https://doi.org/10.7896/j.1703 Studies in Agricultural Economics 119 (2017) 107-114

Alexandre N. ALMEIDA*[1] and Boris E. BRAVO-URETA**

Assessing the sensitivity of matching algorithms: The case of a natural resource management programme in Honduras

A fundamental challenge in impact evaluations that rely on a quasi-experimental design is to define a control group that accurately reflects the counterfactual situation. Our aim is to evaluate empirically the performance of a range of approaches that are widely used in economic research. In particular, we compared three different types of matching algorithms (optimal, greedy and nonparametric). These techniques were applied in the evaluation of the impact of the MARENA programme (*Manejo de Recursos Naturales en Cuencas Prioritarias*), a natural resource management programme implemented in Honduras between 2004 and 2008. The key findings are: (a) optimal matching did not produce better-balanced matches than greedy matching; and (b) programme impact calculated from nonparametric matching regressions, such as kernel or local linear regressions, yielded more consistent outcomes. Our impact results are similar to those previously reported in the literature, and we can conclude that the MARENA programme had a significant, positive impact on beneficiaries.

Keywords: impact evaluation, propensity scores, semiparametric models

* University of Sao Paulo/ESALQ, Av. Pádua Dias, 11, 13418-900 Piracicaba - São Paulo, Brazil. Corresponding author: alex.almeida859@gmail.com
** University of Connecticut, Storrs, Connecticut, USA.

Introduction

Similar to many Central American countries, the Honduran rural sector is characterised by low levels of production and income, which are attributed to a large proportion of landless or near landless rural workers, and low levels of farm family education (López and Valdés, 2000; Bravo-Ureta *et al*., 2011). In 2014, the Honduran agricultural sector contributed only 13.8 per cent to total GDP (WB, 2016). However, a significant part of the total population (44 per cent) lived in rural areas and 82 per cent were below the poverty line (ECLAC, 2009). Moreover, GEF-IFAD (2002) indicated that Honduran rural poverty is largely a consequence of unsustainable land use, which has led to environmental degradation, productivity losses, food insecurity and growing climatic vulnerability.

Recognising these major challenges, the international community has begun to re-adopt the old idea (Johnston and Mellor, 1961) that agricultural productivity growth is an essential component of any development strategy (WB, 2008). Moreover, it is now believed that policy efforts that focus on agricultural development can make a significant contribution to the Millennium Development Goals established by the United Nations in 2000, and to the more recent Sustainable Development Goals (SDSN, 2013; Sachs, 2015).

A key strategy to increase agricultural production and thereby income is the provision of agricultural extension services (Birkhaeuser *et al*., 1991; Anderson and Feder, 2007; WB, 2008). An effective diffusion of knowledge not only reduces the gap between laboratory experiments and farmers' fields, but also develops the skills necessary for good farm management practices and sustainable development (Winters *et al*., 2010).

Although the literature focusing on the evaluation of agricultural programmes in developing countries is growing, there are still very few quantitative studies assessing programme interventions for poverty in Central America (Bravo-Ureta *et al*., 2011). Rigorous measures of the impact of agricultural development programmes that target poor people are necessary not only to contribute to an emerging literature but they can also help donors and government agencies document the impact of their financial contributions and thus improve resource allocation (Heinrich *et al*., 2010; Petrikova, 2014).

Initially applied in medical sciences, treatment evaluation tools have become increasingly popular for analysing policy interventions across disciplines, particularly in economics. A central challenge of all these tools is how to define the counterfactual situation adequately (Ravallion, 2008). Ideally, one would have the outcome of interest for a group of individuals that has been treated and the outcome for the same group without treatment. Yet it is impossible to observe the same group with and without treatment at the same time. When the outcome of non-participants is used as a control, there is a real risk of selection bias that can overestimate or underestimate the impact of the treatment (Duflo *et al*., 2008). A well-executed randomised approach guarantees that, on average, there is no difference between treated and untreated subjects with respect to observable and unobservable characteristics (Ravallion, 2008). However, for technical and ethical reasons, randomised experimental studies in resource economics are difficult to implement (Ravallion, 2008; Heinrich *et al*., 2010). Thus, much of the evaluation work has relied on quasi-experimental designs, often incorporating propensity score matching (PSM) methodologies (WB, 2011).

The purpose of this study is to evaluate the impact of the MARENA (***Ma**nejo de **Re**cursos **Na**turales en Cuencas Prioritarias*) programme implemented in Honduras between 2004 and 2008. For this purpose, we conduct a detailed comparison of impact measures obtained from a range of propensity score functions and matching algorithms currently used in economic research. Overall, no single statistical method has emerged as the principal dominant or superior choice, and the number of applied studies that compares the performance of different matching techniques is very limited (Austin, 2013). In practice, researchers should select methods based on data characteristics to try to optimise the

[1] http://orcid.org/0000-0002-0680-5446

trade-off between the bias and variance of the estimators (Augurzky and Kluve, 2007; Austin, 2013). As a result, it is desirable to assess a variety of matching approaches to examine their robustness when evaluating a given intervention (Caliendo and Kopeinig, 2008; Ravallion, 2008; Imbens and Wooldridge, 2008).

Our analysis focuses on the MARENA programme, which financed activities designed to enhance agricultural production, productivity and the sustainable management of natural resources in predominantly poor rural agricultural areas in Honduras. Details of the programme can be found in Bravo-Ureta (2009) and Bravo-Ureta *et al*. (2011). The ultimate goal of MARENA was to reduce rural poverty while enhancing environmental sustainability. This paper extends the work reported by Bravo-Ureta *et al*. (2011), who relied on quasi-experimental data and traditional matching approaches along with difference-in-difference (DID) techniques, and showed that MARENA had a positive impact on its beneficiaries. An important attribute of MARENA, compared to other natural resource management projects, is that "… the collection of farm-level data to monitor and evaluate the programme was a priority from the beginning" (Bravo-Ureta *et al*., 2011, p.432). This feature offers high quality data assembled on a timely fashion, which makes it possible to conduct a robust evaluation that can then provide useful policy implications. Our goal here is to go beyond this earlier study by conducting an exhaustive analysis of robustness and performance for a variety of matching algorithms with different kinds of propensity scores that are not commonly applied in empirical studies (Khandker *et al*., 2009; Bravo-Ureta, 2014).

In summary, our main results corroborate the findings reported in Bravo-Ureta *et al*. (2011) who, using only two traditional matching techniques (one-to-one nearest neighbour (NN) and kernel regression), found impact estimates ranging from HNL[2] 16,425 to 25,575 in favour of the beneficiaries of the MARENA programme. In addition, based on balancing tests and the stability (i.e. similar of magnitudes) of impact estimates, we find that: (a) propensity scores coming from semiparametric estimation do not produce more robust impact estimates than propensity scores coming from logit or probit models; and (b) optimal matching does not lead to more robust impact estimates than the widely used greedy algorithm. These latter findings are consistent with what is expected based on conceptual grounds (Gu and Rosenbaum, 1993).

Methodology

Matching and quasi-experimental data

For evaluations where the objective is to measure the Average Treatment Effect on the Treated (ATET) and only quasi-experimental data are available, as in our case, it is necessary to generate a control group with observable characteristics for individuals that are as close as possible to those of the treated group (Khandker *et al*., 2009). To satisfy this requirement, the use of PSM has become a useful method of selecting controls to serve as 'perfect clones' of the treated subjects (Gertler *et al*., 2011). This selection is based on a set of observable characteristics (covariates) that are not affected by the treatment (Caliendo and Kopeinig 2008). In this manner, the model satisfies the conditional independence assumption and the common support assumption, as stated by Rosenbaum and Rubin (1983).[3] According to the latter authors, one of the advantages of using the PSM method is computational in nature, particularly when sample sizes are large and matching is time-consuming.

The goal of PSM consists essentially of finding the minimum distance between treated and untreated subjects given by the probability of an individual receiving treatment or not in a 'one dimensional vector' rather than relying on the whole set of observable characteristics (covariates) (Caliendo and Kopeinig, 2008). This minimum distance can be defined in various ways. The most straightforward used matching algorithm is the one-to-one NN method that can be executed with or without replacement of the treated and untreated observations based on the minimisation of the Euclidean distance (Caliendo and Kopeinig, 2008). Earlier, the one-to-one NN matching approaches were based on covariate means (known as Covariate Matching – CM), and were performed based on the Mahalanobis distance metric, a technique that is computationally cumbersome if the number of covariates is large (D´Agostino, 1998).

The NN algorithm, with and without caliper, has been widely implemented in impact evaluation studies and has been called 'greedy'. The main idea is that the odds of a treated unit finding its best match from a reservoir of controls are best for the 'early' units in the search; in other words, first come, first served as described by Rosenbaum (1989). Augurzky and Kluve (2007) explain that a greedy algorithm works by a random selection between treated and untreated units in terms of a specified distance. Once a treated unit finds its control, both are removed from the original sample, and the matching process continues. As a consequence, finding 'good' controls for treated units becomes increasingly difficult as the process unfolds. To overcome this problem, optimal matching has been proposed. This technique "works backwards and rearranges already matched units if some specific treated unit turns out to be a better (closer) match with a control unit previously matched to another treated unit" (Augurzky and Kluve, 2007, p.540). The idea is to attain the optimal minimum distance between treated and untreated units. To date, an empirical study of this matching approach that aims to analyse the impact of development interventions in the context of agriculture does not appear to exist.

In theory, optimal matching should overcome the shortcomings of greedy matching, such as the creation of bad 'late' matches. Gu and Rosenbaum (1993) evaluated the performance of optimal versus greedy matching programmes and found that optimal matching is superior to greedy matching only when the goal is to minimise the average Mahalanobis distance within pairs among covariates. Yet, optimal matching is no better at minimising propensity scores' distances or at producing balanced matched samples. Augurzky and Kluve (2007) tested the relative efficiency of greedy and optimal matching, along with different types

[2] HNL (Honduran Lempiras) 19.50 = USD 1.00 in 2012.

[3] Formal proofs of these assumptions can be found in Rosenbaum and Rubin (1983) and Imbens (2000).

of distance measures, to evaluate the time it takes for high school graduates to complete a bachelor's degree and found that the greedy choice produced a more favourable balancing of covariates than the optimal matching. More recently, Austin (2013) compared several matching algorithms using Monte Carlo simulations, with results very similar to those of Gu and Rosenbaum (1993). Austin (2013) found that if optimal matching resulted in samples in which the mean difference in the propensity scores is less between treated and control units compared to greedy matching, then balancing of covariates was not improved under optimal matching.

As far as we are aware, a good deal of discussion remains but no clear conclusions about the relative performance of the matching algorithms that are commonly used empirically, particularly the optimal matching. Moreover, both greedy and optimal matching systems share a limitation when the common support assumption is imposed (as it should be). In this case, some observations from the treated and/or untreated groups will be dropped, which can be a problem if the sample size is small. Heckman *et al*. (1997) proposed a partial solution to this problem that relied on estimating the treatment effect by comparing the outcome of interest of all treated individuals to a weighted average of the outcomes of all untreated individuals. This comparison is made using a standard nonparametric Nadaraya-Watson regression in which the propensity scores are used as weights.[4]

Regardless of the choice of the matching approach, it is imperative to verify if the balancing property holds. A simple and efficient way is to check the similarities between treated and untreated subjects using two different types of statistics widely used currently: standardised bias and p-values from a standard t-test between the means (D'Agostino, 1998; Lee, 2013).[5] The rule of thumb in such cases is that the standardised bias should not be higher than 20 per cent in absolute value, and p-values should be no lower than the 10 per cent level of statistical significance (Rosenbaum and Rubin, 1985). Moreover, a likelihood-ratio test of the joint significance of all the regressors and the pseudo R^2 after matching are also useful to check the balancing condition (Leuven and Sianesi, 2003; Sianesi, 2004; Caliendo and Kopeinig, 2008). In any case, if the matched sample does not turn out to be balanced, a new specification of the covariates should be considered (Heinrich *et al*., 2010).

Combining PSM and difference-in-difference

As already pointed out, a robust and accurate evaluation of the intervention is possible only if individual characteristics for non-participants are well matched with those of participants. Although matching can eliminate or substantially mitigate biases stemming from observed characteristics, it is possible that biases from unobserved time invariant characteristics, such as managerial skills and motivation of farmers, still remain (Gertler *et al*., 2011; Maffioli *et al*., 2013). As panel data are available for this study, we can combine the DID estimator with alternative propensity scores and the various matching algorithms (Khandker *et al*., 2009; Bravo-Ureta, 2014). The DID approach, as initially suggested by Heckman *et al*. (1998), measures the difference between the expected outcome of treated and control groups at the baseline (in our case 2003-2004) and the difference in the outcome at a point typically close to the end of the intervention (in our case 2007-2008), often referred to as the endline (Ravallion, 2008). The average treatment effect for the treated individual i using DID and combining PSM can be expressed as:

$$DID_i = (Y^T_{i,t} - Y^T_{i,t-1}) - \sum_{j \in c} \omega(i,j)(Y^C_{i,t} - Y^C_{i,t-1}) \quad (1)$$

where $\omega(i,j)$ is the weight (using PSM) given to the jth control individual matched to treated individual i, t is the endline, $t-1$ is the baseline, and T and C stand for treated and control respectively (Khandker *et al*., 2009).

Table 1: Definition of variables used in the analysis.

Variable	Unit	Definition
TVAO	HNL	Total value of agricultural output
BENEF	Dummy	1 if the household is a beneficiary of MARENA
NEIGHBOR	Dummy	1 if the household is not a beneficiary of MARENA and lives within its area of influence
AGLAND	Hectares	Total land devoted to agricultural production
DIVER	Dummy	1 if household produces crops in addition to maize and beans
CAFEECO	Dummy	1 if the household produces coffee using ecological practices
ALTITUD	Dummy	1 if the farm is located at an altitude higher than the mean
AGE	Years	Age of household head
EDUC	Years	Years of schooling of the household head
NUMBER	Number	Number of people in the household
ORGA	Dummy	1 if the household head participates in farmer organisations
ASSIST	Dummy	1 if the household receives technical assistance
YEAR	Dummy	0 = 2004, 1 = 2008

Source: own compilation

Implementation of the empirical analysis for the MARENA intervention

The implementation of the empirical analysis is as follows:

Step 1. Estimate a binary choice model to calculate the probability (propensity score) that the farmer is a beneficiary of MARENA, using data for the 2003-2004 baseline year. The function to be estimated can be written in general terms as:

$$BENEF = f(AGLAND, CAFEECO, NUMBER, ALTITUD, AGE, EDUC, ORGA, ASSIST, DIVER) \quad (2)$$

where $BENEF = 1$ if beneficiary and 0 if non-beneficiary. The covariates are defined in Table 1.

Step 2. Using the propensity score vectors from step 1, matched samples are constructed based on Euclidean distance using different algorithms without replacement.[6] Fig-

[4] Local linear matching is also a version of kernel matching and is implemented in the same fashion as the Heckman approach (see Caliendo and Kopeinig, 2008).

[5] The standardised bias is the size of the difference in the means of covariates between treated and untreated units, scaled by the square root of the average of their sample variances (Heinrich *et al*., 2010).

[6] Austin (2013) discourages the use of matching with replacement, because it seems to induce a higher mean square error (higher variance) of the estimated impact than matching without replacement.

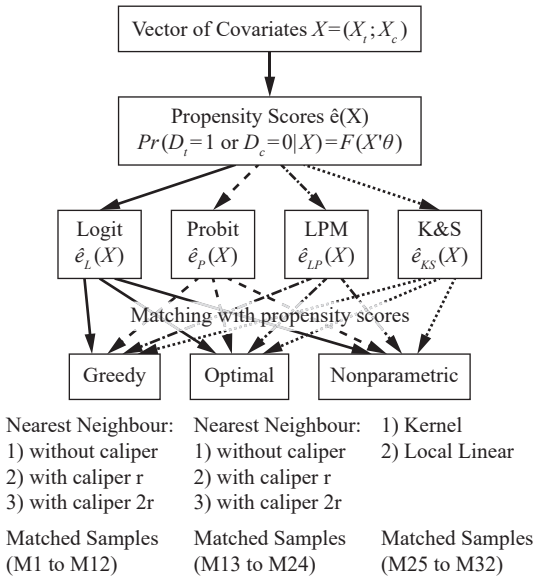

Figure 1: Matched sample generation process using propensity score vectors from the estimation of logit, probit, linear probability and Klein and Spady models.
Source: own composition

ure 1 shows all the combinations considered among matching algorithms and propensity scores generating a total of 32 matched samples. The common support condition is imposed in all cases.

Step 3. Check whether the covariates of treated and untreated units are balanced. If they are not, a new specification of the score function regarding covariates should be tested.

Step 4. Calculate the ATET by combining PSM from the 32 matched samples and DID method (equation 1). See Caliendo and Kopeinig (2008) for a detailed review on ATET.

Data

The data used in this research are from a two-round panel covering 366 households, of which 109 were beneficiaries of the MARENA programme, while the remaining 257 constitute the untreated or control group. Of the untreated group, 143 households (neighbours) are located within MARENA's area of intervention, and 114 households are located outside of that area (non-neighbours).[7] Data were collected during the 2003-2004 agricultural year (baseline) and then four years later, for the 2007-2008 production cycle (endline). The dataset includes information on socioeconomic and demographic household characteristics, alternative sources of income, and a detailed description of farm inputs, outputs, expenses and revenues. Table 2 reports the means and standard deviations of the MARENA programme for the agricultural year 2003-2004 (baseline) for beneficiaries versus non-beneficiaries. The key outcome of interest for the evaluation is the total value of agricultural output (TVAO). TVAO includes revenues from the production of maize, beans, coffee and horticultural crops and the value of any farm products consumed by the household. Before the programme, the TVAO (not shown) was much larger for the control group (around HNL 45,000) than for the treatment group (HNL 27,786).

Table 2: Group comparisons prior to matching (beneficiaries vs. control, baseline 2004).

Covariate	Treatment		Control		Two-sample t-statistic	Standardised difference (%)
	Mean	SD	Mean	SD		
AGLAND	1.80	1.26	2.25	2.29	1.91**	-24.07
CAFFECO	0.02	0.13	0.00	0.06	-1.40	13.76
NUMBER	6.20	2.68	5.96	2.52	-0.81	9.11
ALTITUD	0.47	0.50	0.53	0.50	1.00	-11.46
AGE	46.61	14.45	48.14	14.10	0.95	-10.78
EDUC	3.50	2.74	3.24	2.97	-0.77	8.91
ORGA	0.74	0.44	0.26	0.44	-9.60***	109.79
ASSIST	0.44	0.50	0.25	0.43	-3.78***	41.89
DIVER	0.52	0.50	0.44	0.50	-1.46	16.66
Observations	109		257			

For definitions of the variables see Table 1
***, **, * Significant at the 1%, 5% and 10% levels respectively
Source: own calculations

The last two columns of Table 2 contain statistics (t-test and the standardised bias difference in per cent) that were used to compare the treated and untreated groups with regard to observable characteristics before the matching at the baseline. As stated, large statistical differences among observable characteristics can lead to biased estimates of the real impact of the intervention. We observed that only three variables have shown such distortion in our sample. They are: (1) total land devoted to agricultural production (AGLAND); (2) participation in farmer organisations (ORGA); and (3) technical assistance (ASSIST). Therefore, special attention is given to these three variables below.

Results

Estimating propensity scores

In practice, discrete choice models, such as logit and probit, have been widely used to estimate propensity scores before matching (Cameron and Trivedi, 2005). On the other hand, Smith (1997) and Caliendo and Kopeinig (2008) argue that because propensity score models are used only for classification, a simple linear probability model (LPM) could also be used. However, one of the drawbacks of the LPM is that it is likely to yield predicted outcomes that lie outside of the common support condition, resulting in a loss of information (observations), thereby compromising the quality of the matching (Zhao, 2007). For our study, we also use the model developed by Klein and Spady (1993) that has the major advantage of relaxing the assumption that the error term follows a logistic (logit) or normal (probit) distribution, which can be restrictive and can produce inconsistent estimates in practice (Li and Racine, 2007). The coefficients of the semiparametric K&S model, logit, probit, and LP models are displayed in Table 3.

The next step is to calculate the predicted probabilities

[7] Similar to Bravo-Ureta *et al*. (2011), the spillover effect (indirect effect) of the MARENA programme between neighbours and non-neighbours was also investigated. Our estimates show that the spillover effect (on neighbour), although positive, was not statistically significant in any of our simulations. According to Bravo-Ureta *et al*. (2011), if the skills or incentives required to implement the farming practices by the programme are sufficiently complex, it is not unexpected that the knowledge diffusion between beneficiaries and non-participants neighbours might be inefficient.

(propensity scores). A simple Q-Q plot method[8], which compares the quantiles of these scores for the four models, indicate that there is no statistical difference between them. However, we find that propensity scores coming from different models do affect the magnitude of the final impact of the intervention after matching, as shown below.

The impact of the MARENA programme and robustness checking

Table 4 reports the impact of MARENA in HNL on the TVAO between 2003-2004 and 2007-2008 for 32 samples matched using different matching algorithms and distance measures. Matching is combined with the DID estimator and is applied in all cases. The ATET estimates are identified in Table 4 by a superscript with the capital letter M along with a number from 1 to 32 for each matched sample (See Figure 1 for a review of the matched sample generation process). The ATET results constructed using a one-dimensional vector are shown; that is, matching is performed only using the predicted propensity scores estimated from logit, probit, linear probability and K&S functions. We used these vectors of propensity scores to perform the matching based on the following algorithms: (1) greedy and optimal one-to-one NN with no caliper; (2) greedy and optimal one-to-one NN with caliper r, where r is one quarter of a standard deviation of the propensity score (Rosenbaum and Rubin, 1983); (3) greedy and optimal one-to-one NN with caliper 2r; (4) kernel regression; and (5) local linear regression.

The greedy matching, nonparametric kernel, and local linear were performed in a STATA do-file procedure (*psmatch2.ado*) published by Leuven and Sianesi (2003).[9] Another STATA do-file procedure (*optmatch2.ado*) developed by Mark Lunt at the University of Manchester was used for optimal matching.[10]

As mentioned above, three variables in our unmatched sample (Table 2) for the baseline year were unbalanced: AGLAND, ORGA and ASSIST. Of the total 32 matched samples constructed, eight (matched samples 1 to 4; 13 to 15) did not yield balance for one of the covariates (AGLAND or ORGA) and one matched sample (M16) exhibited two unbalanced covariates (ORGA and ASSIST), i.e., they did not pass the balancing tests (p-value < 0.10 and standardised bias < 20 per cent) after matching. Therefore, all the ATET estimates from these eight matched samples were omitted in the analysis in Table 4; The indicators of covariate balancing for these eight matched samples are shown in Annex 1.

Firstly, all statistically significant ATET estimates for matched data were higher than for the unmatched data (HNL 13,886, not shown). Secondly, based on the balancing tests and stability of coefficients (i.e. quite similar on magnitude values), we found more consistent matching results of the impact of the programme (1) under non-parametric matching approaches, whether kernel or local linear, and (2) under the greedy algorithms, particularly when a caliper is imposed.

Table 3: Logit, probit, LPM, and K&S results for participation in the MARENA programme using baseline data (2004) (N=366).

Covariate	Logit	Probit	LPM	K&S (1993)
	Coefficient			
AGLAND	-0.375***	-0.215***	-0.048***	-0.380***
	(0.094)	(0.056)	(0.011)	(0.061)
CAFFECO	4.035***	2.346**	0.578**	4.604***
	(1.034)	(1.114)	(0.243)	(0.732)
NUMBER	0.038	0.025	0.006	0.042***
	(0.050)	(0.031)	(0.008)	(0.008)
ALTITUD	-0.474*	-0.259*	-0.076*	-0.607***
	(0.281)	(0.160)	(0.043)	(0.097)
AGE	-0.013	-0.008	-0.002	-0.036***
	(0.010)	(0.006)	(0.002)	(0.006)
EDUC	-0.035	-0.015	-0.006	-0.089***
	(0.050)	(0.029)	(0.008)	(0.013)
ORGA	2.296***	1.340***	0.418***	3.423***
	(0.295)	(0.163)	(0.044)	(0.535)
ASSIST	0.673**	0.403**	0.117**	0.524***
	(0.290)	(0.168)	(0.047)	(0.086)
DIVER	0.523*	0.280*	0.068	0.721***
	(0.299)	(0.167)	(0.045)	(0.116)
CONSTANT	-1.056*	-0.635	0.270**	-
	(0.624)	(0.404)	(0.108)	
LR $chi^2(9)$	78.95***	105.36***	14.42***	
Wald $chi^2(9)$				46.19***
Pseudo R^2	0.242	0.236	0.248	0.221
Log likelihood	-169.947	-170.21		-158.307

For definitions of the variables see Table 1
***, **, * Significant at the 1%, 5%, and 10% levels respectively; SE are shown in parentheses
Source: own calculations

Table 4: The impact of the MARENA programme on total value of agricultural output in HNL constructed from matched samples using propensity score (PS) vectors from the estimation of logit, probit, linear probability and K&S models.

Outcome: $TVAO = TVAO_t - TVAO_{t-1}$	PS (Logit)	PS (Probit)	PS (LPM)	PS (K&S)
(1)	(3)	(4)	(5)	(6)
Greedy matching				
NN with no caliper£	FBT^{M1}	FBT^{M2}	FBT^{M3}	FBT^{M4}
NN with caliper (0.06)†£	$18,629^{M5}$ (10,072)**	$18,153^{M6}$ (10,034)*	$20,594^{M7}$ (9,710)**	$18,390^{M8}$ (9,877)*
NN with caliper (2x0.06)†£	$20,408^{M9}$ (9,860)**	$18,948^{M10}$ (9,813)*	$20,427^{M11}$ (9,534)**	$17,120^{M12}$ (9,895)*
Optimal matching				
NN with no caliper£	FBT^{M13}	FBT^{M14}	FBT^{M15}	FBT^{M16}
NN with caliper (0.06)†£	$23,126^{M17}$ (11,313)*	$8,594^{M18}$ (1,068)	$16,091^{M19}$ (9,404)*	$18,188^{M20}$ (11,224)*
NN with caliper (2x0.06)†£	$19,963^{M21}$ (10,770)*	$12,548^{M22}$ (10,893)	$24,263^{M23}$ (10,278)**	$22,649^{M24}$ (11,558)*
Nonparametric matching				
Kernel€	$19,845^{M25}$ (9,852)**	$18,977^{M26}$ (9,811)*	$20,661^{M27}$ (9,519)**	$17,000^{M28}$ (9,822)*
Local linear€	$17,882^{M29}$ (9,933)*	$17,231^{M30}$ (9,886)*	$18,414^{M31}$ (9,855)*	$17,736^{M32}$ (9,970)*

The DID estimator is combined with PSM; SE are shown in parenthesis and superscripts identify the matched samples; FBT stands for failed balancing tests
***, **, * Significant at the 1%, 5% and 10% levels respectively
Notes: £ One-to-one matching; †Size of the caliper used is a quarter of a standard deviation of the propensity score as suggested by Rosenbaum and Rubin (1983). For the four models, the standard deviation ranged from 0.2315 to 0.2426; € The optimal bandwidths (Logit=0.067; Probit=0.066; LPM=0.065; K&S=0.081) were calculated based on the rule of thumb of Silverman (1986)
Source: own calculations

[8] Not reported here but available from the authors upon request.

[9] Stata v.13 has introduced a new *teffects* command for estimating ATE and ATET with the advantage compared to *psmatch2* for which standard errors take into in account that propensity scores are estimated rather than known (http://www.ssc.wisc.edu/sscc/pubs/stata_psmatch.htm). One limitation, however, is that the procedure does not allow the use of propensity scores different from those estimated using logit or probit models. Moreover, the non-replacement option is also not allowed.

[10] http://personalpages.manchester.ac.uk/staff/mark.lunt

Discussion

In this paper, we extended the study by Bravo-Ureta *et al*. (2011) who found that the MARENA programme had a significant positive impact on beneficiaries. To corroborate those earlier results, and given that there is no agreed-upon best approach, we used several matching approaches designed specifically for quasi-experimental data (Caliendo and Kopeinig, 2008). We also tried to extend our analysis by relaxing the major assumptions imposed in the logit and probit models that are widely used to calculate propensity scores, and also by comparing different algorithms (e.g. greedy versus optimal). We observed that the use of propensity scores from a semiparametric estimation (Klein and Spady), for example, might provide estimated coefficients that are quite similar to those obtained using the logit, probit or linear probability models. Nevertheless, as stated by Smith (1997), and Caliendo and Kopeinig (2008), because the goal of propensity scores is only classification, the choice of the model to be estimated might not be crucial. Overall, our evaluation lends support to the positive impacts reported in the literature for a family of natural resource management interventions that have been implemented in recent years in Latin America (Solís *et al*., 2009, Cavatassi *et al*., 2011) and to similar programmes that are currently under preparation.

We did not corroborate the hypothesis, coming from the theoretical literature, that optimal matching produces 'better-balanced' matched samples and consequently more stable results than the greedy matching. We found that the balancing property holds equally well for both the greedy and optimal algorithms, particularly when calipers are imposed. However, based on the stability of the various impact estimates, we did find significant differences in terms of ATET values when propensity scores are compared, particularly in the same optimal matching approach. Moreover, the ATET calculated from nonparametric regressions, such as kernel or local linear, not only presented very consistent outcomes but also satisfied the balancing property for all selected covariates.

One of the potential reasons that optimal matching has no advantage over greedy matching in producing balanced matched samples is because both methods select more or less the same controls (Gu and Rosenbaum, 1993). We also find the optimal matching generates unstable results across differently estimated propensity scores than greedy matching even while they have equally better-matched samples. That being said, our findings support the view that the way propensity scores are estimated, whether parametrically or semiparametrically (K&S model), matters, particularly when these scores are used to execute the optimal matching. For example, by comparing the same impact from different propensity scores based on the same matching algorithm, it is clear that there is greater variability across the estimates from one-to-one optimal matching with caliper r (from HNL 8,594 (M18) to HNL 23,126 (M17)) and one-to-one optimal matching with caliper 2r (from HNL12,548 (M22) to HNL 24,263 (M23)) than the estimates from the greedy algorithms with caliper r and 2r (ATET estimates between HNL 17,120 (M12) and HNL 20,594(M7). Such lower variability was also reported by Bravo-Ureta *et al*. (2011) who, testing a greedy NN matching technique with a caliper arbitrarily chosen at 0.05, found impact estimates ranging from HNL 16,425 to HNL 20,654. Only a logit model based on the same covariate specification, as this study does, was used to generate their vector of propensity scores.

Moreover, in our dataset, two of the estimates (M18 and M22) from the optimal matched samples (NN with caliper of 0.06 and caliper of 0.12 using propensity scores from the probit model) are not statistically significant at the 10 per cent level, even though they have passed the balancing tests.

We also found that when the two non-parametric matching regressions, kernel and local linear, are tested and compared using different propensity scores, the impact estimates are similar in magnitude to greedy algorithms with calipers. The respective ATET non-parametric results (shown at the bottom of Table 4) on the TVAO are not only statistically significant and vary in value in a narrow range from 17,000 to 20,661.

Our results point out that analysts should not neglect the application of greedy algorithms and the use of a caliper during matching as also a way to impose common support and consequently avoiding bad matches (Gu and Rosenbaum, 1993; Caliendo and Kopeinig, 2008). Augurzky and Kluve (2007) and Austin (2013), using Monte Carlo simulations to examine different algorithms, found that when a caliper is used, it is possible to achieve balance, both for continuous and binary covariates, as we found for all matching approaches used. Note that the imposition of too narrow a caliper can result in the loss of observations and, consequently, an increased variance of the estimates so that non-parametric approaches (kernel or local linear) arise as an advantage because they avoid the loss of observations. In fact, based on our analysis, when the common support condition was imposed, between 8 and 14 of 109 treated observations needed to be discarded when performing the matching, depending on the propensity score vector used.

In practice, as seen in the applied literature, clearly some variability in the results are, to some extent, expected because matching techniques are implemented differently; and results depend on characteristics of the data under analysis (Caliendo and Kopeinig, 2008; Ravallion, 2008; Heinrich *et al*., 2010). Moreover, we should keep in mind that an estimator that works well in simulations does not necessarily behave in the same manner in applications with real data, which was the major motivation for this study.

To the best of our knowledge, comparisons of greedy versus optimal matching, although discussed in the literature, have not been well documented and, therefore, warrant further attention in both theoretical and applied work.

References

Anderson, J.R. and Feder, G. (2007): Agricultural extension, in R. Evenson and P. Pingali (eds), Handbook of Agricultural Economics volume 3, 2344-2378. Amsterdam: Elsevier.

Augurzky, B. and Kluve, J. (2007): Assessing the performance of matching algorithms when selection into treatment is strong. Journal of Applied Econometrics **22** (3), 533-557. https://doi.org/10.1002/jae.919

Austin, P.C. (2013): A comparison of 12 algorithms for matching on the propensity score. Statistics in Medicine **33** (6), 1057-1069.

https://doi.org/10.1002/sim.6004

Birkhaeuser, D., Evenson, R.E. and Feder, G. (1991): The economic impact of agricultural extension: a review. Economic Development and Cultural Change **39** (3), 607-650. https://doi.org/10.1086/451893

Bravo-Ureta, B.E. (2009): Evaluación final de resultados del programa MARENA, informe de terminación de proyecto [Final evaluation of the results of the MARENA programme, project completion report]. Washington DC: The Inter-American Development Bank.

Bravo-Ureta, B.E., Almeida A.N., Solís, D. and Inestroza, A. (2011): The economic impact of Marena's investments on sustainable agricultural systems in Honduras. Journal of Agricultural Economics **62** (2), 429-448. https://doi.org/10.1111/j.1477-9552.2010.00277.x

Bravo-Ureta, B.E. (2014): Stochastic frontiers, productivity effects and development projects. Economics and Business Letters **3** (1), 51-58. https://doi.org/10.17811/ebl.3.1.2014.51-58

Caliendo, M. and Kopeinig, S. (2008): Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys **22** (1), 31-72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Cameron, A.C. and Trivedi, P.K. (2005): Microeconometrics: methods and applications. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511811241

Cavatassi, R., Salazar, L., González-Flores, M., and Winters, P. (2011): How do agricultural programmes alter crop production? Evidence from Ecuador. Journal of Agricultural Economics **62** (2), 403-428. https://doi.org/10.1111/j.1477-9552.2010.00279.x

D'Agostino, R.B. (1998): Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Statistics in Medicine **17** (19), 2265-2281. https://doi.org/10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B

Duflo, E., Glennerster, R., and Kremer, M. (2008): Using randomization in development economics research: a toolkit, in T. Schultz and J. Strauss (eds), Handbook of Development Economics volume 4, 3895-3962. Amsterdam: Elsevier.

ECLAC (2009): CEPALSTAT: Bases de datos y publicacionnes estadisticas [Statistical databases and publications]. Santiago: The Economic Commission for Latin America and the Caribbean.

GEF-IFAD. (2002): Tackling land degradation and desertification. Technical Report. Roma: Global Environmental Facility and International Fund for Agricultural Development.

Gertler, P.J., Martinez, S., Premand, P., Rawlings, L.B. and Vermeersch, C.M.J. (2011): Impact Evaluation in Practice. Washington DC: World Bank.

Gu, X.S. and Rosenbaum, P.R. (1993): Comparison of multivariate matching methods: structures, distances, and algorithms. Journal of Computational and Graphical Statistics **2** (2), 405-420. https://doi.org/10.1080/10618600.1993.10474623

Heckman, J.J., Ichimura, H. and Todd, P.E. (1997): Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Review of Economic Studies **64** (4), 605-54. https://doi.org/10.2307/2971733

Heckman, J.J., Ichimura, H., Smith, J. and Todd, P.E. (1998): Characterizing selection bias using experimental data. Econometrica **66** (5), 1017-98. https://doi.org/10.2307/2999630

Heinrich, C., Maffioli, A. and Vázquez, G. (2010): A primer for applying propensity-score matching. Technical Notes 161. Washington DC: Inter-American Development Bank.

Imbens, G.W. and Wooldridge, J.M. (2008): Recent developments in the econometrics of programme evaluation. NBER Working Papers 14251. Cambridge MA: National Bureau of Economic Research.

Johnston, B.F. and Mellor, J.W. (1961): The role of agriculture in economic development. American Economic Review **51** (4), 566-593.

Khandker, S.R., Koolwal, G.B. and Samad, H.A. (2009): Handbook on Impact Evaluation: Quantitative Method and Practices. Washington DC: World Bank. https://doi.org/10.1596/978-0-8213-8028-4

Klein, R.W. and Spady, R.H. (1993): An efficient semiparametric estimator for binary response models. Econometrica **61**, 387-421. https://doi.org/10.2307/2951556

Lee, W.S. (2013): Propensity score matching and variations on the balancing test. Empirical. Economics **44** (1), 47-80. https://doi.org/10.1007/s00181-011-0481-0

Leuven, E. and Sianesi B. (2003): PSMATCH2: STATA module to perform full mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing, Statistical Software Components Series No. S432001. Boston MA: Boston College.

Li, Q. and Racine, J.C. (2007): Nonparametric Econometrics: Theory and Practice, Princeton NJ: Princeton University Press.

López, R. and Valdés, A. (2000): Fighting rural poverty in Latin America: new evidence of the effects of education, demographics, and access to land. Economic Development and Cultural Change **49** (1), 197-211. https://doi.org/10.1086/452497

Maffioli, A., Ubfal, D., Vazquez-Bare, G. and Cerdan-Infantes, P. (2013): Improving technology adoption in agriculture through extension services: evidence from Uruguay. Journal of Development Effectiveness **5** (1), 64-81. https://doi.org/10.1080/19439342.2013.764917

Petrikova, I. (2014): The short-and long-term effects of development projects: evidence from Ethiopia. Journal of International Development **26** (8), 1161-1180. https://doi.org/10.1002/jid.3035

Ravallion, M. (2008): Evaluating Anti-Poverty Programmes, in T. Schultz and J. Strauss (eds), Handbook of Development Economics volume 4, 3787-3846. Amsterdam: Elsevier.

Rosenbaum, P.R. and Rubin, D.B. (1983): The central role of the propensity score in observational studies for causal effects. Biometrika **70** (1), 41-55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P.R. and Rubin, D.B. (1985): Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician **39** (1), 33-38. https://doi.org/10.1080/00031305.1985.10479383

Rosenbaum, P.R. (1989): Optimal matching in observation studies. Journal of the American Statistical Association **84** (408), 1024-1032. https://doi.org/10.1080/01621459.1989.10478868

Sachs, J.D. (2015): The Age of Sustainable Development. New York: Columbia University Press. https://doi.org/10.7312/sach17314

SDSN (2013): Solutions for sustainable agriculture and food systems. Technical report for the post-2015 development agenda. Paris and New York: Sustainable Development Solutions Network.

Sianesi, B. (2004): An evaluation of the Swedish system of active labor market programs in the 1990s. Review of Economics and Statistics **86** (1), 133-155. https://doi.org/10.1162/003465304323023723

Silverman, B.W. (1986): Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability. London: Chapman and Hall. https://doi.org/10.1007/978-1-4899-3324-9

Smith, H.L. (1997): Matching with multiple controls to estimate treatment effects in observational studies. Sociological Methodology **27** (1), 325-353. https://doi.org/10.1111/1467-9531.271030

Solís, D., Bravo-Ureta, B.E. and Quiroga, R. (2009): Technical efficiency among peasant farmers participating in natural resource

management programmes in Central America, Journal of Agricultural Economics **60** (1), 202-219. https://doi.org/10.1111/j.1477-9552.2008.00173.x

Winters, P., Salazar, L. and Maffioli, A. (2010): Designing impact evaluations for agricultural Projects. Impact Evaluation Guidelines: Technical Notes 198. Washington DC: The Inter-American Development Bank.

WB (2008): World Development Report: Agriculture for Development. Washington DC: World Bank.

WB (2011): Impact Evaluations in Agriculture: An Assessment of the Evidence. Washington DC: World Bank.

WB (2016): World Bank development indicators. Washington DC: World Bank.

Zhao, Z. (2007): Sensitivity of propensity score methods to the specifications. Economics Letters **98** (3), 309-319. https://doi.org/10.1016/j.econlet.2007.05.01

Annex

Annex 1: Covariate balancing tests between treated and control farmers of the MARENA programme for eight matched samples which failed the balancing tests.

Matched sample	Comparisons	Covariates									Pseudo R^2	Pr>χ^2
		AGLAND	CAFEECO	NUMBER	ALTITUD	AGE	EDUC	ORGA	ASSIST	DIVER		
M1	p-value (t-test)	0.55	0.00	0.68	0.67	0.80	0.18	0.07	0.47	0.89	0.029	0.432
	SB %	-6.40	0.00	5.80	-6.00	3.60	-20.00	27.30	10.70	2.00		
M2	p-value (t-test)	0.44	0.00	0.35	1.00	0.75	0.30	0.07	0.39	0.89	0.028	0.452
	SB %	-8.10	0.00	13.10	0.00	4.60	-15.10	27.30	12.90	-2.00		
M3	p-value (t-test)	0.08	0.32	0.41	0.68	0.94	0.42	0.17	0.31	0.78	0.037	0.249
	SB %	-29.60	-9.40	11.50	-5.90	1.00	-11.50	20.10	14.80	4.00		
M4	p-value (t-test)	0.05	0.32	0.57	0.89	0.48	0.34	0.44	0.13	0.67	0.032	0.390
	SB %	-27.20	-10.00	-8.50	-2.10	10.70	-13.60	12.00	22.60	-6.30		
M13	p-value (t-test)	0.82	0.16	0.60	0.59	0.80	0.44	0.00	0.22	0.69	0.011	0.946
	SB %	-3.02	19.25	7.18	-7.31	3.47	-10.46	39.06	16.84	5.48		
M14	p-value (t-test)	0.86	0.16	0.51	0.79	0.75	0.68	0.00	0.17	1.00	0.019	0.776
	SB %	-2.40	19.25	8.99	-3.66	4.40	-5.61	40.96	18.77	0.00		
M15	p-value (t-test)	0.12	0.56	0.39	0.50	0.83	0.93	0.02	0.34	0.59	0.020	0.736
	SB %	-20.44	7.84	11.71	-9.14	-2.92	-1.26	31.45	13.03	7.31		
M16	p-value (t-test)	0.15	0.56	0.40	0.59	0.93	0.77	0.01	0.02	0.59	0.011	0.946
	SB %	-19.73	7.84	-11.54	-7.31	1.22	-3.92	35.25	32.73	7.31		

SB: standardised bias (Rosenbaum and Rubin, 1985)
Source: own calculations