129th EAAE Seminar, Castelldefels (Barcelona)
July 12th – 13th 2012

Coping with Heterogeneity in Empirical Rural Development Programme Evaluation: a Matching Approach

Stefan Kirchweger and Jochen Kantelhardt[1]

Abstract

The heterogeneity of farms and the problem of self-selection are challenging the evaluation of treatments in agriculture. This is particularly the case for rural development measures whit voluntary participation and heterogeneous outcomes. But knowledge about the selection mechanisms for a certain treatment, in combination with econometric methods, can help to overcome these problems. One of these promising methods is the *Propensity Score Matching* (PSM) approach. In this paper we apply PSM in order to obtain treatment effects from the agricultural investment support programme in Austria on the farm income. We also test the robustness of the results to hidden bias with sensitivity analysis. Furthermore we split the sample in more homogenous subsamples in order to increase the robustness of the results. The results show that treatment effects differ by a large amount for the subsamples and that splitting leads to slightly more robust results.

Key words: Rural Development programmes, heterogeneity, causal effects, *Propensity-Score Matching,* sensitivity analysis

JEL classifications: Q10, Q12, Q18

[1] *Stefan Kirchweger and Jochen Kantelhardt are with the Institute of Agricultural and Forestry Economics at the University of*

1 Introduction

There are about 187,000 farms located in Austria for the year 2007 (BMLFUW, 2011). Even though there have been structural changes and adaptations in the last few decades, the farms differ in farm structure and production systems. Figure 1 illustrates this heterogeneity by plotting the livestock and utilised agricultural area of around 1,600 farms. The heterogeneity is mainly due to the fact of different site conditions, i.e. mountainous or non-mountainous regions, as well as being the result of farm-manager characteristics or strategies. Furthermore, analyses in agriculture have to take into account that a farm is always built upon a unique relationship between the farm household and the farm enterprise. The heterogeneity of farm units and the unique relationship between farm and farm households leads to heterogeneous responses to support programmes (Pufahl and Weiss, 2009). This results in methodological challenges for researchers in carrying out quantitative analyses in the framework of Rural Development evaluation.

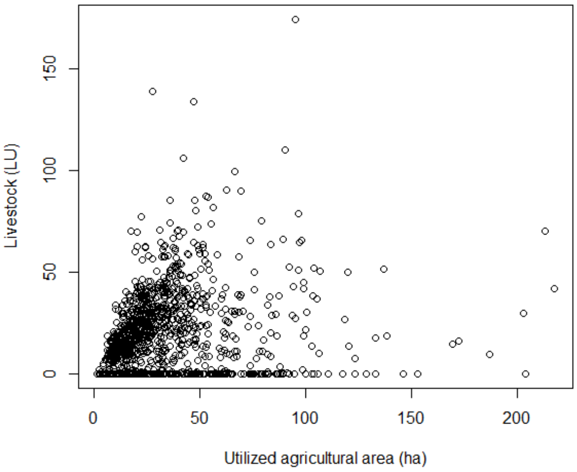

Figure 1: Livestock and utilised agricultural area for a sample of 1,600 Austrians farms.

Quantitative evaluation asks for the causal effect. Therefore the Neyman-Rubin-Holland model, also known as the counterfactual model (Morgan and Winship, 2009), the Neyman-Rubin model (Sekhon, 2009) or Roy-Rubin model (Caliendo and Hujer, 2006) has been developed. The model was originally introduced by Neyman (1923) and is nowadays used in a wide range of topics for microeconomic evaluation (Sekhon, 2009). Under this model the causal effect (Δ_A) for one individual (A) is computed by comparing the outcome in the state of participation (Y_A^1) and the outcome in the state without participation (Y_A^0). This can be formulated as $\Delta_A = Y_A^1 - Y_A^0$. But a fundamental challenge arises, as one of these outcomes is counterfactual because one unit can either be participant or non-participant. When we look for counterfactual for treated units, one solution to this problem is the use of observable non-participants. The treatment effect can then be computed by simply comparing treated and non-treated units. But to follow causal claims, treatment must be independent of the potential outcome and treated and non-treated must be homogenous, only differing by the analysed variable. If these are not fulfilled, the results are biased and/or have high variability. This is not a major issue in randomised experiments, as randomisation of treatment insures the independence of treatment and outcome. To reduce variability, the pairing of treated and untreated units can be used and number of observations can be increased (Rosenbaum, 2005a).

As experiments can hardly been used in agricultural treatment evaluation, we have to rely on observational data (Henning und Michalek, 2008). Observational studies differ from experiments, as the researcher cannot control the assignment of treatment to individuals (Rosenbaum, 2010, 65). Therefore, participants select themselves voluntarily for a certain treatment, which leads to a selection bias in the results. This bias is mainly due to variables (X) disturbing the causal inference of the treatment (T) on the outcome (Y) and therefore violates the independence assumption. Figure 2 illustrates a causal relationship between the treatment T and the outcome Y, but Y is biased through the mutual dependence of T and Y on X.

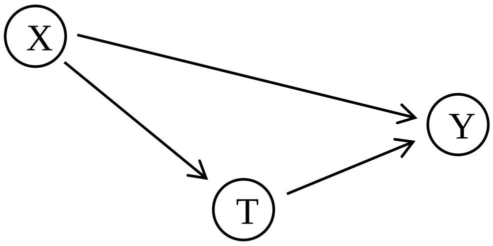

Figure 2: A causal diagram in which the effect of T on Y is disturbed through the back-door path, a mutual dependence on Z. (Source: Morgan and Winship, 2009)

As in heterogeneous observational studies, the increase in observations cannot reduce variability; more homogenous samples are needed (Rosenbaum, 2005a). Therefore the pairing of treated and untreated is needed to reduce both, bias and variability. One approach of pairing is *Propensity Score Matching* where treated and untreated are paired on similar propensity scores. Rubin and Rosenbaum (1983) prove that *matching* on the propensity score is sufficient. As with *Matching*, we only check for observable covariates; there always might be hidden bias caused by unmeasured variables.

The basic objective of this paper is to apply a *Propensity Score Matching* approach to analyse their ability to scope with heterogeneity in agricultural studies. This is exemplified on the agricultural investment support programme in Austria and its effects on the farm income of farms using the time period 2005-09. Therefore further analysis is implemented to reduce, on the one hand, the bias from unobservables and, on the other, to measure the robustness of the results regarding hidden biases. Furthermore we stratify the sample in dairy and granivore farms in order to obtain more homogenous samples and reduce variability as well as increase the robustness of the result. The following specific questions are asked:

- Can *Propensity Score Matching* be a supportive tool to derive causal effects from a farm investment support programme in empirical studies?
- How does *Propensity Score Matching* cope with heterogeneity in agriculture?
- Can bias be reduced by using smaller, more homogenous samples?

In Section 2 we give a brief introduction in farm investment and in the farm-investment support programme of Austria. Section 3 explains the methodological procedure and the database used in this paper. The results of this three-step approach are then displayed in Section 4. This section also includes the application of sensitivity analysis in order to judge on the causality of the different results. The results are discussed in Section 5.

2 Farm investment and the farm-investment support programme in Austria

The farm-investment programme is part of the second pillar of the Common Agriculture Policy and basically concerns improving competitiveness, work conditions, animal welfare and environmental conditions. To achieve these goals, 576 million Euros have been spent in Austria in the period from 2000 to 2009 (Dantler et al., 2010). The number of fostered farms during this period is slightly above 37,000, all mainly located in mountainous regions (see Figure 3). Consequently, forage farms (including mainly dairy and suckler-cow farms) are the main beneficiary of farm-investment payments, with a share of more than 56%. In contrast, in the distribution of farm type of all farms in Austria, forage farms have only a share of 37% (BMLFUW, 2011). In addition, there is an over-representation of granivore farms in contrast to field-crop farms. It is therefore not surprising that more than 50% of these funds foster the construction of barns mainly for dairy farming. Even though participants are mainly mountainous farms, it illustrates a low share of participants in the western federal states of Tyrol and Vorarlberg. This might be due to specific achievements by the federal states.

Furthermore, on average the share of participating farms increases for bigger farms. Hence the means of participants and non-participants differ, especially for the utilised agricultural area (UAA), total livestock units (LU) and milk quota (Dantler et al., 2010). As farm-investment support payments can only be obtained with an investment, and there is hardly any farm investment without support, we have to consider them jointly when evaluation is carried out (see Dirksmeyer et al., 2006 and Dantler et al., 2010). Therefore we also have to consider investment decisions in our analysis. A study done for German farms also points out that investing farmers have a lower share of equity and are older than non-investing farmers (Läpple, 2007). It is evident, therefore, that there has been a selection for participation based on structural and regional variables such as region, farm type, farm size and financial variables.

3 Methodological Approach

For the application of *matching* we use a three-step approach, where we first define the *matching* covariates and estimate the propensity score for the whole sample as well as for the subsamples of dairy, cash crop and granivore farms. Secondly, we match treated and controls based on the propensity score using a suitable Greedy algorithm with calliper *Matching*. As a last step, we calculate the average treatment effect on the treated with a difference-in-difference estimator for all samples. Afterwards sensitivity analysis is applied to judge on the quality of M*atching*.

3.1 *The Propensity Score Matching* approach

Matching follows the Conditional Independence Assumption (CIA) in order to find an adequate control group. Based on the work of Rubin (1977) and Rubin and Rosenbaum (1983), the CIA assumes that under a given vector of observable covariates (Z), the outcome (Y) of one individual is independent of treatment: $\{Y0, Y1 \perp\!\!\!\perp T\} |Z$, where $\perp\!\!\!\perp$ denotes independence. The *matching* procedure is based on conditioning on all covariates influencing T and/or Y (Z_1, Z_2, Z_3,….Z_k). This conditioning on Z should, on the one hand, lead to a reduction in selection bias in the form of a reduced correlation (r) between the errorterm of the treatment T (u) and the errorterm of the outcome Y (e) (see Figure 3).

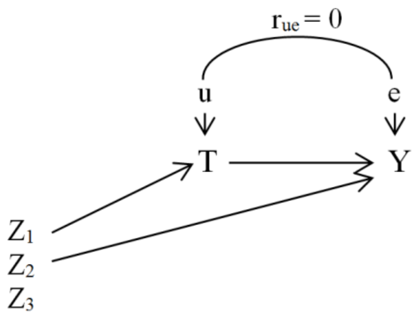

Figure 3: Identification of causal effects through conditioning on observed variables. (Source: Gangl, 2006)

Thus, through *matching* the income of farms are independent of whether the farm participated in the farm-investment programme or not. However, this requires the identification of all those covariates which influence the outcome and the probability of participation but are not influenced by programme participation. The selection of covariates is the most important task in the *matching* procedure. Guidance can be gained from statistical, economical and also practical background in order to choose the appropriate covariates. The influence of the participation on the covariates can be avoided by *matching* on farm variables before the start of treatment.

Another major assumption which needs to be fulfilled is the so-called Common Support Assumption. This basically requires the existence of non-participants having similar Z to all participants. Violation arises especially when covariates are used which predict too well the probability of treatment, but this is simply detected by visual control (Lechner, 2001). Losing observations because of missing common support is not usually a problem when these are not too numerous but might change the quantity of the results.

In order to identify similar controls, PSM use the propensity score ($p(Z)$) of each individual instead of each single covariate. The propensity score is defined as the probability of participation ($\Pr(T=1)$) for one individual given the observed covariates Z, independent of observed participation: $p(Z) = \Pr(T_i=1 \mid Z_1, Z_2, Z_3,\ldots.Z_k)$. Rubin and Rosenbaum (1983) prove that *matching* on the propensity score is sufficient. *Propensity-Score Matching* (PSM) differentiates from exact *matching* as the values of covariates are usually different within the pairs with the same propensity score but are balanced in the treated and control group (Rosenbaum, 2010, 166). The estimation of the propensity score (PS) is commonly based on the fitted values of a binary logit or probit model, using observed treatment assignment (yes/no) as the explained and Z as the explanatory variable. The model must not be linear but may include interactions, polynomials and transformations of the covariates.

There are several algorithms available to pair controls and treated units. In our paper we use a *Greedy algorithm* with calliper pair *matching* without replacement approach. Similarity is therefore established by using a self-defined calliper. A non-participant which is found within the calliper serves as control for one treated and cannot be used as another control. The treated unit is dropped when there is no control available within the calliper. Through this the quality of *matching* rises, as the controls are much more similar in contrast to simple *Nearest Neighbour Matching (*Caliendo and Kopeinig, 2008) and the condition of common support can be fulfilled. Augurzky und Kluve (2004) argue that callipers which are not too narrow are preferable when the heterogeneous effects of treatment are expected (Augurzky und Kluve, 2004). Therefore we set the calliper to 0.2 for our application.

Through the two steps, the estimation of the propensity score and the actual *matching* using a radius algorithm, pairs consisting of participants and controls are built, and a control group which is similar to the participant group is generated. This results in a reduction of systematic mean differences between these groups. Furthermore, *matching* on $\pi(Z)$ does not touch the Y variable until the estimation of the treatment effects in order to prevent it from new biases (Ho et al., 2007). Thus, the average treatment effect on the treated (ATT) can be computed, as the difference of the mean outcome of participants (Y_A^1) and controls (Y_B^0):

$$ATT=\sum_{A=1}^{n}(Y_A^1 \mid p(Z))/n_i - \sum_{B=1}^{n}(Y_B^0 \mid p(Z))/n_j \tag{1}$$

Matching can then be considered successful when the mean of the covariates between treated and control group is balanced. Balance is judged by conventional testing; alternatively, Ho et al. (2007) recommend using QQ-plots which plot the quantiles of a variable of the treatment group against that of the control group in a square plot (Ho et al., 2007). The *matching* algorithm in our analysis is run with the R-package "*Matching*" by J.S. Sekhon (see Sekhon, 2011).

As the independent assumption in *matching* is built on observable covariates, it is often criticised that there might still be hidden bias in the outcome, coming from unobserved variables. Therefore we implement a *difference-in-difference (DiD)* followed by sensitivity analysis considering the amount of hidden bias in the result.

3.2 Estimation of treatment effects

Smith and Todd (2005) recommend for controlling for unobservable covariates the implementation of a *DiD* estimator. The *DiD* relies on the assumption that the differences of participants and non-participants are similar at every time. It is computed as the difference of the progress of the participant and the non-participant from one point before (t') to one point after (t) the time of treatment (t_T) (Heckmann et al., 1998). By implementing the factor time and the before- and after-estimation in the analyses, we can monitor for unobservable, linear and time-invariant effects such as price fluctuations (Gensler et al., 2005). The combination of *matching* and *DiD* results in the *Conditional difference-in*-difference (CDiD) estimation and the used formula can be written as

$$ATT = \sum_{i=1}^{n}(Y_{A,t} - Y_{A,t'}) \,\Big|\, p(Z)/n_A - \sum_{j=1}^{n}(Y_{B,t} - Y_{B,t'}) \,\Big|\, p(Z)/n_B \qquad t' < t_T < t \qquad (2)$$

For our analysis, the pre-treatment situation is in 2003, post-treatment is 2010 and the treatment itself took place between 2005 and 2009. The two-year gap before treatment is necessary, since the year of treatment is the year of payment and the investment usually happens a year or two before payment.

3.3 Sensitivity analysis

In order to investigate the reliability of the results we implement a sensitivity analysis in our model. Therefore we use the so called Rosenbaum bounds (see Rosenbaum 2002, 2005b and 2010). Basically this sensitivity analysis tests for the robustness of results and models. Rosenbaum's approach in particular focuses on the hidden biases from unobservable variables and therefore on the violation of the assumption of independence of treatment and outcome or random assignment of treatment after *matching*. There is hidden bias, when pairs look comparable in their observable characteristics but differ in their actual probability (π) of receiving the treatment.

To measure the departure from random assignment of treatment the parameter Γ is implemented in the odds ratio of the pairs. There is no departure if the odds ($\pi/1$-π) of each unit do not differ within the pair and the Γ=1. When the units k and j have the same probability, the odds ratio was at most:

$$\frac{1}{\Gamma} \leq \frac{\pi j/(1-\pi j)}{\pi k/(1-\pi k)} \leq \Gamma \qquad (3)$$

The parameter of one is given in randomised experiment, but in observational studies this hardly ever appears. If the parameter happens to be 2, this indicates that one of these units is twice as likely to receive the treatment as the other.

It is not possible to compute the parameter; therefore we assume a perfect situation, with a positive treatment and no hidden bias, but we are ignorant of these facts, and perform a sensitivity analysis (Rosenbaum, 2010, 259). In order to start, one selects a series of values for Γ. Then we can either judge the robustness on the p-values and see how the p-value changes for increasing values of Γ or how the magnitude of the treatment effects changes with an higher Γ. High sensitivity to hidden bias appears if the conclusions change for values of Γ just slightly higher than one and low sensitivity is given if the conclusions change at large values of Γ (Rosenbaum, 2005b). The sensitivity analysis in our paper is based on the Wilcoxon sign rank test and the Hodges-Lehmann (HL) point estimate for the sign rank test with an upper and lower bound.[2] The values and estimates of these tests might differ to our results as they deal differently with outliners. We use the R-package "rbounds" by L. Keele (see Keele, 2010).

3.4 Data

We use data from 2000 to 2010 of 1,636 voluntary bookkeeping farms in Austria, where we find 239 farms who only participated in the farm-investment support programme at least once between 2005 and 2009 and 845 farms who did not participate between 2000 and 2010. Farms which did not attend in the years 2000-2004 and 2010, as well as those which received less than 5000 Euros in payments, were dropped from the analysis. Participants and non-participants are matched with data based on the year 2003.

In observational studies, better results can be achieved, when samples are more homogenous (Rosenbaum, 2005a). In order to gain more homogenous samples we split the sample in three subsamples, for dairy and granivore farms. Whereas dairy farms are characterised as farms keeping dairy cows and granivore farms are farms whose sales are mainly due to fattening pigs and steers as well as breeding and fattening hens. We then apply the three-step approach for all three subsamples individually.

4 Empirical Results

The results for the three-step estimation of the average treatment effect on the treated applied in the case of farm-investment support in Austria are displayed in this chapter. Furthermore we show the results of sensitivity analysis and stratification.

[2] A detailed derivation is given in Peel and Makepeace (2009).

4.1 Estimation of the Propensity Score

In order to get the propensity scores of each unit we apply a binary logit model. In our model we include a multinomial variable for the farm type and whether the farm is located in the region west, south and north, a dummy variable for organic farming and metric variables for the age of the farm manager, the labour, the utilised agricultural area (UAA), the share of rented UAA, the livestock density, the share of equity and the non-farm income. The estimates for the coefficients are displayed in Table 1. The results indicate that dairy farms, farms with higher labour and livestock density, as well as more UAA and non-farm income, are more likely to invest and receive farm-investment support but cash-crop farms and farms with older managers are less likely. The model correctly predicts about 78% of the farms attending the programme and is statistically significant at the 0.1% level or better, as measured by the likelihood ratio test.

Table 1: Covariates estimates of logit-models explaining programme participation for the whole sample.

	Estimate	Std. Error	z value	
Intercept	-5.928	1.075	-5.514	***
Dummy permanent crop farms	0.708	0.458	1.546	
Dummy forage farms (exclusive dairy)	-0.030	0.485	-0.061	
Dummy cash-crop farms	-0.639	0.334	-1.911	.
Dummy dairy farms	0.453	0.237	1.910	.
Dummy granivore farms	0.403	0.314	1.284	
Dummy region south	-0.130	0.207	-0.628	
Dummy region west	-0.319	0.291	-1.096	
Dummy konv farming	-0.080	0.215	-0.373	
Age	-0.022	0.009	-2.453	*
Labour	0.565	0.126	4.487	***
Utilised agricultural area (log)	0.713	0.153	4.644	***
Share of rented land	0.587	0.372	1.579	
Livestock density	0.586	0.179	3.270	**
Share of equity	0.801	0.508	1.577	
Non-farm income (log)	0.140	0.039	3.548	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Using this model we estimate the bounded propensity score for each farm, which is the basis for the following *matching* step. The distribution of the propensity scores is quite similar in the treated and the control group (see Figure 4). This is necessary in order to find good matches.

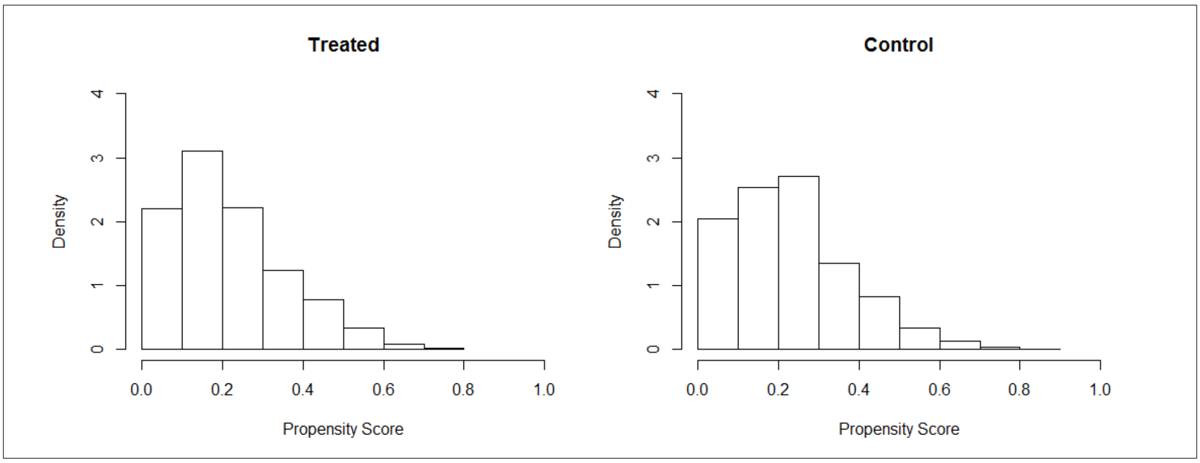

Figure 4: Distribution of propensity scores for treated (left) and controls (right).

4.2 Results from *Matching* and treatment effect estimation

The quality of the *matching* algorithm is based on the achieved balance between treated and control group. The applied Greedy algorithm has the best results regarding the *matching* balance in comparison to other algorithms. Out of 239 potential participants, the *matching* procedure develops a new sample with 227 pairs consisting of one treated and one control. Through this, the sample increased its balance between the two groups (participants and controls) for all variables, which are not statistically significantly different, using conventional levels and the t-test, anymore (see Table 2).

Table 2: Mean values of variables for participants and controls before and after Propensity-Score Matching for the whole sample.

	Potential participants	Potential controls		Selected participants	Selected controls
Number of farms	239	810		227	227
Dummy permanent crop farms	0.050	0.059		0.048	0.048
Dummy forage farms (exclusive dairy)	0.029	0.033		0.031	0.035
Dummy cash-crop farms	0.130	0.279	***	0.137	0.159
Dummy dairy farms	0.452	0.307	***	0.454	0.441
Dummy granivore farms	0.163	0.095	**	0.145	0.163
Dummy region south	0.247	0.247		0.233	0.225
Dummy region west	0.100	0.088		0.101	0.093
Dummy konv farming	0.816	0.819		0.815	0.837
Age	52.280	54.207	**	52.595	51.907
Labour	1.824	1.487	***	1.777	1.814
Utilised agricultural area (log)	3.488	3.309	***	3.465	3.484
Share of rented land	0.287	0.242	**	0.280	0.294
Livestock density	1.125	1.125	***	1.106	1.106
Share of equity	0.905	0.905		0.911	0.903
Non-farm income (log)	7.466	7.375		7.409	7.265
Livestock (log)	3.038	2.344	***	3.003	2.976
Dairy cows (log)	1.549	1.094	***	1.559	1.535
Pigs (log)	1.837	1.363	**	1.769	1.860

t-test for equally of means: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

With the new sample of 227 pairs gained from *matching* approach the ATT is computed by comparing the mean development of the farm income from 2003 to 2010 of participants and controls. This results in an ATT for the farm income of 7197 Euros, which can be interpreted as the amount of farm income which treated farms could increase more than controls. The ATT has a standard error of 2656.4 and t-statistic of 2.71, which indicates a statistical significant difference between the means at the 1% level or better.

4.3 Sensitivity analysis

Even though the ATT for the farm income is positive, we cannot be sure that controlling for observable covariates is enough to draw causal conclusions. Therefore we apply sensitivity analysis to test the robustness of the result. The results of this analysis are displayed in Table 3. The first column of Table 3 is the value of the parameter Γ, which should indicate the difference in the odds of farm participating or not caused by an unobserved variable. In the second and third column the upper and lower bound of the p-value from Wilcoxon Sign ranking test and the fourth and fifth the upper and lower bound of the Hodges-Lehmann point estimates for the sign rank test is shown. In the first row the parameter is set to one, assuming total randomisation through *matching*. The sensitivity analysis shows that through the increase of Γ up to 1.08, the upper bound of the p-value exceeds the 5%-level. This indicates that the result is highly vulnerable to unobserved bias. This also leads to a widening of the HL treatment estimates and therefore increasing the uncertainty through selection bias. When the parameter increases to 1.38, the HL treatment effect is even shown to become negative.

Table 3: Rosenbaum bounds parameters for the results of the whole sample

parameter (Γ)[1]	Wilcoxon p-value		HL treatment estimate	
	Lower bound[2]	Upper bound[3]	Lower bound[4]	Upper bound[5]
1.00	0.021	0.021	4265	4265
1.02	0.015	0.029	4012	4520
1.04	0.011	0.038	3752	4788
1.06	0.008	0.049	3466	5046
1.08	**0.006**	**0.063**	**3230**	**5266**
1.10	0.004	0.079	2938	5521
1.12	0.003	0.098	2682	5807
1.14	0.002	0.119	2449	6036
1.16	0.001	0.143	2213	6255
1.18	0.001	0.169	1995	6468
1.20	0.001	0.198	1752	6712
1.22	0.000	0.229	1519	6911
1.24	0.000	0.262	1302	7134
1.26	0.000	0.297	1060	7340
1.28	0.000	0.333	864	7609
1.30	0.000	0.370	659	7840
1.32	0.000	0.408	458	8052
1.34	0.000	0.446	253	8285
1.36	0.000	0.484	64	8481
1.38	0.000	0.522	-95	8678
1.40	0.000	0.558	-260	8903

[1] Odds of differential assignment due to unobserved factors
[2] Lower bound significance level (on assumption of under-estimation of treatment effect).
[3] Upper bound significance level (on assumption of over-estimation of treatment effect).
[4] Lower bound point estimate (on assumption of under-estimation of treatment effect).
[5] Upper bound point estimate (on assumption of over-estimation of treatment effect).

4.4 Results for stratified subsamples

The subsamples consist of 108 participants and 249 non-participants in the dairy subsamples and 39 treated and 77 non-treated in the granivore subsample. An individual logit model is applied for each subsample. The models are adapted by farm type-specific covariates. The estimates and significance levels of the model can be seen in Table 4. Thus, we included the share of dairy cows in the dairy subsample and the number of pigs variable in the granivore subsample. The estimation shows that in both models these additional covariates are not statistically significant but we are convinced that they play a major role in the decision to participate in the investment support programme (see also Dantler et al., 2010). Furthermore the estimates in both models are similar to the model with the whole sample except for the fact that labour and age are not statistically significant anymore. The models correctly predict about 70% and 76% respectively of the farms attending the programme and both are statistically significant at the 0.1% level or better, as measured by the likelihood ratio test.

Table 4: Covariates estimates of logit-models explaining programme participation for the subsample of dairy and granivor farms

	Dairy subsample				Granivore subsample			
	Estimate	Std. Error	z value		Estimate	Std. Error	z value	
Intercept	-8.771	2.075	-4.227	***	-16.175	4.467	-3.621	***
Dummy region south	-0.258	0.323	-0.798		1.202	0.671	1.792	.
Dummy region west	0.141	0.338	0.417		-11.777	1455.398	-0.008	
Dummy konv farming	-0.033	0.311	-0.105		0.929	1.315	0.707	
Age	-0.010	0.014	-0.708		-0.035	0.030	-1.191	
Labour	0.329	0.272	1.209		0.713	0.579	1.233	
Utilised agricultural area (log)	1.144	0.320	3.576	***	2.574	0.745	3.454	**
Share of rented land	0.693	0.549	1.264		-1.246	1.377	-0.905	
Livestock density	0.802	0.360	2.230	*	0.689	0.396	1.738	.
Share of equity	1.433	0.814	1.760	.	3.810	1.876	2.031	*
Non-farm income (log)	0.250	0.066	3.766	***	0.246	0.130	1.894	.
Share of dairy cows	-0.067	0.855	-0.078					
Number of pigs					0.067	0.320	0.209	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The distribution of the bounded propensity scores is quite similar for treated and controls in the dairy subsample but is more distinctive in the granivore subsample (see Figure 6 and 7). This results in a more challenging *matching* procedure for the granivore subsample in order to fulfill the common-support assumption. The Greedy *matching* algorithm finds 104 pairs for the dairy and 27 pairs for the granivore, which increases the balance of the subsamples for each selected covariate (see Table 5 and 6). Balance of covariates is checked by the t-test, which shows no statistical significant difference on the conventional levels.

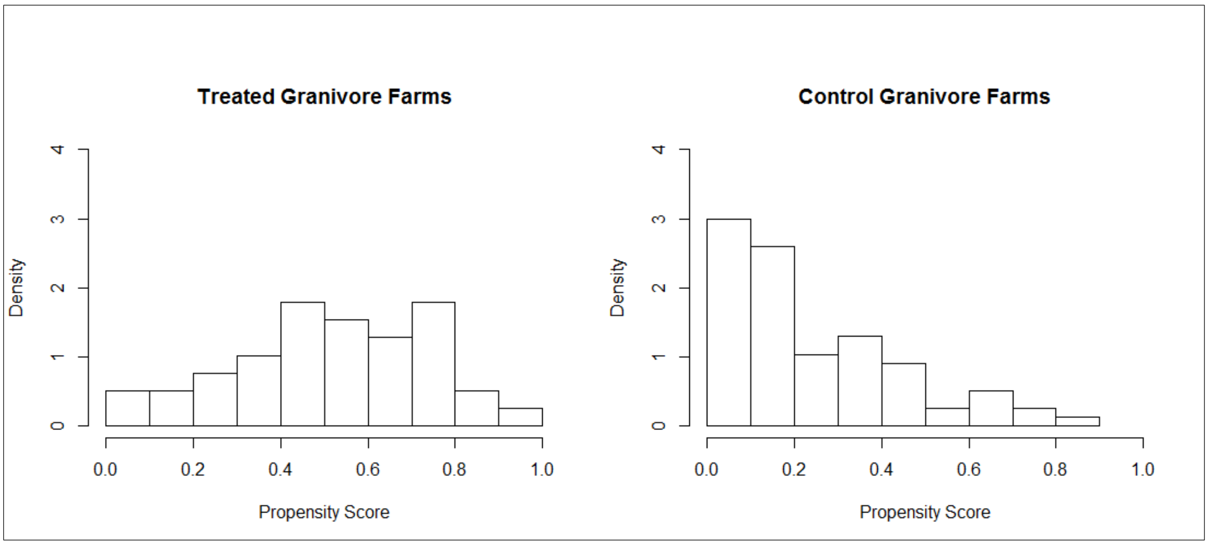

Figure 5: Distribution of propensity scores for treated (left) and controls (right) in the dairy subsample

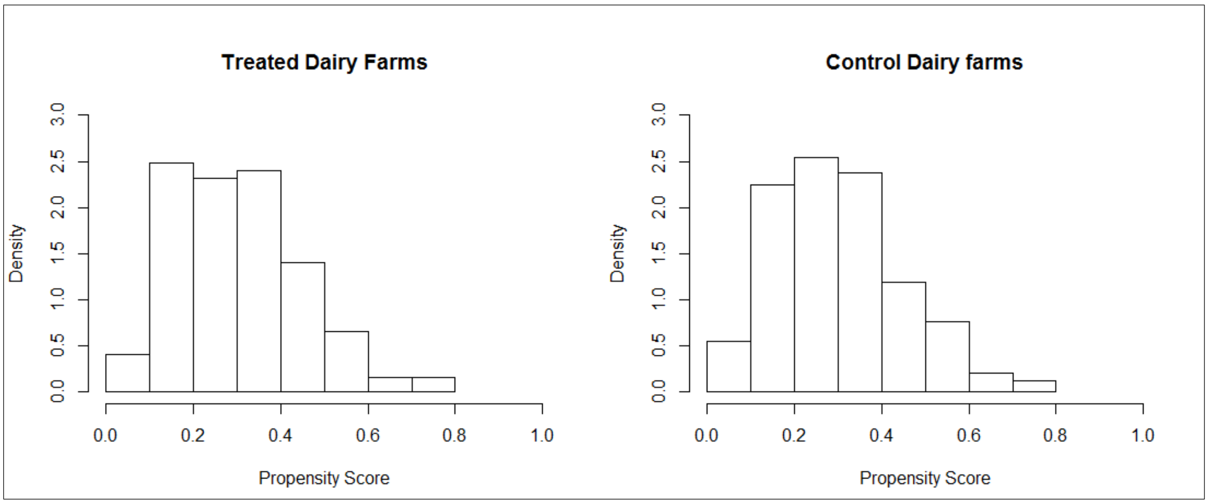

Figure 6: Distribution of propensity scores for treated (left) and controls (right) in the granivore subsample

Using the matched subsamples we can estimate the ATT in the farm income for dairy as well for granivore farms similar to the procedure when the whole sample is used. The farm income of treated dairy farms increases in average in the analysed period by about 1,200 Euros more than the control. The t-statistic is very low and therefore the result is not statistically significant. In contrast, the average development of farm income of treated granivore farms is 18,600 Euros higher and statistically significant at the 1% level or better (see Table 5). This reveals the heterogeneity and variability in the average results when the ATT is estimated with the whole sample.

Table 5: ATT in the farm income (in Euros) for the subsample of dairy and granivore farms

	Estimate	Std. Error	t-stat	
Dairy subsample	1232	2548	0.477	
Granivore subsample	18612	6864	2.711	***

t-test for equally of means: Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Stratification of the heterogeneous sample also leads to an increase in the robustness of the results. This is shown through the sensitivity analysis in Table 6, where the statistical significance and the magnitude of the treatment effect changes at a higher parameter than for the whole sample. For the dairy subsample the ATT is statistical insignificant for the assumption of randomisation but exceed the 5%-level when the parameter increases by 30%. In comparison the parameter has to increase by 50% to change the conclusion of the granivore sample.

Table 6: Rosenbaum bounds parameters for the results for the subsample of dairy and granivore farms

parameter (Γ)[1]	Dairy subsample				Granivore subsample			
	Wilcoxon P-value		HL treatment estimate		Wilcoxon P-value		HL treatment estimate	
	Lower bound[2]	Upper bound[3]	Lower bound[4]	Upper bound[5]	Lower bound[2]	Upper bound[3]	Lower bound[4]	Upper bound[5]
1.00	0.309	0.309	1374	1374	0.007	0.007	17261	17261
1.05	0.237	0.388	790	1892	0.005	0.009	16565	17733
1.10	0.178	0.469	229	2327	0.004	0.012	15856	18014
1.15	0.131	0.547	-321	2868	0.003	0.015	15207	18573
1.20	0.095	0.621	-790	3358	0.002	0.019	14072	19169
1.25	0.068	0.687	-1217	3859	0.001	0.024	13282	19406
1.30	**0.048**	**0.746**	**-1651**	**4310**	0.001	0.029	12766	19979
1.35	0.033	0.796	-2209	4793	0.001	0.035	12400	20817
1.40	0.023	0.839	-2696	5140	0.001	0.041	11948	21456
1.45	0.015	0.874	-3066	5544	0.000	0.048	11497	21786
1.50	0.010	0.903	-3456	6017	**0.000**	**0.055**	**11230**	**22160**
1.55	0.007	0.926	-3901	6348	0.000	0.063	10611	22626
1.60	0.005	0.944	-4293	6748	0.000	0.071	10073	24862
1.65	0.003	0.958	-4693	7036	0.000	0.080	9825	25003
1.70	0.002	0.969	-5025	7389	0.000	0.090	9466	25201

[1] Odds of differential assignment due to unobserved factors
[2] Lower bound significance level (on assumption of under-estimation of treatment effect).
[3] Upper bound significance level (on assumption of over-estimation of treatment effect).
[4] Lower bound point estimate (on assumption of under-estimation of treatment effect).
[5] Upper bound point estimate (on assumption of over-estimation of treatment effect).

5 Discussion and conclusions

The heterogeneity of farms and the problem of self-selection are challenging a evaluation of treatments in agriculture. This is particularly the case for rural development measures, which have voluntary participation and heterogeneous outcomes. But knowledge about the selection mechanisms for a certain treatment, in combination with econometric methods, can help to overcome these problems. Next to Instrumental Variable estimation the *Propensity Score Matching* method has become a popular tool in evaluation.

Basically, *matching* creates a new sample by identifying similar controls for each participating individual based on observed covariates. The selection of these covariates is a central issue and of high sensitivity. It is necessary to identify those variables which have the greatest influence on the decision to participate and on the outcome. PSM uses the probability of participation for each unit, estimated by a binary regression model, to reduce the *matching* dimension to one. In this paper we apply *PSM* in combination with the *Difference-in-Difference Estimator* to assess causal effects in the farm income of the farm-investment programme in Austria.

The results show a statistically significant and positive ATT (227 farms) in farm income per year by roughly 7,000 Euros. This might give a quite positive résumé of the farm-investment support programme in order to enhance the competitiveness of farms. But we cannot be sure if *matching* - including the difference-in-difference estimation - could reduce all the selection bias in the result. Particularly since this analysis deals with heterogeneous data the danger of hidden bias rises (Rosenbaum, 2005a). Therefore we apply sensitivity analysis to measure the effects of violation of the independence assumption. The sensitivity analysis for our model reveals that the causal conclusions are quite vague and can change with only a small amount of hidden bias. We split the sample in subsamples for the most favoured farm types, dairy and granivore farms in order to gain more homogenous samples. Then the *matching* procedure is done individually and the resulting effects differ dramatically. Whereas the effect on farm income for fostered dairy farms (104 farms) is not statistically significant, the effect for treated granivore farms (27 farms) is more than 18,600 Euros and statistically significant. Furthermore the results of the sensitivity analysis show that the models applied for the subsamples are slightly more robust to hidden bias than the model for the whole sample.

The results indicate, on the one hand, that the effect for a small and specific number of farms exceeds the average effect by a high amount. Therefore the splitting of the sample and the effects shows a more accurate picture of the treatment. On the other hand, the increased robustness through sample splitting can be explained by the fact that some group of units, e.g. different farm types, should not be paired with each other in order to derive causal effects, and that homogenous samples might also allow more suitable parametric models and coefficient estimates.

Therefore, especially in the context of agricultural treatment evaluation using observational studies, the need for homogenous samples is of server importance. Much attention needs to be focused on the *Matching* procedure, as the method has to obtain the independence assumption and the homogeneity in the sample. Even though the *Matching* procedure is basically a stratification of the sample, *Matching* on the estimated propensity score might often be misleading and encourage hidden biases. A much more effective method would therefore be the application of exact *Matching*, where treated and non-treated are exactly matched on their covariates and perfect stratification is done. This is especially the case when the inclusion of more covariates cannot describe opting for greater participation. Even though the exact *Matching* approach is limited to a small number of *matching* variables, next to individual adjustments it allows transparency for non-scientific stakeholders in the evaluation process. This is particular necessary as practical information is important for finding covariates. A large amount of work has to be put into pooling information and applying covariates which are plausible for the institutional environment, in which the study is carried out (Lechner, 2002). Transparency is also necessary, when the results are presented, as Rosenbaum (2010) argues*: "An observational study that is not transparent may be overwhelming or intimidating, but it is unlikely to be convincing."* (Rosenbaum, 2010, 147).

All in all, we find that *matching* can help to solve the problems of heterogeneity and self-selection in agricultural studies. *Matching*, at least, confronts the researcher with the process of causal exposure and also the limitations of available data. This is especially relevant in the context of agriculture, where management decisions are always dependent on the unique relationship between farm household and the farm enterprise, on-site and political conditions and also on personal attitudes of the farm manger. All these complex and unobservable factors make it difficult to explain selection mechanism in agriculture. However, *Matching* is definitely a useful tool to balance and pre-process the dataset and understand the direction of causal relationships. In special circumstances, causal claims can be drawn from the result.

Acknowledgements

We are grateful to the Austrian Federal Ministry of Agriculture, Forestry, Environment and Water Management (Division II/5) for funding the project on farm-investment support programme evaluation.

6 Literature

AUGURZKY, B., J. KLUVE (2004): Assessing the Performance of *Matching* Algorithms - When Selection into Treatment Is Strong. Forschungsinstitut zur Zukunft der Arbeit, Diskussionspapier No. 1301

BMLFUW (2010): Grüner Bericht 2010. Wien

CALIENDO, M., R. HUJER (2006): The Mircoeconometric Estimation of Treatment Effects – An Overview. In: Allgemeines Statistisches Archiv 90, 199-215

CALIENDO, M., S. KOPEINING (2008): Some Practical Guidance for the Implementation of Propensity Score Matching . In: Journal of Economic Surveys 22 (1), 31–72

DANTLER, M., S. KIRCHWEGER, M. EDER, J. KANTELHARDT (2010): Analyse der Investitionsförderung für landwirtschaftliche Betriebe in Österreich. Universität für Bodenkultur, Institut für Agrar- und Forstökonomie, Wien.

DIRKSMEYER, W., B. FORSTNER, A. MARGARINA, Y. ZIMMER (2006): Aktualisierung der Zwischenbewertung des Agrarinvestitionsförderprogramms (AFP) in Deutschland für den Förderzeitraum 2000 bis 2004. Länderübergreifender Bericht. Bundesanstalt für Landwirtschaft (FAL), Braunschweig

GENSLER, S., B. SKIERA, M. BÖHM (2005): Einsatzmöglichkeiten der *Matching* Methode zur Berücksichtigung von Selbstselektion. In: Journal für Betriebswirtschaft 55, 37-62

HECKMAN, J. J., H. ICHIMURA, J. A. SMITH, P. E. TODD (1998): Characterizing Selection Bias Using Experimental Data. In: Econometrica 66 (5), 1017–1098

HENNING, C.H.C.A., J. MICHALEK (2008): Ökonometrische Methoden der Politikevaluation: Meilensteine für eine sinnvolle Agrarpolitik der 2. Säule oder akademische Finderübung. In: Agrarwirtschaft 57(3/4), 232-243

Ho, D.E., K. Imai, G. King, E.A. Stuart (2007): Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. In: Political analysis 15, 199-236

Keele, L. (2010): An overview of rbounds: An R package for Rosenbaum bounds sensitivity analysis with matched data. Available at: http://www.personal.psu.edu/ljk20/rbounds%20vignette.pdf (14/3/2012)

Lechner, M. (2001): A Note on the Common Support Problem in Applied Evaluation Studies. University of St. Gallen, Department of Economics, Discussion Paper no. 2001-01

Lechner, M. (2002): Mikroökonomische Evaluation arbeitspolitischer Maßnahmen. University of St.Gallen, Department of Economics, Discussion Paper No. 2002-20

Morgan, S. L., Ch. Winship (2007): Counterfactuals and causal inference: methods and principles for social research. New York: Cambridge University Press.

Neyman, J. [1990 (1923)].On the Application of Probability Theory to Agricultural Experiments Essay on Principles. In: Sec. 9 Stat. Sci. 5(4):465–72. Transl. DM Dabrowska, TP Speed

Peel M.J., G.H. Makepeace (2009): Propensity Score Matching in Accounting Research and Rosenbaum Bounds Analysis for Confounding Variables. Available at SSRN: http://ssrn.com/abstract=1485734 or http://dx.doi.org/10.2139/ssrn.1485734 (14/3/2012)

Pufahl, A., CH.R. Weiss (2009): Evaluating the Effects of Farm Programmes: Results from *Propensity Score Matching* . In: European Review of Agricultural Economics 36 (1), 79–101

Rosenbaum, P.R. (2002): Observational Studies. New York: Springer.

Rosenbaum, P.R. (2005a): Heterogeneity and Causality. In: The American Statistician, 59 (2), 147-152

Rosenbaum, P.R. (2005b): Sensitivity Analysis in Observational Studies. In: Encyclopaedia of Statistics in Behavioural Science, 4, 1809-1814

Rosenbaum, P.R. (2010): Design of Observational Studies. New York: Springer

Rubin, D.B. (1977). Assignment to Treatment Group on the Basis of a Covariate. In: Journal of Educational Statistics, 2, 1–26

Rubin, D.B., P.R. Rosenbaum (1983): The Central Role of the Propensity Score in Observational Studies for Causal Effects. In: Biometrika 70(1), 41–55

Sekhon, J.S. (2011): Multivariate and *Propensity Score Matching* Software with Automated Balance Optimization: The *Matching* Package for R. In: Journal of Statistical Software 42 (7), 1-52

Smith, J.A., P.E. Todd (2005): Does *Matching* Overcome LaLonde's Critique of nonexperimental Estimators? In: Journal of Econometrics 125, 305-353

Appendix

Table 7: *Mean values of variables for participants and controls before and after Propensity-Score Matching for the dairy subsample*

	Potential participants	Potential controls		Selected participants	Selected controls
Number of farms	108	249		104	104
Dummy region south	0.185	0.225		0.192	0.231
Dummy region west	0.213	0.197		0.192	0.240
Dummy konv farming	0.787	0.767		0.788	0.769
Age	52.824	53.964		52.817	52.154
Labour	1.771	1.636	*	1.752	1.812
Utilised agricultural area (log)	3.369	3.149	***	3.341	3.320
Share of rented land	0.285	0.224	*	0.284	0.264
Livestock density	1.292	1.295		1.292	1.291
Share of equity	0.922	0.906		0.925	0.917
Non-farm income (log)	7.718	7.109	*	7.694	7.925
Livestock (log)	3.412	3.192	***	3.404	3.332
Dairy cows (log)	2.789	2.599	**	2.806	2.761
Pigs (log)	0.796	0.734		0.768	0.793

t-test for equally of means: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 8: *Mean values of variables for participants and controls before and after Propensity-Score Matching for the granivore subsample*

	Potential participants	Potential controls		Selected participants	Selected controls
Number of farms	39	77		27	27
Dummy region south	0.256	0.247		0.111	0.259
Dummy region west	0.000	0.013		0.000	0.000
Dummy konv farming	0.974	0.961		0.963	0.963
Age	51.821	54.208		53.630	53.333
Labour	1.730	1.503	*	1.687	1.576
Utilised agricultural area (log)	3.565	3.121	***	3.508	3.413
Share of rented land	0.300	0.241		0.262	0.260
Livestock density	1.687	1.560		1.506	1.728
Share of equity	0.904	0.864		0.932	0.940
Non-farm income (log)	7.490	7.218		7.392	7.207
Livestock (log)	3.969	3.390	***	3.815	3.812
Dairy cows (log)	0.053	0.073		0.077	0.139
Pigs (log)	5.944	5.404	*	5.947	5.915

t-test for equally of means: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1