheme with the 2003 reform. Moreover, though OCM (before FPR) and SFP (after FPR) payments are clearly dominant, some second pillar measures show a considerable overall expenditure. Apparently, a so complex multi-treatment setting is hardly empirically workable (Frölich, 2004).

To simplify the approach, however, we can treat Pillar I measures as an unique aggregate, though articulated in different measures before the FPR, then with the unique common objective after the FPR (market reorientation). Simplifying the scene with respect to second pillar measures is more challenging. It must be acknowledged that farms' production decisions, i.e., the four outcome variables (26)-(29), are somehow affected by these measures. By supporting competiveness and structural adjustment, Axis 1 measures may directly imply investments or production choices that are themselves oriented toward allocative efficiency thus possibly overlapping with the impact of the FPR.[40] Other measures (especially in Axis 2), on the contrary, may actually represent constraints to a rapid adjustment towards a more efficient output vector. In principle, first and second pillar measures may reciprocally interact (positively or negatively, that is, reinforcing or reciprocally offsetting) with respect to the expected outcome such that a simultaneous treatment can generate a different effect compared to the two separate treatments.

Nonetheless, despite the fact that we can not exclude that second pillar measures may interfere with the TE of the FPR, two aspects are worth reminding to properly deal with these multiple-treatment case. First of all, the RDP was reformed in 2005 and such reform was implemented only in 2007. On the one hand, it did not radically change the way the support is delivered and its fundamental objectives, axes, measures and actions are partially analogous between the two programming periods (2000-2006 vs 2007-2013) (Esposti, 2011b). On the other hand, as we are using farm-level 2003-2007 data, here we do not really observe the new (2007-2013) RDP in action. Even in 2007, farm data mostly report second pillar funding that still refers to measures of the former period (2000-2006). So, in practice, there is no second pillar treatment corresponding to the FPR treatment, as no policy regime change really occurs. We only have to take into account that some treated farms, i.e., farms that experienced FPR starting from 2005, also received at least one second pillar payment over the whole period of investigation (2003-2007), while other farms did not. Moreover, the second pillar support was never coupled to particular commodities and, at least apparently, its aim has never been the 'market orientation' that drove the FPR. Therefore, the impact of second pillar support on the FPR's outcome is an involuntary side-effect of its implementation.

[39] See Annex (Table A1) for a description of the measures reported in Figure 2.

[40] "Measures relating to structural adjustment of farming […] enhancing the economic viability of agriculture through investment and modernisation" (European Commission, 2011, p. 8).

Following these arguments, it is reasonable here to argue that these two policies are currently implemented and delivered almost independently, following different selection procedures and aiming at different objectives. Therefore, we are not really in a case of multiple TEs identification and estimation with all consequent complications (Frölich, 2004; Imbens and Wooldridge, 2009; Chabé-Ferret and Subervie, 2013). In fact, in the present case only one TE is of interest (that of FPR)[41] and, while it can not be excluded that second pillar support may affect FPR assignment and outcome, the other way round can be ruled out, that is, the assignment to second pillar support is not affected by FPR assignment and outcome. As a consequence, here the second pillar support can be considered just as one of those pre-treatment covariates in **X** that affect the identification and estimation of the TE of the FPR by influencing both assignment and outcome but are independent from it.

[Figure 2 here]

5.4. The Cofounding Factors

The confounding factors here considered are those pre-treatment variables (covariates **X**)[42] expected to incorporate all the relevant aspects that may affect the production choices before the treatment (thus, affecting the outcome regardless the treatment itself) as well as the treatment assignment. We selected these variables to capture, with the minimum redundancy, three different types of factors (Table 3).

First of all, we consider the relevant *individual characteristics* of the farmer (AGE) and of the farm (Altitude - ALT). Secondly, the economic (ES, FC)[43] and physical (AWU, HP, UAA and, at least partially, LU) *size* of the farm clearly matters. All these variables evidently affect the outcomes but presumably are not directly correlated with the treatment assignment. Still, they are definitely linked to production choices and, since pre-treatment production choices are unquestionably correlated with the treatment assignment, this correlation indirectly occurs even with respect to these first two categories of covariates.

The third typology of confounding factors, in fact, consists of those variables (TF and, in part, LU) that directly express the *production specialization* of the farm. The linkage between these covariates and the treatment assignment is evident as this actually concerns those farms that were interested by specific OCM measures while, on the contrary, farms not involved in first pillar are those whose production specialization was less (or not at all) targeted by specific policy measures. To express farm production specialization, the 4-digit "Type of Farm" (TF) FADN classification is adopted (2000 classification).[44]

[41] A similar case is presented in Chabé-Ferret and Subervie (2013).

[42] Pre-treatment variables have been observed in 2003, the only exceptions being FC, for which the 2003-2004 average has been considered since this variable may largely vary on a yearly base, and the dummy RDP.

[43] The relative (with respect to net value added) amount of fixed costs expresses the importance of fixed factors (especially labour and physical capital) within the farm and, therefore, it is a proxy of the scale of the farm business itself.

[44] By itself, however, this qualitative variable is not suitable in this empirical exercise as it has not a monotonous linkage with the treatment assignment. For instance, class 4210 (beef production) is more dependent on first pillar support than classes 2022 (flowers' production) and 6010 (horticulture); therefore, farms belonging to the former class are more likely to be

A final confounding variable included in the analysis is the already mentioned dummy expressing second pillar support (RDP). The dummy takes value 1 if the farm received at least one second pillar payment over the whole period of investigation (2003-2007), 0 otherwise. In the original balanced 2003-2007 sample very few farms (less than 30) were non-treated cases but still received second pillar support. Therefore, finding a common support and satisfying the balancing condition is going to be very difficult when this RP dummy is included among covariates. For this reason, the adopted sample of 6542 farms actually excludes those very few non-treated farms receiving second pillar support. More generally, identifying the contribution of the RP dummy to both the treatment assignment and the outcome is particularly difficult in the binary case because this variable actually corresponds to the treatment. Consequently, the RD dummy is here considered as one of the variables in **X** only in the multivalued treatment.

6. THE EMPIRICAL APPLICATION

6.1. Comparing treated and non-treated (control) units: descriptive evidence

Table 1 illustrates how the sample farms distribute across the FPR treated and non-treated groups, also distinguishing those farms that received, in addition, second pillar payments. As mentioned, most farms in the sample received the FPR treatment (83%). Most of them (56% on total farms) also received second pillar support during the period. Therefore, all the abovementioned group comparisons are inherently unbalanced, with the number of control units being different (much lower) than the number of treated units.[45] The treatment level (i.e, the amount of payments) strongly varies within the treated sample. This heterogeneity evidently depends on the large farm size (either in physical or economic terms) heterogeneity. If we measure the treatment level in terms of treatment intensity (*TI*), this heterogeneity sharply falls but it remains significantly high, as also emerged in Figure 1.[46]

The overall support per farm increases by about 17% from 2003 and 2007 in nominal terms, though this increase is lower than 10% in real terms. The dispersion of support, on the contrary, is reducing though remains remarkably high. This reduction of variability can be partially attributed to the decoupling of first pillar support that, in fact, stabilized the pre-reform differences among farms. The group of treated farms with also second pillar support receives a larger first pillar support, on average. This difference increases over years and can be explained by the fact that farms of this group show a larger size on average, thus first pillar support is itself expected to be larger.

---

assigned to the treatment than farms of the latter classes. To overcome this problem, the official TF classification has been reclassified by assigning to any 4-digit class a number (ranging from 1 to 7) expressing its dependency on first pillar CAP support. This number expresses a qualitative monotonous variable (TF_R) that increases as the dependency on CAP support declines. The Annex (Table A2) details this reclassification of the TF variable.

[45] This justifies the fact that an appropriate matching methodology to estimate the ATT might require comparison with replacement (see below).

[46] For instance, the coefficient of variation of *TI* (support on farm's value of production) is 1.9 in 2003 and 2.4 in 2007.

Table 2 reports some descriptive statistics by treatment group for the four outcome variables. It may be easily appreciated that, comparing the two extreme years 2003-2007, for all outcome variables the average values tend to increase moving from the non-treated (or control) group to the treatment group. Moreover, within the treatment group, farmers also receiving second pillar payments show an additional positive impulse. Nonetheless, if we look at comparisons between other couples of years, the picture becomes less clear. While y_i^1 and y_i^2 tend to confirm higher values, more mixed evidence emerges for y_i^3 and y_i^4 especially when years considered are those around the treatment year (2005). This can be explained by the presence of anticipation and/or lagged effects, but another explanation could simply be that the observed differences in outcome variables are not caused by the FPR.
In fact, the most relevant evidence emerging in Table 2 is the high variability of all outcome variables in the whole sample, as well as in treatment groups. In practice, if we constructed a conventional 95% confidence interval around the sample averages we would notice that these intervals are largely overlapping across the groups for all outcome variables. More generally, though these differences in outcome variables are mostly consistent with the expectations in terms of policy TE, these simple statistics can not be conclusive on the fact that such differences across groups can be indisputably attributable to FPR.
Table 3 reports some descriptive statistics of the pre-treatment covariates (or confounding variables) considered in the present analysis. Even in this case, for most variables the dominating evidence concerns the large variability observed in both the whole sample and in treatment groups, and this prevents from clear-cut statements about structural differences across the groups. Only for few variables a difference between treatment groups' clearly emerges. In particular, non-treated units tend to show some peculiarities compared to the treated ones while the difference between the treated units with and without second pillar payments seems mostly negligible. Non-treated units show a smaller physical size (UAA) but this is not necessarily true if we consider the economic size (ES). Moreover, as expected, the production specialization of the non-treated group is evidently less dependent on first pillar support (TF_R), it practically excludes livestock activities (LU) while it favours activities mostly run in flat areas (ALT). The immediate interpretation is that most of these non-treated units are farms with small area and high output values strongly specialized in a certain kind of production (e.g., horticulture) largely disregarded by coupled first pillar payments. This reinforces the idea that these farms might not be reliable as control units in identifying and estimating the TE of the FPR.

[Tables 1-3 here]

6.2. Results[47]

6.2.1. PSM estimation

6.2.1.1. The estimated Propensity Score (PS)

Table 4 reports the estimates of the PS equation (5). The specification adopted is a conventional binomial probit with the treatment assignment as the dependent binary variable and the covariates as explanatory variables. After the PS estimation, the balancing condition has been checked, and always satisfied, on the common support.[48]

Parameter estimates of the probit equation are all statistical significant with the only exception of the parameter associated to variable FC. This can be reasonably interpreted with the fact that fixed costs might be strongly collinear with other explanatory variables, in particular those expressing the amount of labour (AWU) and capital (HP). Older farmers seem to have an higher probability to be treated (i.e., to receive the first pillar support). The altitude operates in the same direction as the propensity to receive the first pillar support increases moving from plain areas to hilly and mountainous farms. Evidently, farms excluded from first pillar support (therefore from the treatment) due to their production decisions, are positioned in relatively more favourable areas. These non-treated farms tend also to be more intensive as the propensity to receive $T=0$ declines with the amount of working units. The increasing substitution of agricultural labour with physical capital (machinery), as indicated by the parameter associated to HP, is evident moving from the non-treated ($T=0$) to the treated ($T=1$) group.

To partially confirm this effect, the propensity score is positively affected by the physical size of farms (UAA), i.e. larger farms tend to have an higher propensity to receive the first pillar payments. However, this effect is actually of limited magnitude and may also be misleading. As a matter of fact, whenever the economic size (ES) is taken into account, the propensity score actually decreases with the size. We can argue that farms with larger size (therefore arguably more professional farms) show a dualistic attitude towards the policy treatment in consideration here. On the one hand, we have some labour intensive farms of relevant economic size and prevalently positioned in plain and well-endowed areas that tend to be excluded from first pillar support due to their specific production specialization. On the other hand, we find capital intensive farms with large physical size and sometimes positioned in less favoured hilly and mountainous areas for which we observe an higher probability to receive first pillar payments.

Variables expressing production specialization (LU and, above all, TF_R) may explain this apparent dualism. They confirm that the prevailing production decisions eventually bring about self-selection in

[47] All estimations have been performed using software STATA12.

[48] The balancing condition is tested by stratifying farms (within the common support) in blocks of equal $p(\mathbf{X})$ range. Whenever balancing is not found, the number of blocks is adjusted until balancing is satisfied. The common support consists in that range of variation of $p(\mathbf{X})$ where we can observe both treated and no treated farms. Limiting the balancing check to the common support is justified by the fact that the following matching and ATT estimation excludes the units that are outside the common support.

terms of treatment assignment. Moving towards TF with a lower dependency on the CAP is associated to a significantly lower propensity to receive the treatment. This evidence also suggests that the way TF variable has been reclassified is appropriate. The role of production specialization is also confirmed by LU variable suggesting that livestock activities tend to be more present in the treated group. In general term, more than any other structural or idiosyncratic characteristic of the farm, production orientation and specialization seems to be the factor that primarily induces selection of farms in terms of treatment assignment.

[Table 4 here]

6.2.1.2. Matching and the estimated Average Treatment effect on the Treated (ATT)

Table 5 reports the PSM estimation of the *ATT* in (6) according to the four alternative matching procedures presented in section 3.1.2. The results are juxtaposed to facilitate comparisons and, thus, to assess their robustness with respect to how matching is performed. Stratification Matching has been performed using the blocks identified in PS equation estimation. Nearest Neighbour Matching is obtained sorting all units by the respective estimated propensity score, and then searching forward and backward for the closest control unit(s). In the case of multiple nearest neighbours, either the forward or backward matches are randomly drawn.[49] Radius Matching is performed by taking a 0.05 radius of the PS, while for Kernel Matching the conventional Gaussian Kernel function is adopted.[50] As illustrated by Becker and Ichino (2002), analytical standard errors can be obtained only for some of these *ATT* estimates. For the sake of comparison, however, here standard errors of the *ATT* estimates are always computed through bootstrapping with 1000 replications.

Some regularities among these estimated *ATT*s emerge. First of all, in the case of outcome variables y^1 and y^2, results are quite robust across the four matching procedures. Moreover, for these outcome variables *ATT* estimates are in most cases statistically different from 0. On the contrary, estimates are more variable and often not statistically significant in the case of the outcome variables related to investment decisions, that is, y^3 and y^4. Secondly, estimates obtained with Kernel and, above all, Radius matching tend to be systematically higher than *ATT* estimated with the other two matching procedures. Radius matching estimates are also dependent on the definition of the radius as a lower radius may significantly change the results.[51] Limiting the attention to statistically significant *ATT* estimates, all the results are concordant in sign and similar in magnitude. Therefore, common robust conclusions can be drawn from the four different matching approaches.

---

[49] An alternative procedure consists in taking into account, and equally weight, both forward and backward matches. These alternative estimates do not substantially differ from what presented here and are available upon request. More technical details on these PSM procedures can be found in Becker and Ichino (2002).

[50] The alternative Epanechnikov Kernel function has been also tested by imposing a bandwidth of 0.05. These alternative estimates do not substantially differ from what presented here and are available upon request.

[51] Estimates under alternative values of the radius are available upon request.

The (binary) treatment under consideration here (FPR) always induces a change in the output vector and, less clearly, induces investment decisions. If we consider the average value of y^1 across the treated group, the estimated ATT implies a quite strong (>50%) impact. Moreover, the estimated *ATT* is very robust across the different matching approaches ranging between 0.37 and 0.43. This latter conclusion holds true even more evidently for y^2 whose response ranges between 0.09 and 0.10. In relative terms, this impact of the FPR is even stronger as it amounts to almost 60% of the observed average value within the treated group. The estimated ATT indicates an impact of the FPR on production choices of relevant magnitude also on the two outcome variable expressing investment decisions (y^3 and y^4) but both the low statistical significance and the limited robustness across alternative matching procedures prevent from a conclusive empirical evidence in this respect. Though positive, this ATT seems affected by the strong heterogeneity (variability) observed for these variables across both treated and non-treated farms. In such production decisions, as mentioned, their timing and the timing of the treatment plays a decisive role. Anticipation effects, as well as lagged effects due to financial and technical constraints preventing an earlier investment response to the treatment, make the identification and estimation of the ATT harder even if a relevant and positive impact actually occurred.

The policy interpretation of these results brings us back to the original research questions of the present paper. We can state that results support the idea that the FPR really oriented farms to market as induced the treated farms to a stronger change in their output composition. However, the actual nature of this response to the treatment more clearly emerges by comparing the results for the four outcome variables. On the one hand, the response of the revenue composition (y^1) is more relevant than the response of the output composition (y^2). Apparently, and this is fully expected as a consequence of market reorientation, treated farms reorient their activity to market by changing their output mix and this changes even more strongly the revenue composition as the new activities are supposed to be those with higher market convenience (for instance due to a temporary positive price dynamics) and adds new important sources of revenue to the preexistent farm's core activities. This is even more true if the farm's response in terms of production choices does not consist in adding new activities (and quitting previous ones) but just in giving more space to existing activities showing higher market convenience (thus increasing their revenue share). On the other hand, the impact on investment decisions is much less evident thus suggesting that the FPR convinced the treated farms to change their short-run production decisions but did not concern as much the long-run production choices.

This apparently limited scope of the farm response to the FPR may somehow depend on the fact that most farms opted for conservative strategies and reduced their investment level over the period under study, regardless the policy treatment (see Table 2 in this respect). In a period of persistent market crises or difficulties, many Italian farms suspended production decisions with medium and long-term implications regardless the change in the policy regime. The conclusion could be that the capacity of the FPR to orient farms to market is substantially reduced whenever major production changes (and consequent adaptations) are potentially costly and whenever farms tend to assume a "wait-and-see" attitude both due to market uncertainty and, possibly, to further policy signals confirming (or denying) the direction taken by the CAP first pillar support.

[Table 5 here]

6.2.1.3. Robustness of the PSM ATT estimations

ATT estimates obtained through PSM seems to provide a positive answer on the capacity of this estimation approach to indentify that part of the observed difference between treated and non-treated farms that can be really attributed to the treatment. Not only for the positive and statistically significant ATT for y^1 and y^2, but also for the capacity to distinguish the response to the treatment in terms of short-term production choices (y^1 and y^2) and of medium or long investment decisions (y^3 and y^4). Nonetheless, one could still argue that this empirical evidence on the FPR ATT is not really due to the treatment but to the fact that the control units are not appropriate counterfactuals. As already mentioned, even though matching estimation is conditioned on a large set of observables, the balancing condition is satisfied and matching is limited to the common support, this only ensures against the *selection-on-observables bias* but, evidently, can do nothing against the *selection-on-unobservables bias*. It could be argued that the difference between treated and non-treated farms, especially in terms of many unmeasurable aspects (personal beliefs and attitudes, family histories and traditions, just to make few examples) can not be entirely taken into account by the set of conditioning factor $\mathbf{X}$. As a consequence, the difference between treated and non-treated units, also when conditioned on $\mathbf{X}$, depends on these unmeasured aspects and not only on the treatment.

Even if this *selection-on-unobservables bias* is not entirely ruled out, however, it does not mean that the estimated ATT is meaningful or misleading. The bias can be small or irrelevant and, therefore, the estimated ATT maintains its validity. The Rosenbaum (1987; 2002) bounding approach assesses how strongly an hypothetical unmeasured disturbing variable must influence the selection process to undermine the results of matching. By itself, this bounding approach does not test for the presence of a selection bias.[52] Instead, Rosenbaum bounds provide evidence on the degree to which the significance of a matching result depends on an hypothetical unmeasured disturbing factor, that is, on the untestable CIA. Here, we adopt the DiPrete and Gangl (2004) implementation of the Rosenbaum intuition. Table 6 reports both the Wilcoxon signrank tests, providing upper and lower bound estimates of significance levels at given levels of hidden bias (the Γ parameter) and the corresponding Hodges-Lehmann point estimates and confidence intervals. These statistics are computed for the to the Nearest

[52] This would mean testing that no (unobserved) variable influences the selection into treatment.

Neighbour Matching ATT estimate for all the four outcome variables. Results suggest that in the case of y^1 and y^2 only a very high hidden bias (that is, Γ=4) would bring the estimated ATT statistically to 0. On the contrary, as already noticed, the estimated ATT for y^3 and y^4 are much weaker and a relatively small hidden bias (Γ=2) is enough to make the ATT null or not statistically significant.[53]

The robustness of the ATT estimation for y^1 and y^2 even in the presence of an hidden bias is confirmed by Table 7 showing the sensitivity analysis performed according to the approach proposed by Nannicini (2007) and Ichino et al. (2008). The table reports the ATT estimation under an "artificial" variable that restores the CIA assumed it is not valid under the observed covariates $\mathbf{X}$. The comparison of the estimates obtained with and without matching on this "artificial" variable (the confounder) shows to what extent the original results are robust to specific sources of failure of the CIA (Nannicini, 2007). The statistical distribution of this confounder across the sample is simply expressed by the two parameters *d* and *s*.[54] These parameters can be identified following two different criteria. The first consists in calibrating them in such a way that the distribution of the confounder across the sample follows the distribution of one observed binary variable. Here, these parameters have been calibrated by assuming that the confounder has a distribution corresponding to the binary variable obtained by distinguishing two different levels of economic size (ES≤6; ES>6). Alternatively, *d* and *s* can be simulated to obtain increasing underlying effects (outcome and selection effects) and different combinations between the two. In particular, such simulation allows identifying that combination of *d* and *s* that "kill" the estimated ATT, i.e, that make the ATT statistically null. How large *d* and *s* must be to "kill" the estimated ATT provides an evidence of its robustness.

Results show that, in the case of y^1 and y^2 (and, at least partially, for y^4), the estimated ATT obtained though matching on $\mathbf{X}$ and on the confounder does not differ much with respect to the original results (Table 5) unless the simulated confounding factor is pushed up to a very strong violation of the CIA, that is, to generate a very high impact on both the selection into treatment (*selection effect*) and on the outcome of non-treated units (*outcome effect*). On the contrary, for y^3 a relatively mild confounding factor (and, consequently, minor *outcome* and *selection effects*) substantially change the estimated ATT.

Eventually, though performed with different empirical strategies, this robustness analysis is quite concordant. One may still argue that, in the case of the FPR, the ATT PSM estimation is still affected by some unobservable variable and, therefore, by an hidden bias. Nonetheless, results suggest that the

---

[53] It is worth reminding that Γ=4(2) implies that there is an unobserved covariate that quadruples (doubles) the probability of receiving the treatment for control units matched with treated units, that is, units having statistically identical observed covariates $\mathbf{X}$.

[54] Parameter *d* can be interpreted as a measure of the effect of the confounder on the outcome of non-treated units; *s* as a measure of the effect of the confounder on the selection into treatment. These two effects can be expressed as average odds ratios of the simulated confounder and are also called "outcome effect" and "selection effect", respectively (Nannicini, 2007) and Ichino et al., 2008).

estimated short-term production response to the treatment (y^1 and y^2) is fairly robust as it remains positive and significant unless the hidden bias becomes particularly and implausibly high.

[Tables 6-7 here]

6.2.2. DID estimation

The results of Tables 6 and 7 can be interpreted in favor of the robustness of the PSM ATT estimates even when the CIA does not hold true and an hidden selection-on-unobservable bias occurs. Nonetheless, in principle, one way to get rid of such hidden bias is the DID estimation. However, as mentioned, such solution requires several years of observation and, at the same time, a decision has to be taken on the proper couple of observations (before and after treatment) to be used. Comparing single-year observations on the outcome variable *Y* and attributing all the observed difference to the treatment can be, especially in agricultural production, hazardous and generate not robust estimations. More robust evidence can be obtained by comparing several couples of years/periods.

Here, years 2003 and 2004 unquestionably represent before-reform (thus, before-treatment) years as the application of the reform started in 2005 in Italy. At the same time, years 2005, 2006 and 2007 can be considered as after-reform (after-treatment) years. Consequently, the following pairs of years can be candidate for a before and after-treatment comparison: 2005-2004, 2006-2004 and 2007-2004, but also 2005-2003, 2006-2003 and 2007-2003. However, as in the present case the outcome variable is a year-by-year variation, the only before-treatment observation (or *baseline*) is the 2003-2004 variation, though, in principle, also the 2006-2005, 2007-2006 and 2007-2006 variations are outside the treatment. The available after-treatment observations (or *follow-up*) are the 2005-2004, 2006-2004 and 2007-2004 variations. Following the Chabé-Ferret and Subervie (2013) suggestion to perform a *symmetric DID estimator*, the follow-up here considered is the 2005-04 variation.

At the same time, an alternative comparison is considered to take two aspects into account. First of all, to allow enough time to adopt production decisions adjusting to the treatment (especially for investment decisions) the 2006-04 variation is considered as follow-up observation. Secondly, to exclude year 2003 with its peculiar agricultural performances especially due to extreme weather conditions in Italy (Esposti, 2011), the 2006-07 is considered as baseline observation. Therefore, the DID estimation is here performed for two couple of comparisons: 2004-2003 and 2005-2004; 2007-2006 and 2007-2004. In addition, the DID estimation is also performed for the couple of variations 2004-2003 and 2007-2006. Both are expected to be baseline (or no-treatment) observations. Consequently, this latter DID estimation is expected to behave as a placebo testing procedure: a significant ATT would indicate that the CIIA does not hold true at least for the years under consideration.

Table 8 displays the results of the DID ATT estimation for the two abovementioned comparisons. The unconditional and the conditional (PSM included) DID estimates are reported (i.e., (7), (10) and

(11)).[55] Two preliminary remarks are worth noticing. First of all, the DID estimation confirms that for outcome variable y^3 the estimated ATT is never significant. On the contrary ATT estimation is statistically significant in one comparison for both y^2 and y^4 and it is always statistically significant for y^1. Apparently, this would confirm what already emerged in the case of PSM ATT estimation: the response to the treatment is appreciable in terms of short-term changes of the output mix composition, while it becomes less evident in terms of changes in the revenue composition and in the longer term production choices (investments). The second preliminary remark concerns the fact that conditioning the DID estimation on $\mathbf{X}$ (or $p(\mathbf{X})$) does not change much and, above all, does not significantly improve the statistical goodness of the results. This may be interpreted as an indirect evidence in favour of relatively robust ATT estimates with respect to the possible selection-on-observable bias.

What really surprises in DID estimates reported in Table 8, however, is the patent difference emerging between the two comparisons, that is, between 2004-03(follow-up) on 2004-03(baseline) and 2007-04(follow-up) on 2007-06(baseline). Limiting the attention to the statistically significant result, in the former case a negative and counterintuitive ATT is obtained for y^1 and a positive ATT for y^2. In the latter comparison, on the contrary, the estimated ATT for y^1 is positive, as expected, and also a positive ATT for the investment rate, y^4, is observed. The little robustness observed in the DID estimates across different before and after-treatment periods may indicate, in fact, that time strongly affects the outcome variables here considered and this effect of time is evidently not controlled by the set of covariates and differs between treated and non-treated units.

In practice, this lack of robustness could be interpreted as an evidence against the validity of the CIIA in the present case due to the year-by-year strongly unpredictable and highly differentiated (across crops, territories, types of farm and farming) variations in market and environmental (e.g., weather) conditions. This plausible interpretation can be explicitly assessed through a placebo test (Chabé-Ferret and Subervie, 2013). It consists in assessing whether a significant DID ATT estimation persists even when the comparison is made between two baseline (no-treatment) periods (comparison between 2007-06 and 2004-03 in the present case). In such circumstance, since the observed ATT can not be directly attributed to the treatment we should conclude that the CIIA is violated.

Table 9 reports the results of this test and clearly shows that the present DID estimates (unconditional or conditional) fail to pass the test at least for outcome variables y^1 and y^2. There is a placebo effect as a significant ATT is observed when, in fact, it should not. Moreover, results are very similar to the case reported in Table 8 with 2004-03 as a baseline and 2005-04 as a follow-up. The former baseline, which is common in the two cases, evidently affect so much the DID results to overcome the possibly

[55] In the conditional DID estimation the balancing condition is assessed and always respected. PS estimation and matching is perfomed as in the PSM estimation presented above adopting the Kernel matching. The coefficient estimates of the PS (probit) function are available upon request.

real effects of the treatment. This might be attributed to some very peculiar figures emerging for year 2003 but, more generally, the placebo test is not passed even when the baseline does not include year 2003.[56] This confirms how critical the choice of years to be compared can be in DID estimation especially when agricultural data, typically showing strong year-by-year variations independent on farmers' choices, are under study.

Such kind of evidence in an agricultural context is not new and is considered as a clear demonstration of the violation of the CIIA in such circumstances. This is possible because treated and non-treated units tend to be structurally different in terms of production orientation and specialization. Therefore, it is almost unavoidable that year-by-year external variations (from weather to market conditions) are strongly differentiated across these diverse orientations and specializations. In addition, however, as emphasized by Chabé-Ferret and Subervie (2013),[57] the apparent violation of the CIIA can be rather attributed to the elusive timing of the response to the treatment, that is, to the presence of anticipation effects (potentially occurring in the 2004-03 baseline) or of lagged effects (potentially occurring in the 2007-06 baseline). In any case, whatever the cause of the failure of the placebo test is, the conclusion remains the same: DID estimation (conditional or not) in the present case shows little robustness as it too strongly depend on the years choosen to define the baseline and the follow-up observations.

It is here worth noticing that a further reason why ATT DID estimates (but the same argument could also apply to ATT PSM estimation) show little robustness may depend on the fact that the response to the treatment is largely heterogeneous (in magnitude and sign) across farms. Estimating the average treatment effect, therefore, may be misleading as other parameters (or statistics) expressing the statistical distribution of the treatment effect on the treated may be more informative. The median treatment effect, for instance, could be more robust than the ATT whenever many outliers (both treated and non-treated) can be observed within the sample. To investigate this aspect, Table 10 reports the PSM DID estimation of the Quantile Treatment Effects on the Treated (QTT). These are the ATT estimated on the quantiles of the four observed outcome variables. Three quantiles are here considered, the 25%, the 50% (the median), the 75% quantiles. Comparing these QTT it is possible to conclude whether the estimated ATT results from a combination of very heterogeneous and offsetting responses over sub-samples or it is regularly observed across all sub-samples.

Results reported in Table 10 would confirm that for the investment decisions (variables y^3 and y^4) all the estimated QTT are inconclusive (not statistically different from 0) in all quantiles. On the contrary, the response in revenue composition, y^1, monotonically varies across quantiles but it always remains positive or negative according to the adopted comparison as already observed in Table 8. Eventually, a first group of farms does not respond at all to the treatment while in the case of the other quantiles the response becomes increasingly larger though in opposite direction and with a much

[56] These further placebo tests are available upon request.

[57] "We find some evidence that the common trend assumption may not hold in our data" (Chabé-Ferret and Subervie, 2013, p. 18). The authors interpret this as a consequence of anticipation effects.

larger effect in the 75% quantile when the baseline period is 2004-03 and the follow-up is 2005-04. For output composition, y^2, we observe a similar behaviour with the ATT becoming significant only in the higher quantiles but always showing, when significant, a positive response.

This evidence of a differentiated response across the treated farms, however, should not surprise. One easy explanation is that it depends on the different treatment intensity these farms receive. This aspect can not be evidently investigated within a binary treatment approach. A proper analysis of such heterogeneous response necessarily requires acknowledging the multivalued nature of the treatment and adopting the consequent appropriate TE methodologies.

[Tables 8-10 here]

6.2.3. Multivalued ATE estimation

Under a multivalued treatment, one could easily estimate the ATE within a classical regression specification like (12). In such case, the estimated α or, more generally, the estimated first order derivatives of $E\left[(Y_i | T_i, \mathbf{X}_i\right]$ with respect to *T* would indicate the ATE. However, not only this estimate potentially incurs the selection bias. It also assumes an homogenous (i.e., equal across farms) linear relationship between Y_i and T_i while it is arguably non-linear and heterogeneous across treated units.[58]

Though acknowledging, as discussed, a likely monotone concave response, for the sake of comparison the lower part of Table A3 (in the Annex) still reports the OLS estimate of (12). Results would suggest a significant ATE for y^1, y^2 and y^4. However, the sign of this effect is not consistent with the expectation. The ATE on y^1 is negative while it is positive but almost negligible in magnitude for y^2, thus suggesting no influence of the FPR on short-term production choices, i.e., no market reorientation. On the contrary, the estimated TE parameter is positive for y^4 and this would indicate that the impact of FPR on production choices is only observed on longer-term decisions which is, in fact, at odds with what obtained in the ATT estimation under binary treatment.[59]

The two multivalued treatment methodologies discussed in section 3.3 are expected to overcome these limitations through nonparametric and, above all, consistent ATE estimates varying, without constraints, over the whole rank of observed *TI*.

[58] In fact, more flexible parametric specification than (12) could be adopted. It remains true that the implicit assumption underlying (12) is that the parameters expressing the response of the treated units to the treatment are constant regardless the treatment intensity.

[59] It is also worth noticing that similar evidence and analogous issues emerge also in the case of a specification of (12) where T_i is not expressed by the TI but by the binary treatment variable T = 0,1 (Table A3, upper part). Selection bias arises also in this case for the same reasons but even here results would suggest an highly statistical significant ATT for y^1, y^2 and y^4. Interestingly, though coefficient estimates associated to $\mathbf{X}_i$ are quite similar in the two specifications, with binary treatment all the significant estimated treatment effects are positive, thus consistent with expectations.

6.2.3.1. GPS estimation

The initial stage that is common to the two multivalued estimation methodologies here considered consists in the GPS estimation. Nonetheless, the specification of the GPS function, $GPS_i = r(T_i, \mathbf{X}_i)$ is necessarily different in the two cases. In the Hirano-Imbens approach, *T* (i.e., the *TI*) is a continuous variable and, therefore, $r(T_i, \mathbf{X}_i)$ is specified as a linear flexible functional forms, i.e., a fully interacted polynomial in both arguments, T_i and $\mathbf{X}_i$. In the Cattaneo approach, on the contrary, *T* (*TI*) is a categorical variable thus $r(T_i, \mathbf{X}_i)$ necessarily takes a multinomial specification. A Multinomial Logit Model (MLM) is here adopted. Even for this MLM a fully interacted polynomial specification in both arguments of the MLM is adopted. In both cases, however, the functional forms actually estimated are the best fitting specifications selected according to the AIC starting from a fully interacted second-order polynomial (quadratic specification).

Table 11 reports the GPS function estimates in the continuous (linear regression) case. It emerges that most estimated parameters are significantly different from 0, the only exceptions being those concerning variables ALT and FC. A higher generalized propensity (that is, a higher probability to receive a higher *TI*) is associated to larger size (in both economic and physical terms) farms and older farmers while, on the contrary, a lower propensity is associated to farms with higher labour intensity and specialization in less supported activity included a higher presence of livestock activities.

These results are consistent with the expectation given the well-known distribution of first pillar's coupled payments in Italy (Povellato and Velazquez, 2005) and with what observed in the binary treatment case. The reliability of these results can be assessed by testing the balancing condition over the space of the estimated GPS and distinguishing the *TI* in the abovementioned 7 intervals. The balancing tests on the common support accept the balancing condition at the 95% confidence level: for no regressor and in no treatment level we observe a mean difference between the units belonging to that treatment group and the respective average that is statistically different from 0 (see Table A4).[60]

Moreover, indirectly, the robustness of these results is also confirmed by the other GPS function, i.e., the MLE estimation of the MLM implied by the Cattaneo (2010) approach (Table A5). In this latter case, a further difference with respect to the continuous case consists in the exclusion of the categorical independent variables (ES, TF_R, RDP) as they would prevent the MLE estimation of the MLM to reach convergence. Parameter estimation of the MLM actually refer to the different *TI* levels and, as well known (Cattaneo et al., 2013.), can not be directly interpreted as marginal effects passing from one treatment level to the subsequent.[61] Nonetheless, also these results indicate that a higher *TI* is associated with larger farms (in physical and economic terms), lower labour intensity and older

[60] More details on these aspects can be found in Bia and Mattei (2008, p. 368-369; 2012) and Flores and Mitnik (2008). Bia and Mattei (2012) and Bia et al. (2013) also put forward some first attempts to perform a sensitivity analysis assessing for the validity of the weak unconfoundedness under multivalued treatments.

[61] Estimated marginal effects are available upon request.

farmers, while on the contrary lower *TI* is associated with flat-areas agriculture and more labour-intense activities included livestock productions.[62]

As a matter of fact, these estimates of the GPS function's parameters do not provide any particular novel evidence compared to expectations and to what already emerges from a simple descriptive analysis and from the binary treatment case. Nonetheless, this is just the first necessary step to achieve the multivalued ATE estimation and its reliability and robustness is needed to make ATE estimation itself reliable. Moreover, unlike the binary-treatment estimations, here the RDP dummy can be included in the analysis. This is an additional advantage of adopting a multitreatment approach as it can explicitly assess how the presence of a second, and independent, binary treatment may affect the probability to receive a given intensity of the first multivalued treatment (*TI*).[63] In particular, Table 11 reports the results of the GPS estimation also with the RDP dummy among regressors. Estimates indicate that, though this dummy does not influence much the coefficients associated to the other covariates, it is itself statistically significant and implies that the propensity to receive an higher *TI* (first pillar payment per unit of agricultural GPV) remarkably declines if a second pillar payment is also present.

[Table 11 here]

6.2.3.2. The Hirano-Imbens approach: estimation of the DRF

Table 12 reports the results of the second estimation step in the Hirano-Imbens approach for the four outcome variables under consideration here. It consists of the estimation of *uDRF*, $g(G\hat{P}S_i, T_i) = E\left[Y_i \middle| G\hat{P}S_i, T\right]$. Again, the empirical parametric specification of the estimand function started from a fully-interacted quadratic form in both arguments $G\hat{P}S_i$ and T (see (15)), where $G\hat{P}S_i$ is the GPS estimated in the first stage (Table 11). Then, the best fitting specification, according to the AIC, has been adopted. As evident in Table 12, this adopted specification includes both the quadratic terms and the interaction term of the two arguments. It is well emphasized in this empirical literature (Hirano and Imbens, 2004; Bia and Mattei, 2008) that these estimates can not be given any direct economic interpretation and are just intermediate results. However, they are needed to achieve the final estimation step providing the estimation of $aDRF(T) = E[(\hat{g}(T)]$ in (16) and, consequently, of the ATE we are mainly interested in.

Estimation results are of good quality for the first outcome variables, y^1, where parameters of most terms are statistically significant expressing a complex relationship among $G\hat{P}S_i$, T and the *uDRF*.

[62] Due to space limitations, as these results are repeated over the 7 *TI* groups, the respective balancing tests are not reported here. They are available upon request and indicate that balancing condition is always accepted at the 95% confidence level.

[63] In fact, as mentioned above, the inclusion of the RDP dummy within the GPS function is possible only the linear-regression specification (Table 11).

Starting with the second outcome variable (y^2), however, results become statistically weaker and the treatment level significantly affect the *uDRF* only interacting with the estimated GPS: the DRF comes more from the propensity to receive a given *TI* rather than by the *TI* itself. This is even more clear in y^3 and y^4, where few parameters (just one for y^3) are statistically significant and they all involve the estimated GPS, alone or in interaction with the *TI*.

It remains true, however, that results seem statistically weaker passing from the first two outcome variables, dealing with the production mix, to those related to investment choices. Generally speaking, the fact that the impact on the DRF mostly comes more from the propensity to receive a given *TI* is a further indirect evidence of how, even in the multivalued case, the selection-on-observables bias may be relevant. We can also notice, from Table 12, that results obtained with and without the inclusion of the RDP dummy do not substantially differ. Even though receiving a second pillar payment in addition may be specific of farms with some significant differences in the observed covariates and may affect the treatment assignment (consequently, the GPS), this does not relevantly affect how the outcome variable responds to the *TI* and to the estimated GPS.

Figures 3-6 display both the estimated average DRF and ATE (point estimates and confidence intervals) for the four outcome variables over the observed continuous domain of TI (therefore, we can also call the latter *average TE function*).[64] Again, we may appreciate a significant different behavior of the first two outcomes compared to y^3 and y^4. For y^1 and y^2 we observe similar results in several respect. First of all, in both cases DRF and TE estimates appear to be statistically robust as standard errors are small and, consequently, the confidence intervals thin. The response increases in the initial TI levels and reaches its maximum at about *TI* = 10% and *TI* = 30%, respectively. However, the corresponding TE, though positive, is declining. Once the maximum response is reached, it starts declining regularly and this evidently implies a negative TE. This behaviour is particular evident and relevant in the case of y^1 where the TE rapidly vanishes to 0, while, on the contrary, in the case of y^2 a positive TE is observed until TI levels that encompass the large part of the observed farms.

The interpretation of this evidence seems particularly interesting. The FPR induced the farms' response in terms of production mix/market reorientation, but this mostly regards those farms receiving a relatively limited CAP payment with respect to its GVP (i.e., a low TI). Moreover, this TE is particularly evident in terms of composition of the production mix; on the contrary, it is quite limited if this change of production mix is measured in terms of revenue composition. Evidently, the most responding farms are those that were already market-oriented before the FPR as demonstrated by the limited incidence of the CAP support on their GPV. Whenever we move towards farms strongly dependent on the CAP their response is weaker. Moreover, this response may be more apparent than

[64] These results concern the case not including the RDP among covariates. Very similar results are obtained for the specification with RDP and are available upon request.

substantial. It implies that new productions are activated while others are quitted and this affects y^2. At the same time, this change might not provide any significant change in revenue composition, that is, in y^1. In practice, many farms, especially those strongly dependent on first pillar support, remain locked-in their original rent-seeking production choices and only timidly attempt some minor changes that actually result in marginal adjustments in revenue composition (*lock-in effect*).

As could expected, the results emerging for the other two outcome variables, y^3 and y^4, are much less robust and interpretable. Expectations come from the fact that investment decisions may depend on many strategic long-term aspects, not considered here, beside the decoupling of CAP support. Moreover, as discussed in previous sections, an investment response to the policy change may have a timing (anticipation and lagged effects) that makes a proper identification of the ATE in the present circumstance much more difficult. It is still worth noticing that the response, though with an increasing statistical variability, linearly increases over the observed TI domain. This occurs in both cases thus suggesting that farm size (that is, what really makes the difference between y^3 and y^4) does not qualitatively influence the response.

The slight but regular and statistically significant increase of the TE across the *TI* can be interpreted in two directions. On the one hand, we can conclude that the TE is constant over the *TI* domain, so the ATE corresponds to the TE observed in any different treatment level. This seems particularly reasonable in the case of y^4, as it implies that an increase of *TI* has a constant positive impact on investment rates. On the other hand, we can conclude that, though slightly, the TE is increasing with the TI and this seems consistent with the decoupling of support: the higher the *TI*, the higher the amount, relative to the farm's GPV, of free financial resources provided by the FPR to make investments. The fact that these resources are free, i.e. not associated to specific production activities, may indirectly suggest that these investments are dedicated to other/new activities. This may represent a sort of *pure financial effect* of the decoupling of support at higher *TI* values that can be also reinforced for highly risk-adverse farmers. As a matter of fact, decoupled support also represents a less risky contribution to farm income (Moro and Sckokai, 2011). An higher *TI* is thus associated to an higher insurance in taking brand new investments thus becoming a sort of incentive to invest.

However, this is not granted given the definition of the two outcome variables under question (y^3 and y^4). Even if we acknowledge that a positive and maybe increasing TE occurs, this does exclude that investments are made on the same activities carried out by the farm even before the FPR.

[Table 12 here]

[Figures 3-6 here]

6.2.3.3. EIF and IPW estimation

As discussed in section 3.3, the robustness of the DRF and ATE estimation obtained with the Hirano-Imbens approach can be hardly assessed. Once the balancing condition is confirmed in the GPS estimation stage, the remaining steps strongly depend on arbitrary (parametric) specification assumptions. To indirectly evaluate to what extent these assumptions may influence results, the Cattaneo (2010) approach can be helpful as it is based on a substantially different sequence of estimation steps.

The second step of this approach consists in the estimation of $e_j(\mu_j, \mathbf{X}_i)$, that is, the mean of the potential outcome conditional on covariates $\mathbf{X}$. Again, a flexible form is initially specified and the best fitting specification, selected with AIC, is then adopted. Here, the selected specification is the fully interacted quadratic form that implies the estimation of an high number of parameters. Therefore, due to space limitations, these estimates are not commented here and are reported in Table A6. After all, again, this is an intermediary estimation step, without particular economic content, to achieve the final stage, that is, the estimation of the average potential outcome for any given treatment level under the two alternative estimation procedures ($\mu_{EIF,j}$ and $\mu_{IPW,j}$).

Table 13 reports the $\mu_{EIF,j}$ and $\mu_{IPW,j}$ estimates for all outcome variables and respective treatment levels, while Figures 7 and 8 display the respective variations across the treatment levels, that is, the estimated ATE according to (25). In general terms, results confirm what emerged within the Hirano-Imbens approach. For both y^1 and y^2 the potential outcome increases with the treatment for the first *TI* levels, then we have the inversion in this relationship with potential outcome declining with the *TI*. This inversion comes quite soon in the case of y^1, i.e. when *TI* exceeds 10%, while it comes later for y^2. This implies that the estimated TE is positive in the first levels of treatment and then declines to zero and even becomes negative. Compared to the DRF estimation of the Hirano-Imbens approach, the most evident difference consists in the fact that confidence intervals are here larger, thus indicating a less efficient estimate of the ATE. Nonetheless, the tests of "zero TE" reported in Table 13 clearly indicate that such hypothesis can be excluded for both y^1 and y^2.

In the case of y^3 and y^4 it seems more difficult to conclude that results tend to be qualitatively concordant with what obtained with the Hirano-Imbens approach. In particular, confidence intervals around the estimated potential outcome are large such that, for some treatment levels, the it is not statistically different from zero. This is particularly true for y^3 and in the case of the IPW estimation. Therefore, also the consequent estimated ATE are often not statistically significant and do not necessarily show a regular pattern across the treatment levels. In the case of y^3, though we can reject the hypothesis of no TE, the only robustly significant ATE is that observed passing from the first to the second treatment level, that is, around 5% of *TI*. However, this effect is negative. For all other TI the ATE is statistically zero. Better evidence is obtained for y^4, as expected due to the lower statistical noise caused by heterogeneous size. The ATE is positive passing from the first to the second treatment levels than it goes to 0, but it comes back into positive territory and with an apparent increasing pattern after the forth treatment level. In this respect, these results confirm what obtained in the case of Hirano-Imbens approach. The investment response to the treatment is much less clear than the production choice response. Nonetheless, when the investment rate, rather than the investment absolute value, is considered the response seems to be positive and, to a certain extent, slightly increasing with the increase of the *TI*. Higher *TI*, as mentioned, can be associated to an higher amount of free financial resources to make brand new investments.

A further general evidence that is worth noticing concerns the comparison between EIF and IPW estimates. Though results are qualitatively similar, IPW estimates are clearly less efficient. Therefore, despite the asymptotical equivalence of these two estimators, in the finite sample and in this specific application, the EIF estimation appears to be generally superior.

[Table 13 here]

[Figures 7-8 here]

7. CONCLUDING REMARKS

This paper primarily aims at showing how, even for policy treatments that apparently poorly fit into this kind of approaches, the TE toolkit developed in the last decades still allows performing and comparing different estimation methods, assessing the robustness of results, carrying out sensitivity analysis. Without necessarily opting for one method rather than another, the contemporaneous use of all these methods and the comparison of their results may still provide reliable evidence on the impact of policies or programmes.

The empirical approach here proposed aims at assessing the effect of the FPR on farms' production choices. Results show that, as expected, the farms' response in terms of market reorientation tends to be limited to short-run choices and has been declining with the treatment intensity. More surprisingly, it emerges that this response is significant and positive just for the lower levels of treatment intensity, that is, for those farms that are less dependent on first pillar support. In more dependent farms the response rapidly declines to 0 and even becomes negative. Moreover, this response is more evident in terms of introduction of new productions or, more generally, of change in the production mix. It significantly reduces when the response is measured in terms of change in revenue composition thus indicating that, even when present, this response remains fairly conservative.

On the one hand, results show how, with a long and large enough balanced panel and an appropriate definition of the outcome variable, the impact of the reform on farms' market reorientation can be properly estimated by taking advantage of the fact that it can be considered either a binary or a multivalued treatment. Recent estimation approaches suitable for the latter case have been also applied. On the other hand, however, while results provide quite robust evidence about the effect of the reform, some steps forward could be proposed with respect to the present approach.

The adopted methodological toolkit seems rich enough not only to check for robustness of results by comparing different estimation approaches, but also to push the investigation further by refining the estimation approaches, the definition of the treatment and outcome variables (as well as of the relevant covariates), and the construction of the balanced panel. For instance, it could be argued that some second pillar measures substantially interfere with production choices, thus with the FPR TE itself. As a consequence, a more sophisticated articulation of treatment groups could be attempted. Actually, the analysis of multiple continuous treatments is at the forefront of the current TE econometrics literature (Frölich, 2004; Imbens and Wooldridge, 2009) and is well beyond the scope of the present paper. Nonetheless, some methodological solutions accompanied by appropriate matching and estimation techniques could be proposed and attempted in future research.

ACKNOWLEDGMENT

The author is grateful to the National Institute of Agricultural Economics (INEA) for having provided him with the FADN Italian database.

Table 1. Distribution of sample farms across the treatment groups and descriptive statistics of the respective amount of CAP support (First Pillar)

Treatment group:	*Not treated*	*Treated*		*Whole Sample*
		Only First Pillar	*Both Pillars*	
Number of farms (%)	1112 (17%)	3664 (56%)	1766 (27%)	6542 (100%)
1st Pillar support (2003) (€)				
Avg. amount of support	0	13319	12517	10838
Standard Deviation	0	75765	53609	56903
Coefficient of Variation (CV)	0	5.69	4.28	5.25
Minimum	0	0	0	0
Maximum	0	2205000	2004153	2205000
1st Pillar support (2004) (€)				
Avg. amount of support	0	14698	14641	12184
Standard Deviation	0	71779	41462	51391
Coefficient of Variation (CV)	0	4.88	2.83	4.22
Minimum	0	0	0	0
Maximum	0	2062500	616554	2062500
1st Pillar support (2005) (€)				
Avg. amount of support	0	14823	15894	12592
Standard Deviation	0	67897	47025	50719
Coefficient of Variation (CV)	0	4.58	2.96	4.03
Minimum	0	0	0	0
Maximum	0	1539952	686356	1539952
1st Pillar support (2006) (€)				
Avg. amount of support	0	11220	12425	9638
Standard Deviation	0	56237	53801	46019
Coefficient of Variation (CV)	0	5.01	4.33	4.77
Minimum	0	0	0	0
Maximum	0	1296061	1078351	1296061
1st Pillar support (2007) (€)				
Avg. amount of support	0	14978	15671	12619
Standard Deviation	0	69993	43941	51060
Coefficient of Variation (CV)	0	4.67	2.80	4.05
Minimum	0	0	0	0
Maximum	0	1635650	584075	1635650

Table 2. Sample averages of the outcome variables over treatment groups (standard deviation in parenthesis)

Treatment group:	*Not treated*	*Treated*		*Whole Sample*
		Only First Pillar	*Both Pillars*	
Outcome variables:				
y^1 (distance index)				
2007 vs. 2003	0.169 (0.435)	0.505 (0.653)	0.780 (0.920)	0.522 (0.688)
2004 vs. 2003	0.084 (0.322)	0.469 (0.671)	0.756 (0.998)	0.481 (0.700)
2005 vs. 2004	0.096 (0.311)	0.157 (0.342)	0.251 (0.417)	0.172 (0.357)
2006 vs. 2004	0.092 (0.316)	0.177 (0.379)	0.186 (0.385)	0.165 (0.370)
2006 vs. 2005	0.096 (0.311)	0.159 (0.339)	0.273 (0.464)	0.179 (0.368)
2007 vs. 2006	0.054 (0.249)	0.123 (0.305)	0.112 (0.300)	0.108 (0.294)
y^2 (distance index)				
2007 vs. 2003	0.006 (0.014)	0.016 (0.019)	0.019 (0.021)	0.015 (0.019)
2004 vs. 2003	0.003 (0.009)	0.011 (0.016)	0.012 (0.020)	0.010 (0.016)
2005 vs. 2004	0.004 (0.011)	0.014 (0.018)	0.020 (0.023)	0.014 (0.019)
2006 vs. 2004	0.005 (0.012)	0.018 (0.019)	0.019 (0.021)	0.016 (0.020)
2006 vs. 2005	0.004 (0.010)	0.013 (0.018)	0.019 (0.022)	0.013 (0.019)
2007 vs. 2006	0.007 (0.012)	0.020 (0.016)	0.021 (0.019)	0.018 (0.016)
y^3 (in €)				
2007 vs. 2003	-20477 (179122)	-7998 (132044)	-615 (147010)	-8126 (171088)
2004 vs. 2003	-8862 (163765)	-2524 (88680)	11407 (136897)	160 (141463)
2005 vs. 2004	-4787 (35010)	697 (115710)	-1643 (117391)	-867 (117835)
2006 vs. 2004	-7567 (53784)	-3097 (94343)	-4597 (115999)	-4262 (120295)
2006 vs. 2005	-2779 (37919)	-3974 (71351)	-2586 (116160)	-3396 (88566)
2007 vs. 2006	-4048 (35240)	-2378 (106188)	-7423 (119181)	-4024 (103035)
y^4 (in €)				
2007 vs. 2003	-0.401 (1.533)	-0.267 (1.633)	-0.283 (2.363)	-0.294 (2.083)
2004 vs. 2003	-0.101 (1.193)	0.003 (1.757)	-0.083 (1.790)	-0.038 (1.883)
2005 vs. 2004	0.044 (3.882)	-0.044 (2.081)	-0.033 (2.166)	-0.026 (2.410)
2006 vs. 2004	-0.172 (1.213)	-0.079 (2.850)	0.057 (3.262)	-0.058 (3.223)
2006 vs. 2005	-0.216 (3.914)	-0.035 (1.960)	0.097 (2.863)	-0.030 (3.076)
2007 vs. 2006	-0.128 (0.958)	-0.191 (2.491)	-0.260 (2.982)	-0.199 (2.903)

Table 3. Sample averages of the pre-treatment variables over treatment groups (standard deviation in parenthesis)

Treatment group:	*Not treated*	*Treated*		*Whole Sample*
		Only First Pillar	*Both Pillars*	
Pre-treatment variables (X):				
AGE (of the holder) (years)	51.85 (13.84)	53.95 (15.05)	50.05 (14.07)	52.54 (14.71)
Altitude (ALT) (m)	154.28 (192.11)	234.95 (215.60)	405.33 (359.34)	267.24 (273.79)
Annual Working Units (AWU)	2.79 (6.22)	1.90 (2.71)	2.38 (3.30)	2.18 (3.60)
Economic Size (ES) (classes)	6.41 (2.19)	5.85 (2.42)	6.68 (1.90)	6.17 (2.29)
Fixed Costs (on Net Value Added) (FC)	2.79 (36.41)	2.12 (17.95)	1.18 (9.27)	1.98 (20.00)
Horse Power (HP)	93.23 (129.58)	173.94 (206.31)	215.68 (260.50)	171.49 (217.15)
Livestock Units (LU)	5.73 (50.53)	41.20 (214.38)	58.09 (277.26)	39.73 (220.44)
Utilized Agricultural Area (UAA) (ha)	7.50 (24.34)	30.52 (63.46)	53.16 (85.15)	32.72 (67.93)
Type of Farm (TF) (4-digits)[a]	3211 (fruits)	1310 (arable crops)	4110 (dairy)	1310 (arable crops)
Type of Farm (reclassified) (TF_R)	5.07 (1.67)	3.40 (1.59)	3.31 (1.40)	3.66 (1.65)
Pillar II support dummy (RDP)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.28 (0.15)

[a] In this case the Table reports the higher frequency class

Table 4. Propensity Score (PS): binomial probit estimation[a]

Pre-treatment variables (X):	**Coefficient estimate** (standard error in parenthesis):
Constant term	1.805* (0.1283)
AGE (of the holder) (years)	0.0039* (0.0015)
Altitude (ALT) (m)	0.0009* (0.0001)
Annual Working Units (AWU)	-0.0503* (0.0076)
Economic Size (ES) (classes)	-0.0407* (0.0110)
Fixed Costs (on Net Value Added) (FC)	-0.0006 (0.0006)
Horse Power (HP)	0.0020* (0.0002)
Livestock Units (LU)	0.0022* (0.0003)
Utilized Agricultural Area (UAA) (ha)	0.0010* (0.0001)
Type of Farm (reclassified) (TF_R)	-0.3011* (0.0148)

[a] Balancing condition satisfied on the common support (6528 observations) at the 0.01 significance level
*Statistically significant at 0.05 level

Table 5. *ATT* estimates for the four outcome variables (standard errors in parenthesis)[a,b]

Outcome variables:	*Stratification Matching*	*Nearest Neighbour Matching*	*Radius Matching*[c]	*Kernel Matching*
y^1	0.371*	0.429*	0.426*	0.380*
	(0.052)	(0.048)	(0.021)	(0.037)
y^2	0.010*	0.010*	0.010*	0.009*
	(0.001)	(0.001)	(0.001)	(0.001)
y^3	10525	6208[b]	8468	13156
	(10267)	(16141)	(12182)	(13292)
y^4	0.061	0.064	0.148	0.092
	(0.097)	(0.108)	(0.073)	(0.085)

[a] All estimates are performed on the common support (6528 observations)
[b] Bootstrap standard errors obtained with 1000 replications
[c] Radius = 0.05
*Statistically significant at 0.05 level

Table 6. Rosebaum bounding approach to the analysis of deviations from the CIA: robustness of the PSM estimated ATT to hidden bias [a]

		Overestimation of TEs		*Underestimation of TEs*	
Outcome variables:	Γ	*Wilcoxon signed ranked test - significance level*	*Upper bound Hodges-Lehmann point estimate*	*Wilcoxon signed ranked test - significance level*	*Lower bound Hodges-Lehmann point estimate*
y^1	1	0	-	0	-
	2	< 0.001	0.092	< 0.001	0.681
	3	< 0.001	0.019	< 0.001	0.851
	4	0.4534	-0.000	< 0.001	0.948
y^2	1	0	-	0	-
	2	< 0.001	0.007	< 0.001	0.013
	3	< 0.001	0.000	< 0.001	0.020
	4	0.821	-0.000	< 0.001	0.026
y^3	1	0.981	-	0.981	-
	2	1	-14121.5	< 0.001	10053
y^4	1	< 0.001	-	< 0.001	-
	2	1	-0.269	< 0.001	0.461

[a] Tests and calculations refer to single Nearest Neighbour Matching PSM estimation

Table 7. Deviations from the CIA: sensitivity analysis as effect of different kinds of confounder on PSM ATT estimates (standard errors in parenthesis)[a,b]

	Calibrated confounder (dummy ES>6)[c]			*Simulated confounder*					
				d = 0.3, s = 0.6			*d = 0.6, s = 0.75*		
Outcome variables:	*ATT*	*Outcome Effect*	*Selection Effect*	*ATT*	*Outcome Effect*	*Selection Effect*	*ATT*	*Outcome Effect*	*Selection Effect*
y^1	0.385* (0.064)	1.181	0.823	0.305* (0.090)	3.645	17.945	-0.017 (0.142)	23.100	46.466
y^2	0.010* (0.002)	1.098	0.712	0.008* (0.003)	3.680	17.239	0.001 (0.005)	23.068	41.829
y^3	-522 (20268)	0.979	0.868	7656 (37734)	3.625	16.867	-12708 (25996)	21.953	40.827
y^4	0.070 (0.172)	1.220	0.822	0.107 (0.202)	3.526	16.285	-0.320 (0.206)	21.528	41.059

[a] Sensitivity analysis refers to Nearest Neighbour Matching PSM estimation and is performed on the common support
[b] Bootstrap standard errors obtained with 1000 replications
[c] This confounder implies $d = 0.05$ and $s = -0.05$
*Statistically significant at 0.05 level

Table 8. Unconditional and conditional DID estimates (standard errors in parenthesis)[a]

Outcome variables:	*DID*		*CDID*		*PSM-DID* [b]	
	(1)	*(2)*	*(1)*	*(2)*	*(1)*	*(2)*
y^1	-0.376*	0.022*	-0.379*	0.023*	-0.313*	0.026*
	(0.017)	(0.010)	(0.018)	(0.011)	(0.069)	(0.011)
y^2	0.004*	0.001	0.004*	0.001	0.004*	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
y^3	-6033	3829	-5807	3853	-18356	-1154
	(5471)	(2881)	(5996)	(3955)	(9800)	(5744)
y^4	-0.154	0.219*	-0.151	0.219*	-0.346	0.085
	(0.108)	(0.076)	(0.130)	(0.076)	(0.187)	(0.161)

[a] Bootstrap standard errors obtained with 1000 replications
[b] Kernel-based PSM
(1) Baseline= 2004-03, Follow-up = 2005-04
(2) Baseline= 2007-06, Follow-up = 2007-04
*Statistically significant at 0.05 level

Table 9. Placebo test: DID estimates across two baselines 2004-03 and 2007-06 (standard errors in parenthesis)[a]

Outcome variables:	*DID*	*CDID*	*PSM-DID* [b]
y^1	-0.401*	-0.398*	-0.313*
	(0.026)	(0.025)	(0.031)
y^2	0.005*	0.004*	0.003*
	(0.001)	(0.001)	(0.001)
y^3	-11391	-10477	-18004
	(6173)	(6168)	(15318)
y^4	-0.151	-0.148	-0.145
	(0.122)	(0.121)	(0.131)

[a] Bootstrap standard errors obtained with 1000 replications
[b] Kernel-based PSM
*Statistically significant at 0.05 level

Table 10. Quantile treatment effects (QTT): PSM-DID estimates (standard errors in parenthesis)[a,b]

	(1)			*(2)*		
Quantiles:	*25%*	*50%*	*75%*	*25%*	*50%*	*75%*
Outcome variables:						
y^1	-0.001*	-0.072*	-0.903*	0.000	0.008*	0.048*
	(0.000)	(0.006)	(0.027)	(0.000)	(0.001)	(0.012)
y^2	0.000	0.014*	0.015*	-0.012	0.000	0.022*
	(0.000)	(0.002)	(0.002)	(0.008)	(0.001)	(0.001)
y^3	-159	80.9	160	4728	1936	263
	(1232)	(283)	(130)	(1050)	(534)	(241)
y^4	-0.030	0.003	0.058	0.040	0.073*	0.107*
	(0.051)	(0.013)	(0.046)	(0.048)	(0.020)	(0.023)

[a] Bootstrap standard errors obtained with 1000 replications
[b] Kernel-based PSM
(1) Baseline= 2004-03, Follow-up = 2005-04
(2) Baseline= 2007-06, Follow-up = 2007-04
*Statistically significant at 0.05 level

Table 11. GPS estimation: linear regression of the continuous treatment (*TI*) on the covariates (standard errors in parenthesis)[a,b]

Specification:	*Without RDP dummy*	*With RDP dummy*
Pre-treatment variables (X):		
Constant term	4.365*	4.4038*
	(0.1620)	(0.1618)
AGE (of the holder) (years)	0.0132*	0.0124*
	(0.0019)	(0.0019)
Altitude (ALT) (m)	-0.0000	0.0001
	(0.0001)	(0.0001)
Annual Working Units (AWU)	-0.1882*	-0.1901*
	(0.0123)	(0.0123)
Economic Size (ES) (classes)	0.0405*	0.0516*
	(0.0130)	(0.0131)
Fixed Costs (on Net Value Added) (FC)	-0.0018	-0.0019
	(0.0013)	(0.0013)
Horse Power (HP)	0.0004*	0.0005*
	(0.0002)	(0.0002)
Livestock Units (LU)	-0.0003*	-0.0003*
	(0.0001)	(0.0001)
Utilized Agricultural Area (UAA) (ha)	0.0067*	0.0069*
	(0.0005)	(0.0005)
Type of Farm (reclassified) (TF_R)	-0.5639*	-0.5607*
	(0.0188)	(0.0188)
RDP	-	-0.4360*
		(0.0628)

[a] The BoxCox transformation of the treatment variable is used
[b] The assumption of Normality is statistically satisfied at 0.05 level
*Statistically significant at 0.05 level

Table 12. DRF coefficient estimates (GPS = estimated Generalised Propensity Score; TI = Treatment Intensity)[a]

Specification:	*Without RDP dummy*				*With RDP dummy*			
	y^1	y^2	y^3	y^4	y^1	y^2	y^3	y^4
TI	0.0182*	0.0001	-729.8	-0.0103	0.0200*	0.0001	-562.2	-0.0096
	(0.0037)	(0.0001)	(903.9)	(0.0118)	(0.0037)	(0.0001)	(896.0)	(0.0117)
TI^2	-0.0003*	0.0000	9.437	0.0002	-0.0004*	-0.0000	6.963	0.0002
	(0.0001)	(0.0000)	(16.616)	(0.0002)	(0.0001)	(0.0000)	(16.484)	(0.0002)
GPS	4.921*	-0.0575*	107956	-4.092*	5.039*	-0.0551*	170548	-4.172*
	(0.9162)	(0.0261)	(221262)	(1.879)	(0.9158)	(0.0261)	(221230)	(2.079)
GPS^2	-9.239*	0.2389*	-703988	17.82*	-9.8090*	0.2328*	-974590	16.67*
	(3.603)	(0.1024)	(870085)	(8.320)	(3.603)	(0.1026)	(870418)	(8.326)
GPS*TI	-0.1378*	0.0023*	7411*	0.0467	-0.1459*	0.0019*	6570*	0.0500
	(0.0168)	(0.0005)	(3065)	(0.0529)	(0.0166)	(0.0005)	(3018)	(0.0523)
Constant term	0.2094*	0.0165*	-11528	-0.1136	0.2041*	0.0163*	-14251	-0.0829
	(0.0549)	(0.0016)	(13262)	(0.1725)	(0.0549)	(0.0016)	(13259)	(0.1725)

[a] The BoxCox transformation of the treatment variable is used
*Statistically significant at 0.05 level

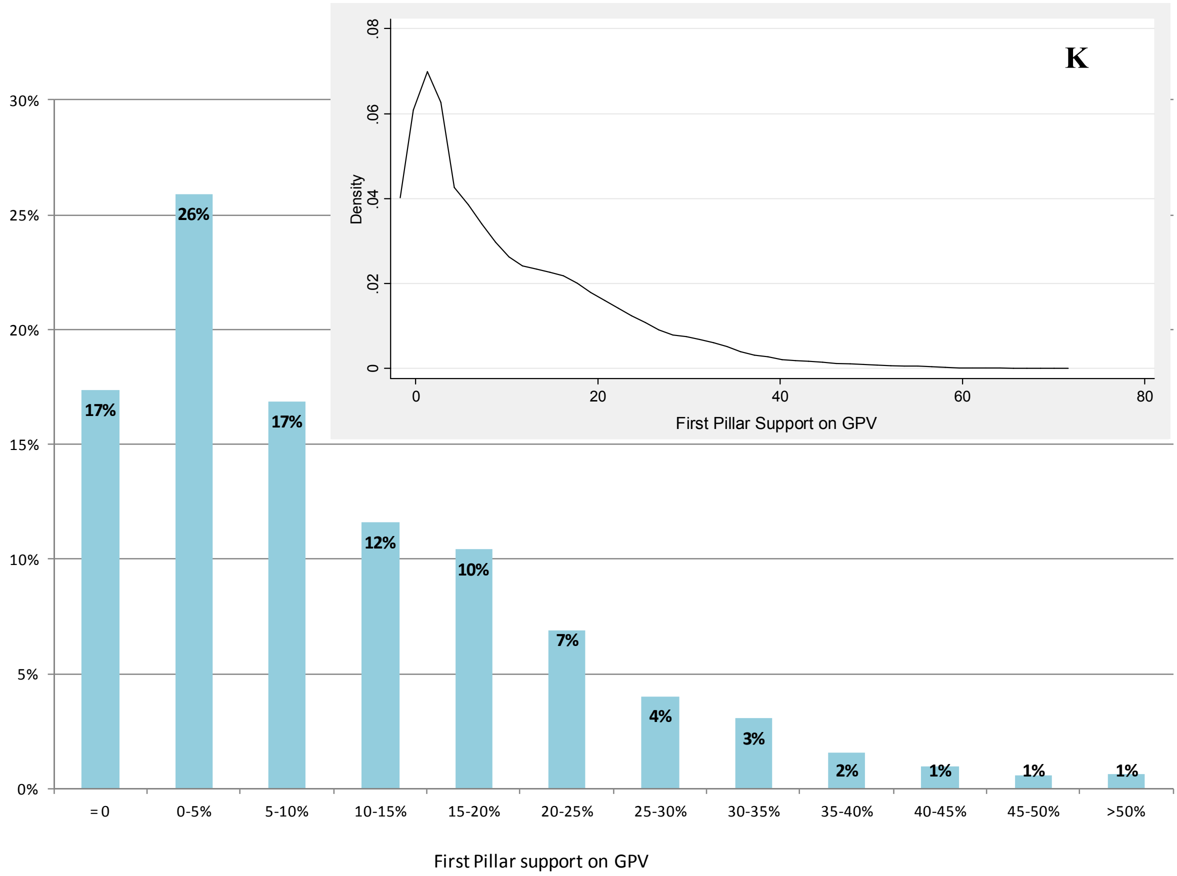
Table 13. Average potential outcome for each treatment level: EIF and IPW estimates (standard errors in parenthesis)

Outcome variable:	y^1	y^2	y^3	y^4
Treatment level:				
	EIF estimates			
1	0.6342*	0.0141*	-2799	-0.3877*
	(0.0163)	(0.0004)	(3502)	(0.0747)
2	0.8540*	0.0175*	-21658*	-0.3551*
	(0.0203)	(0.0006)	(6110)	(0.0782)
3	0.7885*	0.0195*	-8364	-0.3540*
	(0.0240)	(0.0008)	(2945)	(0.0617)
4	0.7008*	0.0195*	-11230*	-0.3681*
	(0.0294)	(0.0009)	(9281)	(0.0495)
5	0.6609*	0.0190*	6892	-0.3234*
	(0.0287)	(0.0011)	(11470)	(0.0729)
6	0.6883*	0.0186*	-20934*	-0.2265*
	(0.0289)	(0.0013)	(9764)	(0.1130)
7	0.6404*	0.0146*	-29781*	-0.1909*
	(0.0263)	(0.0014)	(6480)	(0.0684)
Test of zero treatment effect[a]	126.9*	47.18*	12.23*	12.34*
	IPW estimates			
1	0.6155*	0.0142*	-5581	-0.4045*
	(0.0187)	(0.0010)	(6925)	(0.0910)
2	0.8289*	0.0177*	-27916*	-0.4010*
	(0.0219)	(0.0007)	(3946)	(0.0966)
3	0.6458*	0.0194*	-13203	-0.1351
	(0.0814)	(0.0076)	(20588)	(0.1036)
4	0.5324*	0.0194*	-11968*	-0.3166*
	(0.0550)	(0.0037)	(5767)	(0. 0901)
5	0.4921*	0.0196*	5393	-0.2712
	(0.1728)	(0.0043)	(12822)	(0.1576)
6	0.4849	0.0178	-7816	-0.3831
	(0.2797)	(0.0172)	(50054)	(0.6418)
7	0.4094*	0.0171	-8617	-0.1441
	(0.4674)	(0.0126)	(76096)	(0.9844)
Test of zero treatment effect[a]	90.1*	41.1*	16.5*	7.5

[a] The test distributes as a χ^2 with 6 d.o.f. under the H_0 that the potential outcome is equal for any treatment level

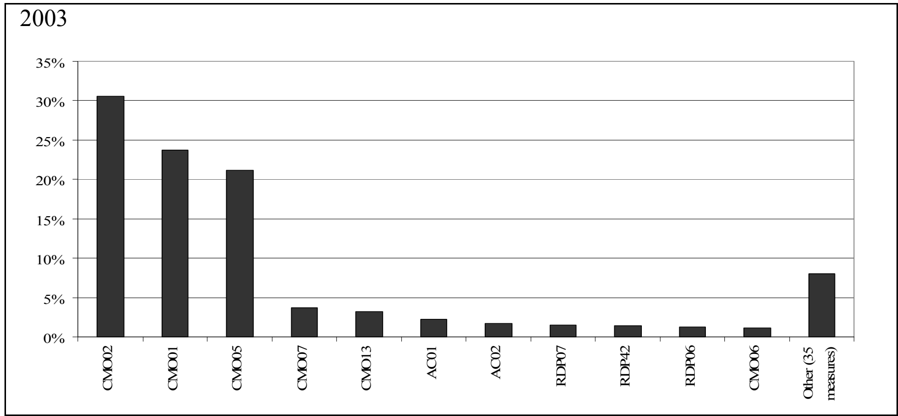
*Statistically significant at 0.05 level

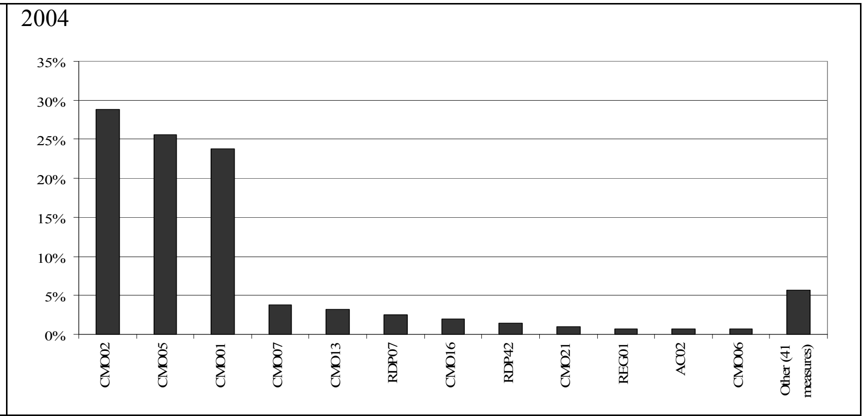
Figure 1. Distribution of the continuous treatment (*TI*), First Pillar support on farm's GPV (in %): Kernel density (**K**)[a] and frequency histogram (avg. over 2003-2007 period)

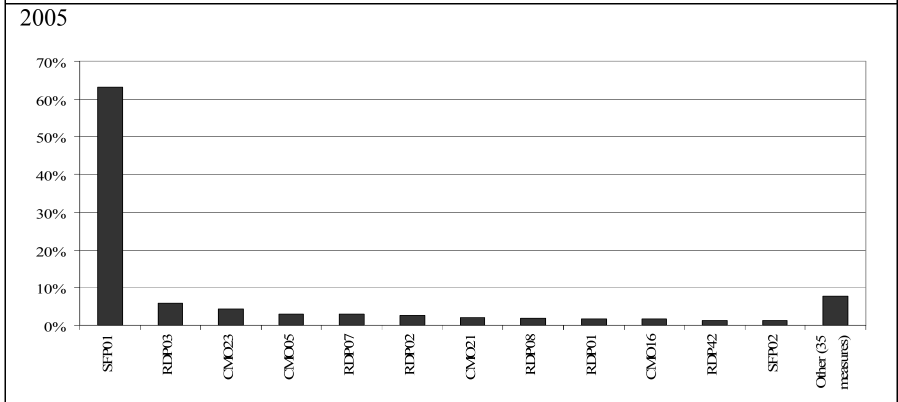


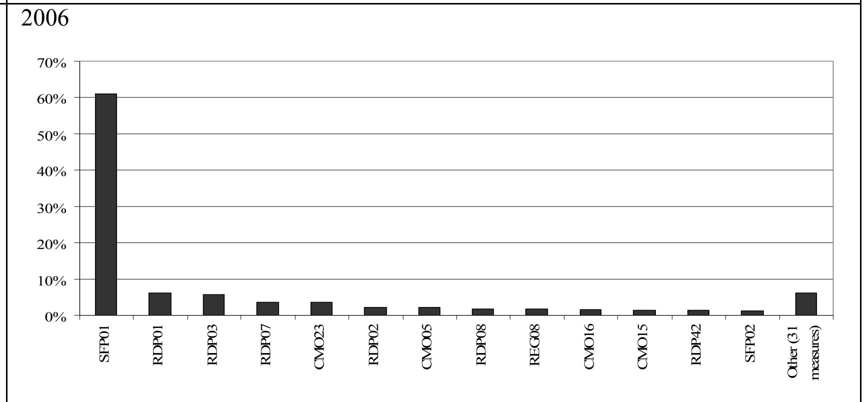
[a] Epanechnikov Kernel; Bandwidth = 1.7315

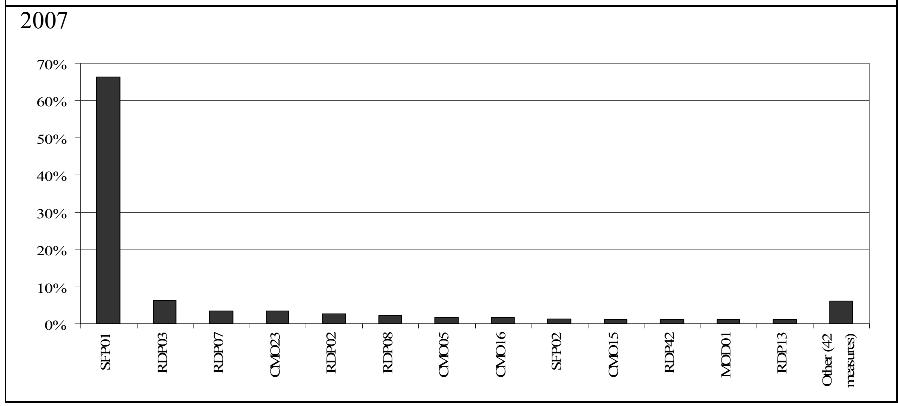
Figure 2. Distribution (%) of the total support across CAP measures (significant measures: >1% on total support) within the sample (Table A2)











Total number of payments:

2003 = 16892 (2,6 per farm on avg.). Total support = 82,8 mln € (12700 € per farm on avg.)

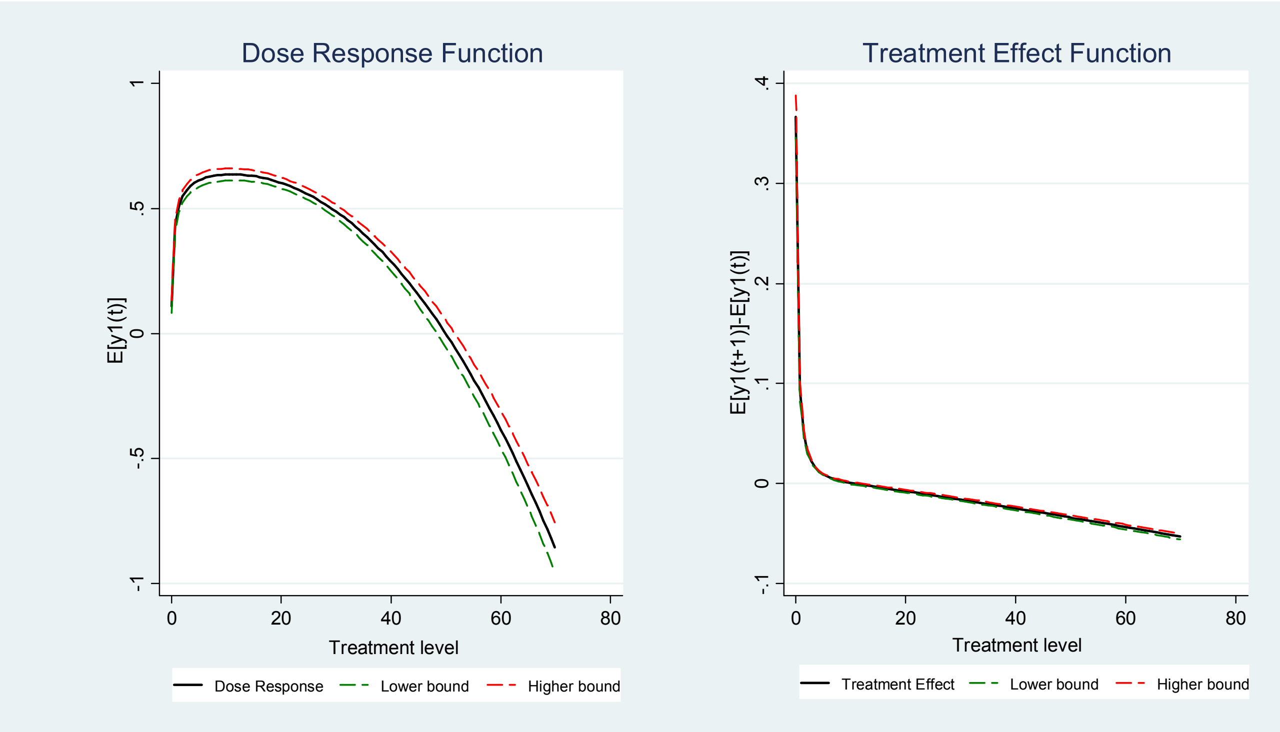
2004 = 18864 (2,9 per farm on avg.). Total support = 92,4 mln € (14100 € per farm on avg.)

2005 = 20203 (3,1 per farm on avg.). Total support = 97,2 mln € (14900 € per farm on avg.)

2006 = 10092 (1,5 per farm on avg.). Total support = 62,4 mln € (9600 € per farm on avg.)

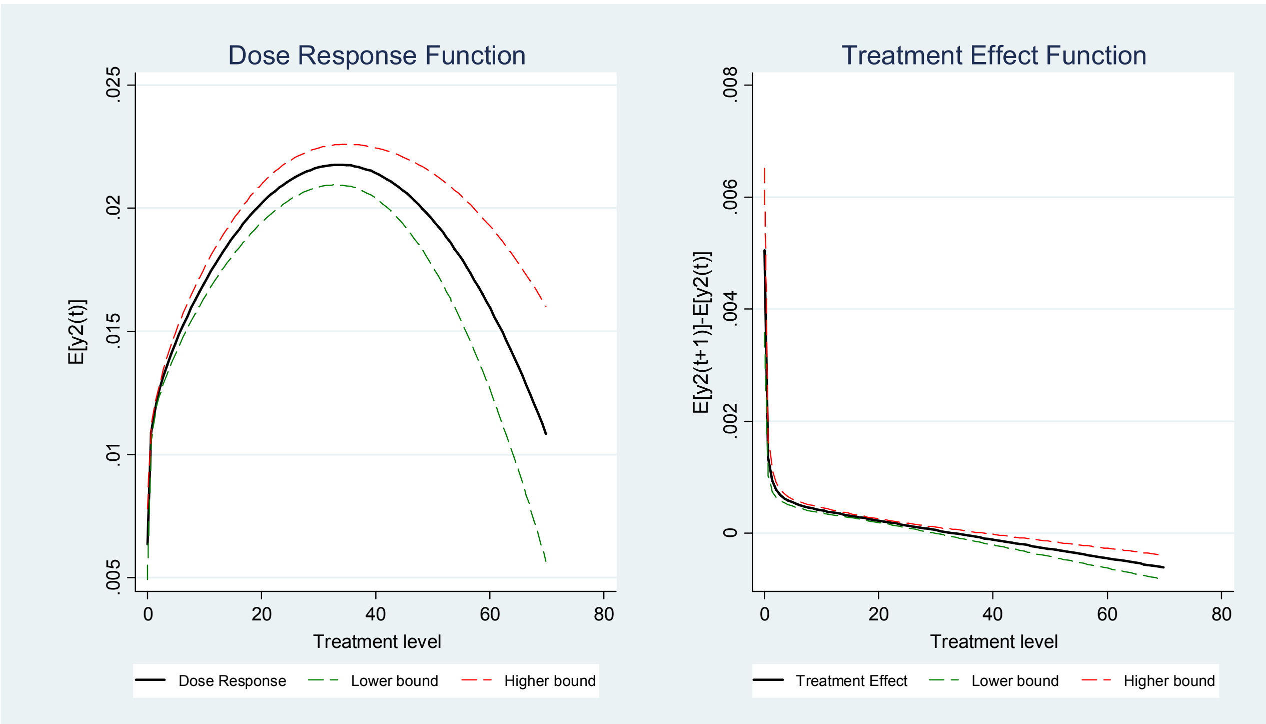
2007 = 17430 (2,7 per farm on avg.). Total support = 96,3 mln € (14700 € per farm on avg.)

Figure 3. Estimated average DRF and TE for outcome variable $y^{1\ a,b}$



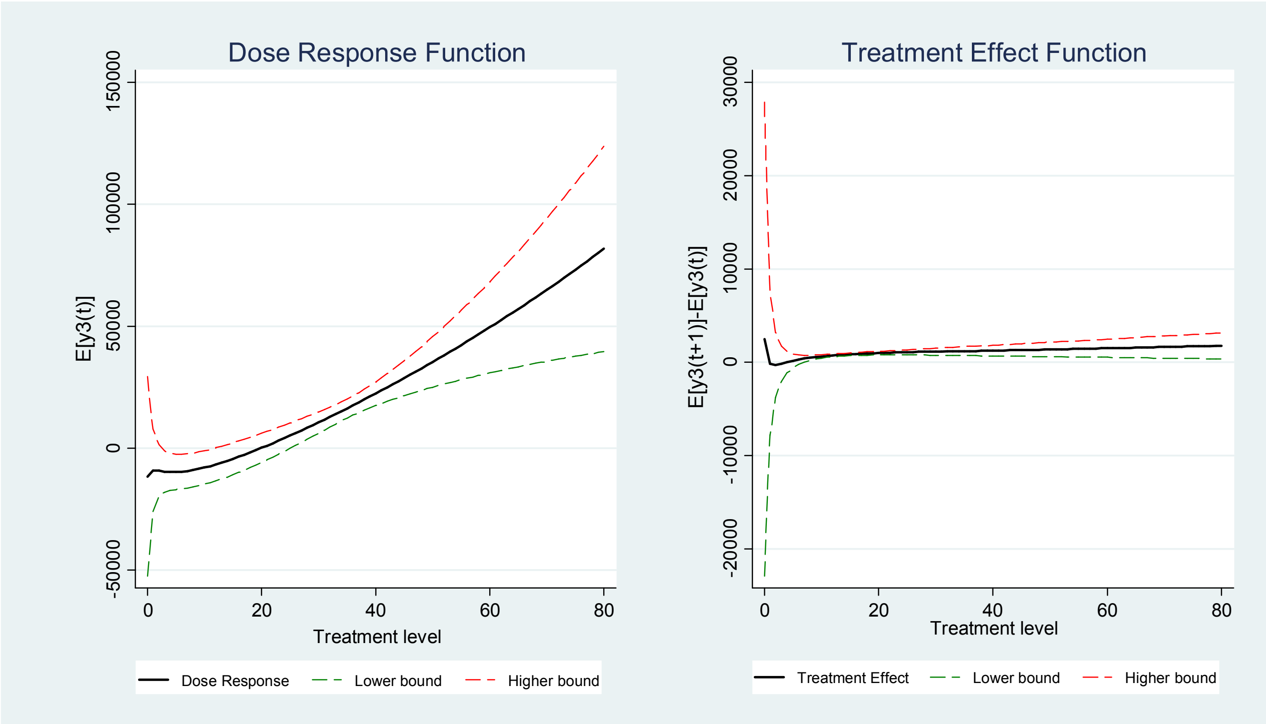
[a] Bootstrap standard errors (100 replications)
[b] Confidence bounds at 0.95 % level

Figure 4. Estimated DRF and TE for outcome variable $y^{2\ a,b}$



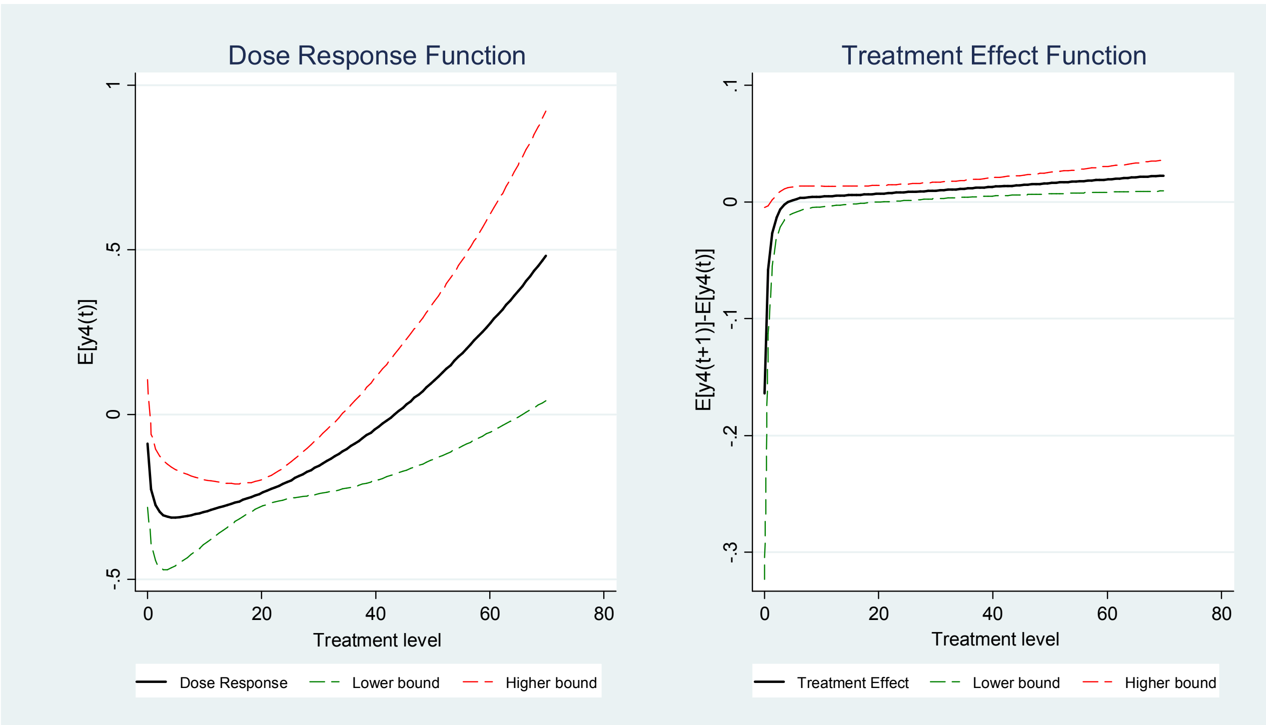
[a] Bootstrap standard errors (100 replications)
[b] Confidence bounds at 0.95 % level

Figure 5. Estimated DRF and TE for outcome variable $y^{3\ a,b}$



[a] Bootstrap standard errors (100 replications)
[b] Confidence bounds at 0.95 % level

Figure 6. Estimated DRF and TE for outcome variable $y^{4\ a,b}$



[a] Bootstrap standard errors (100 replications)
[b] Confidence bounds at 0.95 % level

Figure 7. Potential outcome variation across the treatment levels: EIF estimation of the ATE

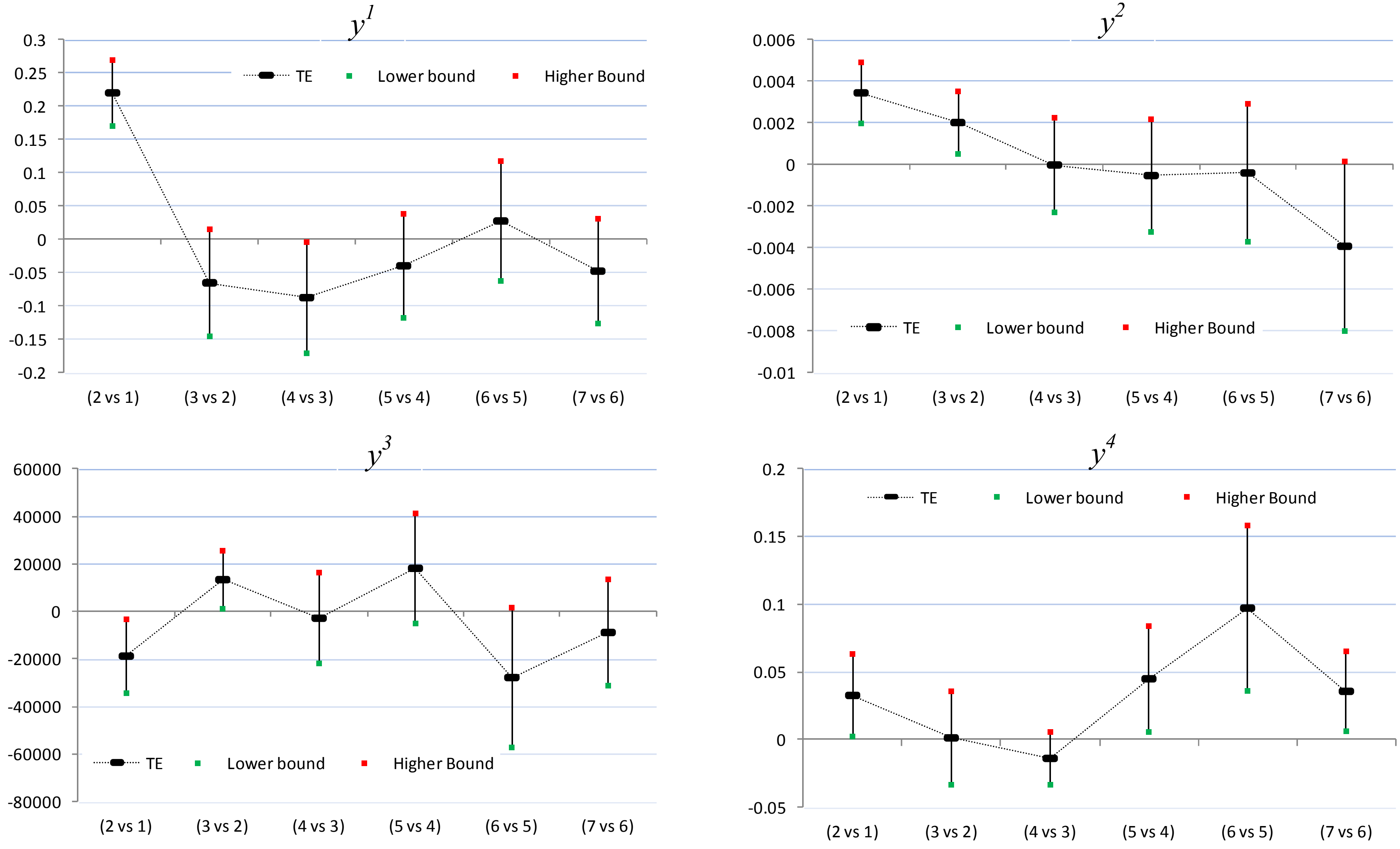
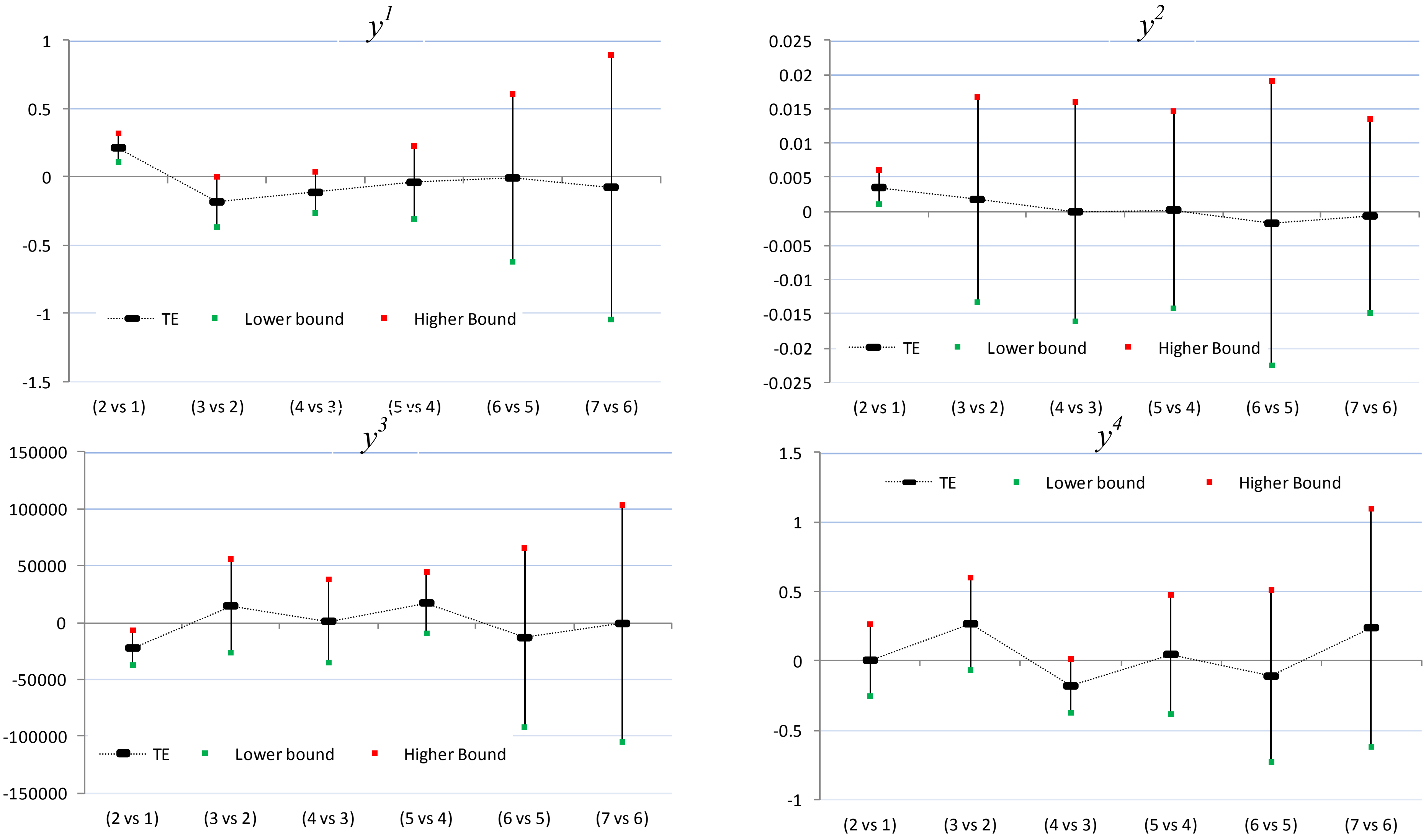


Figure 8. Potential outcome variation across the treatment levels: IPW estimation of the ATE



REFERENCES

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *Review of Economic Studies*, 72(1), 1–19.

Abadie, A., Drukker, D., Herr, J.L. and Imbens, G.W. (2004). Implementing Matching Estimators for Average Treatment Effects in Stata. *The Stata Journal*, 4(3), 290-311.

Abadie, A. and Imbens, G.W. (2002). Simple and bias-corrected matching estimators. Technical report, Department of Economics, University of California, Berkeley. http://emlab.berkeley.edu/users/imbens/.

Becker, S. O. and Ichino, A. (2002). Estimation of Average Treatment Effects Based on Propensity Scores. *The Stata Journal*, 2(4), 358–77.

Becker, S.O., Caliendo, M. (2007). Sensitivity analysis for average treatment effects. *The Stata Journal*, 7(1), 71-83.

Bia, M., Mattei, A. (2007). *Application of the Generalized Propensity Score. Evaluation of public contributions to Piedmont enterprises*. POLIS Working Paper 80, Department of Public Policy and Public Choice, Università del Piemonte Orientale, Alessandria.

Bia, M., Mattei, A. (2008). A Stata package for the estimation of the dose–response function through adjustment for the generalized propensity score. *The Stata Journal*, 8(3), 354-373.

Bia, M., Mattei, A. (2012). Assessing the Effect of the Amount of Financial Aids to Piedmont Firms Using the Generalized Propensity Score. *Statistical Methods and Applications*, 21 (4), 485-516.

Bia, M., Flores, C.A., Flores-Lagunes, A., Mattei, A. (2013). Stata package for the application of semiparametric estimators of dose-response functions. CEPS/INSTEAD, Working Paper No 2013-07, Luxemburg.

Cagliero, R., Cisilino, F., Scardera, A. (2010). *L'utilizzo della RICA per la valutazione di programmi di sviluppo rurale*. Roma: Rete Rurale Nazionale.

Cattaneo, M.D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155, 138–154.

Cattaneo, M.D., Drukker, D.M., Holland, A.D. (2013). Estimation of multivalued treatment effects under conditional independence. *The Stata Journal*, 13(3), 407–450.

Cattaneo, M.D., Farrell, M.H. (2011). Efficient Estimation of the Dose-Response Function under Ignorability using Subclassification on the Covariates. In Drukker, D.M. (ed.), *Missing Data Methods: Cross-sectional Methods and Applications*. Series: Advances in Econometrics, Volume 27A1, Emerald Group Publishing Limited, Bingley.

Chabé-Ferret, S. (2010). To Control or Not to Control? Bias of Simple Matching vs Difference-In-Difference Matching in a Dynamic Framework. Paper presented at the 10th World Congress of the Econometric Society, Shanghai, 17-21 August.

Chabé-Ferret, S., Subervie, J. (2013). How much green for the buck? Estimating additional and windfall effects of French agro-environmental schemes by DID-matching. *Journal of Environmental Economics and Management*, 65, 12–27.

De Castris, M. (2013). Are the R&D Subsdies Effective? An Empirical Analysis of the Italian Fund for Technological Innovation. Paper presented at the 54th Annual Conference of the Italian Economic Association (SIE), Bologna, 24-26 October.

De Filippis, F., Romano, D. (a cura di) (2010). *Crisi economica e agricoltura*. Gruppo 2013-

Coldiretti, Roma.

Di Porto, E., Elia, L., Tealdi, C. (2014). Undeclared Work in a Flexible Labour Market. CEIS, University of Rome Tor Vergata (mimeo).

DiPrete, T., Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*, 34, 271–310.

Duflo, E., Glennerster, R., Kremer, M. (2006). *Using Randomization in Development Economics Research: A Toolkit*. NBER Technical Working Papers 0333, Cambridge: National Bureau of Economic Research.

EENRD (2010). *Approaches for assessing the impacts of the Rural Development Programmes in the context of multiple intervening factors*. European Evaluation Network for Rural Development (EENRD), European Commission, Brussels.

Esposti, R. (2007). Regional Growth and Policies in the European Union: Does the Common Agricultural Policy Have a Countertreatment Effect? *American Journal of Agricultural Economics,* 89, 116–134.

Esposti, R. (2011a). La chiave e la luce: perché valutare la riforma del primo pilastro della PAC è difficile. *ARE – AgriRegioniEuropa*, 7 (25), 9-13

Esposti, R. (2011b), Reforming the CAP: an agenda for regional growth? In: Sorrentino, S., Henke, R., Severini, S. (eds.), *The Common Agricultural Policy after the Fischler Reform. National Implementations, Impact Assessment and the Agenda for Future Reforms,* Farnham: Ashgate, 29-52.

Esposti, R. (2011c). Evaluating the CAP Reform as a multiple treatment effect: evidence from Italian farms. Atti 122nd EAAE Seminar, *Evidence-Based Agricultural and Rural Policy Making: Methodological and Empirical Challenges of Policy Evaluation*, 17-18 Febbraio, Ancona.

Esposti, R., Listorti, G. (2013). Agricultural Price Transmission across Space and Commodities during Price Bubbles. *Agricultural Economics*, 44 (1), 125.

European Commission (2005). *Putting rural development to work for jobs and growth*. Special Edition Newsletter, Directorate-General for Agriculture and Rural Development, Brussels.

European Commission (2006). *Rural Development 2007-2013. Handbook on common monitoring and evaluation framework. Guidance document*. DG Agriculture and Rural Development, European Commission, Brussels.

European Commission (2011). The CAP in perspective: from market intervention to policy innovation. *Agricultural Policy Perspectives Briefs*, Brief nº 1, European Commission, Directorate-General for Agriculture and Rural Development, Brussels.

Flores, C.A., Mitnik. O.A. (2009). Evaluating Nonexperimental Estimators for Multiple Treatments: Evidence from a Randomized Experiment. Paper presented at the 2009 ASSA (Allied Social Science Association) meeting, San Francisco, 3-5 January.

Frascarelli A. (2008). *L'Ocm unica e la semplificazione della Pac*. Gruppo 2013-Coldiretti, Roma.

Frölich, M. (2004). Programme evaluation with multiple treatments. *Journal of Economic Surveys*, 18, 181-224.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66(2), 315–31.

Heckman, J. J. (2005). The Scientific Model of Causality. *Sociological Methodology*, 35(1), 1–97.

Heckman, J. J., Ichimura, H., Todd, P. (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies*, 64(4, Special Issue: Evaluation of Training and Other Social Programmes), 605–654.

Heckman, J. J., Ichimura, H., Smith, J., Todd, P. (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica*, 66, 1017–1099.

Hirano, K., Imbens, G.W. (2004). The propensity score with continuous treatment. In: Gelman, A., Meng, X.L. (eds.) *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. West Sussex: Wiley InterScience, 73-84.

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–960.

Ichino, A., Mealli, F., Nannicini, T. (2008). From Temporary Help Jobs to Permanent Employment: What Can We Learn From Matching Estimators and their Sensitivity? *Journal of Applied Econometrics*, 23, 305- 327.

Imbens, G.W. and Wooldridge, J.M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47 (1), 5–86.

Kirchweger, S., Eder, M., Kapfer, M., Kantelhardt, J. (2012). Evaluating measures for improving farm competiveness in the European Rural Development programme: a comparison of different matching approaches. Selected poster prepared for presentation at the IAAE Triennal Conference, Foz du Iguaçu, 18-24 August.

Kluve, J., Schneider, H., Uhlendorff, A., Zhao, Z. (2012). Evaluating continuous training programs using the generalized propensity score. *Journal of the Royal Statistical Society*, 175, 587-617.

Listorti, G., Esposti, R. (2012). Horizontal price transmission in agricultural markets: fundamental concepts and open empirical issues. *Bio-based and Applied Economics*, 1(1), 81-108.

Lukesch R., Schuh B. (eds.) (2010). Approaches for assessing the impacts of the Rural Development Programmes in the context of multiple intervening factors. Working Paper, March 2010, Findings of a Thematic Working Group established and coordinated by The European Evaluation Network for Rural Development, Brussels.

Magrini, E., Montalbano, P., Nenci, S., Salvatici, L. (2013). Agricultural trade distortions during recent International price spikes: what implications for food security? Paper presented at the 54th Annual Conference of the Italian Economic Association (SIE), Bologna, 24-26 October.

Martini, A., Sisti, M. (2009). *Valutare il successo delle politiche pubbliche*. Bologna: Il Mulino.

Michalek, J. (2012). Counterfactual impact evaluation of EU rural development programmes – Propensity score matching methodology applied to selected EU member states: vol. 1, a micro-level approach. Joint Research Center (JRC), Sevilla.

Moro, D., Sckokai, P. (2011). The impact of pillar I support on farm choices: conceptual and methodological challenges. Paper presented at the 122nd EAAE Seminar “Evidence-Based Agricultural and Rural Policy Making: Methodological and Empirical Challenges of Policy Evaluation”, Ancona (Italy), February 17-18.

Nannicini, T. (2007). Simulation–based sensitivity analysis for matching estimators. *The Stata Journal*, 7(3), 334-350 .

Nichols, A. (2007). Causal inference with observational data. *The Stata Journal*, 7(4), 507-541.

OECD (2011). *Evaluation of Agricultural Policy Reforms in the European Union*. Paris: OECD Publishing.

Povellato A., Velazquez B.E. (2005). *La riforma Fischler e l'agricoltura italiana*. Roma: INEA.

Powell, J.L. (1994). Estimation of Semiparametric Models. In: Engle, R.F., McFadden, D.L., *Handbook of Econometrics*. Volume IV, New York: Elsevier Science, 2443-2521.

Pufahl, A., Weiss, C. (2009). Evaluating the effects of farm programmes: results from propensity score matching. *European Review of Agricultural Economics*, 36 (1), 79-101.

Renwick, A., Revoredo-Giha, C. (2008). Measuring Cross-Subsidisation of the Single Payment Scheme in England. Paper presented at the 109th EAAE Seminar "The CAP after the Fischler Reform: National Implementations, Impact Assessment and the Agenda for Future Reforms", Viterbo (Italy), November 20-21.

Rosenbaum P. (1987). Sensitivity Analysis to Certain Permutation Inferences in Matched Observational Studies. *Biometrika*, 74(1), 13-26.

Rosenbaum P. (2002). *Observational studies*. New York: Springer Verlag.

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

Rossi, P.H., Freeman, H.E. (1993). *Evaluation. A Systematic Approach*. London: SAGE Publications.

Rubin, D. (1974). Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies. *Journal of Educational Psychology*, 66, 688-701.

Salvioni, C., Sciulli, D. (2011). Impact of Rural Development Policy and Less Favoured Areas Scheme: A Difference in Difference Matching Approach. Paper presented at the EAAE 2011 Congress, Change and Uncertainty Challenges for Agriculture, Food and Natural Resources, August 30-September 2, Zurich.

Sorrentino, S., Henke, R., Severini, S. (eds.) (2011). *The Common Agricultural Policy after the Fischler Reform. National Implementations, Impact Assessment and the Agenda for Future Reforms,* Farnham: Ashgate.

Smith, J., Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators. *Journal of Econometrics*, 125 (1–2), 305–353.

Sotte, F. (2005). La natura economica del PUA. *AgriRegioniEuropa* (*ARE*), 1(3), 15-18.

Sotte, F. (2006). Imprese e non-imprese nell'agricoltura Italiana. *Politica Agricola Internazionale*, 1/2006, 13-30.

Sotte, F., Arzeni, A. (2013). Imprese e non-imprese nell'agricoltura italiana. *AgriRegioniEuropa* (*ARE*), 9(32), 65-70.

StataCorp (2013). *Stata Treatment-Effects Reference Manual: Potential Outcomes/Counterfactual Outcomes*. Stata Press, College Station, Texas.

Villa, J.M. (2012). Simplifying the estimation of difference in differences treatment effects with Stata. MPRA Paper No. 43943, University Library of Munich.

Winters, P., Maffioli, A., Salazar, L. (eds.) (2011). Evaluating the Impact of Agricultural Projects in Developing Countries. Special Feature of the *Journal of Agricultural Economics*, 62(2), 393-492.

ANNEX

Table A1 - Abbreviations of policy measures in Figure 2

AC01	Accompanying measure: integrated agriculture (2078/92)
AC02	Accompanying measure: organic agriculture (2078/92)
CMO01	Common Market Organization: arable crops' compensatory payment (1251/99 and others)
CMO02	Common Market Organization: arable crops' supplementary payment (1251/99)
CMO05	Common Market Organization: other crops
CMO06	Common Market Organization: bovine special premium (1254/99)
CMO07	Common Market Organization: suckler cow premium (1254/99)
CMO13	Common Market Organization: sheep premium (3013/89)
CMO15	Common Market Organization: fruits - investments
CMO16	Common Market Organization: durum wheat special quality premium (1782/03)
CMO21	Common Market Organization: dairy premium (1782/03)
CMO23	Common Market Organization: supplementary quality aid for arable crops (art.69, 1782/03)
MOD01	Modulation: supplementary payment (art.12 1782/03)
RDP01	Rural Development Plan: investments in agricultural holdings
RDP02	Rural Development Plan: settlement of young farmers
RDP03	Rural Development Plan: training
RDP06	Rural Development Plan: low environmental impact
RDP07	Rural Development Plan: organic farming
RDP08	Rural Development Plan: breeds in danger of being lost to farming
RDP13	Rural Development Plan: afforestation of agricultural land
RDP42	Rural Development Plan: compensatory payment for less favour areas
REG01	Regional measure: other payments for livestock activities
REG08	Regional measure: rehabilitation and prevention for livestock activities
SFP01	Single Farm Payment (1782/03)
SFP02	Mandatory set-aside

Table A2 - Type of Farm (TF) reclassification (TF_R)

TF 2000 4-digit Classification (TF)	TF reclassified (TF_R)	TF 2000 4-digit Classification (TF) *(continues)*	TF reclassified (TF_R) *(continues)*
1310	1	5011	6
1320	1	5012	6
1330	1	5013	6
1410	2	5021	6
1420	2	5022	6
1430	2	5023	6
1441	2	5031	6
1442	2	5032	6
1443	2	6010	5
2011	7	6020	5
2012	7	6030	5
2013	7	6040	5
2021	7	6050	5
2022	7	6061	5
2023	7	6062	5
2031	7	7110	6
2032	7	7120	6
2033	7	7210	6
2034	7	7220	6
3110	4	7230	6
3120	4	8110	5
3130	4	8120	5
3141	4	8130	5
3143	4	8140	5
3211	4	8210	5
3212	4	8220	5
3213	4	8231	5
3220	4	8232	5
3230	4		
3300	4		
3400	4		
4110	3		
4120	3		
4210	3		
4220	3		
4310	3		
4320	3		
4410	3		
4420	3		
4430	3		
4440	3		

Table A3 - Linear regression of the outcome variables on the covariates and the treatment in the two alternative specifications: binary treatment (T = 0,1) and continuous treatment (TI = treatment intensity) (robust standard errors in parenthesis)

Covariates: / Outcome variable:	y^1	y^2	y^3	y^4
Binary Treatment (T)				
Constant term	0.3149* (0.0495)	0.0103* (0.0016)	26114 (14178)	-0.5052* (0.1821)
AGE (of the holder) (years)	-0.0042* (0.0005)	-0.0000 (0.0000)	-280.8* (106.3)	-0.0036 (0.0020)
Altitude (ALT) (m)	0.0006* (0.0000)	0.0001* (0.0000)	0.7704 (8.042)	0.0000 (0.0001)
Annual Working Units (AWU)	-0.0086* (0.0018)	-0.0002* (0.0001)	-3874 (2170)	0.0078* (0.0034)
Economic Size (ES) (classes)	0.0040 (0.0041)	0.0002 (0.0002)	-68.70 (1636)	0.0278* (0.0139)
Fixed Costs (on Net Value Added) (FC)	-0.0001 (0.0001)	-0.0000 (0.0000)	-41.08 (34.00)	0.0003* (0.0001)
Horse Power (HP)	0.0002* (0.0001)	0.0001* (0.0000)	-28.32 (32.52)	0.0000 (0.0001)
Livestock Units (LU)	0.0005* (0.0001)	0.0001 (0.0000)	-56.26 (48.08)	0.0001 (0.0001)
Utilized Agricultural Area (UAA) (ha)	-0.0001* (0.0000)	0.0001* (0.000)	2.154 (2.638)	0.0001 (0.0001)
Type of Farm (reclassified) (TF_R)	-0.0142* (0.0050)	-0.0010* (0.0002)	-2922 (1668)	0.0202 (0.0144)
T (0, 1)	0.2950* (0.0180)	0.0074* (0.0006)	8199 (6687)	0.1598* (0.0612)
Continuous Treatment (TI)				
Constant term	0.7028* (0.0489)	0.0159* (0.0015)	34260* (13647)	-0.4250* (0.1708)
AGE (of the holder) (years)	-0.0036* (0.0005)	-0.0001* (0.0000)	-278* (106)	-0.0039 (0.0021)
Altitude (ALT) (m)	0.0007* (0.0000)	0.0001* (0.0000)	2.196 (8.218)	0.0000 (0.0001)
Annual Working Units (AWU)	-0.0128* (0.0019)	-0.0002* (0.0001)	-3931 (2155)	0.0085* (0.0033)
Economic Size (ES) (classes)	0.0036 (0.0041)	0.0001 (0.0001)	-149.6 (1671)	0.0241 (0.0137)
Fixed Costs (on Net Value Added) (FC)	-0.0001* (0.0000)	-0.0000 (0.0000)	-42.43 (33.87)	0.0002* (0.0001)
Horse Power (HP)	0.0003* (0.0001)	0.0001* (0.0000)	-26.64 (32.77)	0.0000 (0.0001)
Livestock Units (LU)	0.0005* (0.0001)	0.0001* (0.0000)	-54.99 (48.05)	0.0001 (0.0001)
Utilized Agricultural Area (UAA) (ha)	-0.0001 (0.0001)	0.0000 (0.0001)	2.2036 (2.6393)	0.0000 (0.0001)
Type of Farm (reclassified) (TF_R)	-0.0508* (0.0053)	-0.0011* (0.0002)	-3418* (1560)	0.0261 (0.0174)
TI	-0.0054* (0.0008)	0.0001* (0.0000)	27.67 (189.3)	0.0058* (0.0028)

*Statistically significant at 0.05 level

Table A4 - Estimated GPS (linear regression specification) and balancing tests on the 7 treatment intervals (standard deviations in parenthesis)[a]

Specification:	*Without RDP dummy*		*With RDP dummy*	
	Mean difference	*t-value*	*Mean difference*	*t-value*
Treatment interval 1 (TI = 0.001-4.998):				
GPS[b]	0.093		0.093	
AGE (of the holder) (years)	0.413 (0.501)	0.824	0.396 (0.505)	0.783
Altitude (ALT) (m)	-39.414 (35.381)	-1.114	-30.253 (29.627)	-1.021
Annual Working Units (AWU)	-0.645 (0.936)	-0.689	-0.611 (0.920)	-0.665
Economic Size (ES) (classes)	0.051 (0.078)	0.659	0.046 (0.078)	0.589
Fixed Costs (on Net Value Added) (FC)	0.845 (0.731)	1.155	0.740 (0.729)	1.016
Horse Power (HP)	27.983 (18.539)	1.509	27.914 (18.490)	1.510
Livestock Units (LU)	-29.542 (18.448)	-1.601	-28.061 (18.544)	-1.513
Utilized Agricultural Area (UAA) (ha)	5.680 (3.015)	1.884	6.635 (3.870)	1.715
Type of Farm (reclassified) (TF_R)	-0.172 (0.091)	-1.890	-0.166 (0.096)	-1.725
RDP	-	-	-0.010 (0.015)	-0.664
Treatment interval 2 (TI = 5.005-9.992):				
GPS[b]	0.171		0.171	
AGE (of the holder) (years)	1.275 (0.866)	1.472	1.470 (0.866)	1.698
Altitude (ALT) (m)	-10.795 (10.208)	-1.058	-15.322 (10.292)	-1.489
Annual Working Units (AWU)	-0.128 (0.123)	-1.042	-0.140 (0.123)	-1.139
Economic Size (ES) (classes)	0.174 (0.096)	1.813	0.142 (0.086)	1.650
Fixed Costs (on Net Value Added) (FC)	-2.282 (1.289)	-1.770	-2.143 (1.175)	-1.826
Horse Power (HP)	-27.791 (19.536)	-1.433	-33.662 (19.543)	-1.722
Livestock Units (LU)	-33.899 (20.840)	-1.627	-36.441 (20.745)	-1.757
Utilized Agricultural Area (UAA) (ha)	-7.219 (4.226)	-1.708	-6.676 (4.217)	-1.583
Type of Farm (reclassified) (TF_R)	-0.006 (0.055)	-0.105	-0.027 (0.055)	-0.493
RDP	-	-	-0.038 (0.027)	-1.395
Treatment interval 3 (TI = 10.006-14.987):				
GPS[b]	0.166		0.165	
AGE (of the holder) (years)	-0.775 (0.581)	-1.333	-0.512 (0.580)	-0.883
Altitude (ALT) (m)	10.128 (10.821)	0.936	7.891 (10.793)	0.731
Annual Working Units (AWU)	0.193 (0.110)	1.755	0.174 (0.114)	1.525
Economic Size (ES) (classes)	0.169 (0.092)	1.842	0.160 (0.092)	1.749
Fixed Costs (on Net Value Added) (FC)	0.964 (0.903)	1.068	1.003 (0.907)	1.105
Horse Power (HP)	10.364 (9.309)	1.113	9.809 (9.410)	1.042
Livestock Units (LU)	6.578 (10.216)	0.644	5.613 (10.587)	0.530
Utilized Agricultural Area (UAA) (ha)	-1.679 (2.999)	-0.560	-2.345 (3.053)	-0.768
Type of Farm (reclassified) (TF_R)	-0.226 (0.150)	-1.507	-0.260 (0.151)	-1.719
RDP	-	-	-0.022 (0.018)	-1.229
Treatment interval 4 (TI = 15.003-19.966):				
GPS[b]	0.138		0.137	
AGE (of the holder) (years)	-1.222 (0.649)	-1.883	-1.159 (0.622)	-1.862
Altitude (ALT) (m)	-6.091 (11.927)	-0.511	-8.201 (11.954)	-0.686
Annual Working Units (AWU)	0.223 (0.121)	1.839	0.214 (0.122)	1.759
Economic Size (ES) (classes)	-0.012 (0.100)	-0.116	-0.024 (0.100)	-0.237
Fixed Costs (on Net Value Added) (FC)	0.762 (0.996)	0.765	0.710 (0.990)	0.717
Horse Power (HP)	12.577 (9.758)	1.289	10.872 (9.784)	1.111
Livestock Units (LU)	21.897 (11.525)	1.900	22.428 (11.529)	1.945
Utilized Agricultural Area (UAA) (ha)	-4.559 (3.084)	-1.478	-5.410 (3.104)	-1.743
Type of Farm (reclassified) (TF_R)	-0.021 (0.0318)	-0.649	-0.032 (0.034)	-0.938
RDP	-	-	-0.019 (0.020)	-0.961
Treatment interval 5 (TI = 20.016-24.992):				
GPS[b]	0.106		0.105	
AGE (of the holder) (years)	-3.193 (1.792)	-1.782	-3.054 (1.796)	-1.700
Altitude (ALT) (m)	-31.356 (17.548)	-1.787	-31.821 (17.461)	-1.822
Annual Working Units (AWU)	0.497 (0.259)	1.919	0.453 (0.268)	1.690
Economic Size (ES) (classes)	0.088 (0.129)	0.682	0.077 (0.129)	0.595
Fixed Costs (on Net Value Added) (FC)	-0.882 (1.280)	-0.689	-0.082 (1.339)	-0.061
Horse Power (HP)	30.648 (16.660)	1.840	25.942 (15.814)	1.640
Livestock Units (LU)	28.234 (14.496)	1.948	27.151 (15.449)	1.758
Utilized Agricultural Area (UAA) (ha)	-1.308 (3.539)	-0.369	-3.389 (3.598)	-0.942
Type of Farm (reclassified) (TF_R)	-0.062 (0.038)	-1.616	-0.097 (0.052)	-1.865
RDP	-	-	0.013 (0.026)	0.494

(Table A4 – continues)

Treatment interval 6 (TI = 25.000-29.995):				
GPS[b]	0.079		0.078	
AGE (of the holder) (years)	-1.729 (1.123)	-1.539	-2.226 (1.143)	-1.947
Altitude (ALT) (m)	-15.873 (22.287)	-0.712	-8.214 (22.257)	-0.369
Annual Working Units (AWU)	0.140 (0.243)	0.576	0.084 (0.248)	0.340
Economic Size (ES) (classes)	0.095 (0.185)	0.514	0.034 (0.187)	0.184
Fixed Costs (on Net Value Added) (FC)	0.132 (1.906)	0.069	0.062 (1.945)	0.032
Horse Power (HP)	40.301 (21.219)	1.899	38.416 (20.363)	1.887
Livestock Units (LU)	27.998 (22.027)	1.271	27.894 (22.601)	1.234
Utilized Agricultural Area (UAA) (ha)	-8.161 (4.483)	-1.820	-10.720 (5.573)	-1.923
Type of Farm (reclassified) (TF_R)	0.001 (.058)	0.024	-0.059 (0.062)	-0.937
RDP	-	-	0.021 (0.038)	0.548
Treatment interval 7 (TI = 30.037-69.864):				
GPS[b]	0.045		0.045	
AGE (of the holder) (years)	2.591 (1.962)	1.321	2.129 (1.850)	1.151
Altitude (ALT) (m)	6.789 (18.288)	0.371	7.089 (18.485)	0.383
Annual Working Units (AWU)	0.320 (0.213)	1.506	0.295 (0.218)	1.356
Economic Size (ES) (classes)	-0.206 (0.157)	-1.313	-0.221 (0.160)	-1.385
Fixed Costs (on Net Value Added) (FC)	0.910 (1.612)	0.564	0.880 (1.628)	0.541
Horse Power (HP)	5.857 (13.352)	0.439	2.858 (13.604)	0.210
Livestock Units (LU)	38.995 (21.085)	1.849	39.634 (21.560)	1.838
Utilized Agricultural Area (UAA) (ha)	-2.459 (3.338)	-0.737	-4.136 (3.411)	-1.213
Type of Farm (reclassified) (TF_R)	0.146 (0.091)	1.604	0.120 (0.085)	1.412
RDP	-	-	0.040 (0.032)	1.248

[a] Balancing test is performed on the common support
[b] Mean value

Table A5 - GPS function: MLM (Multinomial Logit Model) estimation of the discrete treatment on the covariates (standard errors in parenthesis)

Treatment level = 1	(base outcome)
Treatment level = 2	
Constant term	-0.5545 (0.1771)*
Altitude (ALT) (m)	0.0004 (0.0001)*
Utilized Agricultural Area (UAA) (ha)	0.0099 (0.0013)*
Livestock Units (LU)	-0.0002 (0.0002)
Horse Power (HP)	0.0020 (0.0003)*
Fixed Costs (on Net Value Added) (FC)	-0.0002 (0.0015)
AGE (of the holder) (years)	0.0003 (0.0028)
Annual Working Units (AWU)	-0.1678 (0.0239)*
Treatment level = 3	
Constant term	-1.179 (0.2036)*
Altitude (ALT) (m)	0.0000 (0.0002)
Utilized Agricultural Area (UAA) (ha)	0.0162 (0.0014)*
Livestock Units (LU)	-0.0009 (0.0003)*
Horse Power (HP)	0.0019 (0.0003)*
Fixed Costs (on Net Value Added) (FC)	-0.0138 (0.0107)
AGE (of the holder) (years)	0.0094 (0.0032)*
Annual Working Units (AWU)	-0.2221 (0.0308)*
Treatment level = 4	
Constant term	-1.491 (0.2144)*
Altitude (ALT) (m)	0.0000 (0.0002)*
Utilized Agricultural Area (UAA) (ha)	0.0189 (0.0014)*
Livestock Units (LU)	-0.0031 (0.0006)*
Horse Power (HP)	0.0027 (0.0003)*
Fixed Costs (on Net Value Added) (FC)	-0.0043 (0.0050)
AGE (of the holder) (years)	0.0132 (0.0033)*
Annual Working Units (AWU)	-0.2856 (0.0346)*
Treatment level = 5	
Constant term	-2.159 (0.2623)*
Altitude (ALT) (m)	0.0002 (0.0002)
Utilized Agricultural Area (UAA) (ha)	0.0217 (0.0015)*
Livestock Units (LU)	-0.0057 (0.0012)*
Horse Power (HP)	0.0034 (0.0004)*
Fixed Costs (on Net Value Added) (FC)	0.0007 (0.0026)
AGE (of the holder) (years)	0.0192 (0.0038)*
Annual Working Units (AWU)	-0.4235 (0.0533)*
Treatment level = 6	
Constant term	-2.793 (0.3204)
Altitude (ALT) (m)	0.0003 (0.0002)
Utilized Agricultural Area (UAA) (ha)	0.0235 (0.0015)*
Livestock Units (LU)	-0.0077 (0.0018)*
Horse Power (HP)	0.0036 (0.0004)*
Fixed Costs (on Net Value Added) (FC)	-0.0010 (0.0045)
AGE (of the holder) (years)	0.0191 (0.0047)*
Annual Working Units (AWU)	-0.3867 (0.0568)*
Treatment level = 7	
Constant term	-1.405 (0.2651)*
Altitude (ALT) (m)	0.0004 (0.0002)*
Utilized Agricultural Area (UAA) (ha)	0.0247 (0.0015)*
Livestock Units (LU)	-0.0168 (0.0024)*
Horse Power (HP)	0.0049 (0.0003)*
Fixed Costs (on Net Value Added) (FC)	-0.0127 (0.0130)
AGE (of the holder) (years)	0.0050 (0.0039)
Annual Working Units (AWU)	-0.5845 (0.0584)*

*Statistically significant at 0.05 level

Table A6 - Coefficient estimates of the conditional mean of the potential outcome: fully interacted quadratic specification

Outcome variable: Covariates (**X**):	y^1	y^2	y^3	y^4
ES				
3	-0.0509	-0.0085*	-25243	-0.6406*
4	-0.0582	-0.0061*	-28364*	-0.5774*
5	-0.0725*	-0.0074*	-27953*	-0.3289*
6	-0.0462	-0.0053*	-22262	-0.1483
7	-0.0239	-0.0041*	-28796*	-0.1757
8	-0.0094	-0.0024*	-26692*	-0.0922
9	-0.0349	-0.0006	-49110*	-0.0859
10	-0.0814*	-0.0001	56449*	0.02094
TF_R				
2	0.0251	0.0036*	5439	0.2773*
3	1.008*	-0.0036*	-14730*	0.0689
4	-0.1974*	-0.0095*	-13212	-0.0206
5	0.1757*	0.0009	-8857	0.1380
6	0.3598*	-0.0011	-35589*	0.5253*
7	-0.1432*	-0.0094*	-24660	-0.1447
ALT	-0.0000	0.0000	-15.45	0.0007
UAA	-0.0002	0.0001*	-13.53*	-0.0003
LU	0.00102*	-0.0000	-383.4*	-0.0008
HP	0.00014*	0.0000*	-25.30	-0.0004
FC	-0.0010	-0.0001	-53.54	-0.0008
AGE	-0.0005	0.0000	-452.9*	-0.0021
AWU	-0.0040	-0.0004*	-3579*	0.0143
ALT*ALT	0.0000*	0.0000	-0.0059	-0.0000
ALT*UAA	-0.0002	-0.0001	0.0537*	0.0003
ALT*LU	0.0000	0.0000	-0.0671	0.0000
ALT*HP	0.0000	0.0000	-0.0289	-0.0000
ALT*FC	0.0000	0.0000	0.1644	-0.0000
ALT*AGE	0.0000	0.0000	-0.0723	-0.0001*
ALT*AWU	0.0000	0.0000	18.70*	0.0000
UAA*UAA	0.0004*	0.0000*	0.0087*	-0.0000
UAA*LU	0.0000	-0.0000	0.0010*	0.0001
UAA*HP	0.0000	-0.0000	0.0005*	0.0000
UAA*FC	-0.0000*	0.0000	0.0172	-0.0000
UAA*AGE	0.0000	0.0000	0.1066	0.0000
UAA*AWU	0.0000	0.0000*	-0.5298*	-0.0001
LU*LU	0.0000 *	0.0000	0.0088	0.0000
LU*HP	1.027	0.0000	-0.2577*	0.0000
LU*FC	0.0000	-0.0000	-0.5892	0.0000
LU*AGE	-.0.0000	0.0000	10.23*	0.0001
LU*AWU	-0.0000	0.0000	-0.0199	-0.0002
HP*HP	-0.0000	-0.0000*	0.0114	0.0000
Constant	0.3970*	0.0233*	65681*	0.0188

*Statistically significant at 0.05 level. Standard errors are available upon request