Analyzing an Active Labor Market Program in Germany: A Regional Approach

An Attempt to Use Propensity Score Matching for the Estimation of Causal Effects on the Level of Counties and Independent Cities

Tim Stegmann
Institute for Work, Skills and Training (IAQ), University of Duisburg-Essen, Germany

ABSTRACT: The Institute for Work, Skills and Training was assigned to evaluate a labor market program aimed at the integration of long-term unemployed individuals aged 50 or older. The integration should have been achieved not only by training and coaching of individuals, but also by building regional networks between labor market stakeholders within a region. To appraise the success of the action undertaken on the regional level in the sense of causal effects, an observational study was used. The experimental- and control-groups were built using propensity score matching. The matching was done using not individual-level data, but data on the regional level because of missing individual data and the aims of the program. The mean growth of the number of employees subject to social insurance contributions over time was chosen as the outcome of the program. The findings are that observational studies are suitable to estimate the causal effects of active labor market programs on the regional (macro) level if individual data are missing or if the aims of the program cannot be observed on the individual level.

KEYWORDS: *active labor market policy; regional labor market program; propensity score matching; observational studies; macro level; potential outcome approach; counterfactuals*

In 2005 the Institute for Work, Skills and Training (IAQ) was assigned to evaluate a labor market program initiated by Germany's former Federal Ministry for Labor and Economic Affairs (now Federal Ministry for Labor and Social Affairs) aimed at the integration of long-term unemployed[1] individuals aged fifty and older. Under the title "Perspektive 50plus—Beschäftigungspakte in den Regionen" ("Perspectives for 50plus—Regional Pacts for Employment"), the program endeavored not only to overcome individual constraints regarding employability, but also to change the minds of personnel managers regarding age-specific human resource policies. These two main goals were pursued by offering training and coaching to individuals and by implementing regional networks, the "pacts for employment," among all relevant labor market stakeholders in the participating cities and counties. The program was not implemented areawide, but designed as an ideas competition for Germany's 442 local job centers that were free to participate in the Ministry's call and to do so alone or with other job centers as partners. As a result of this novel method of launching a federal labor market program, 62

[1] People who are without work for a minimum of twelve months are considered long-term unemployed.

"pacts for employment" with 92 involved job centers were chosen from more than 180 applications.

The downside of the Ministry's selective and flexible approach was that individual participation in the program was not recorded in the administrative database of the Federal Agency for Work. Individual data on participants and the individual treatment they received had to be channeled directly from the pacts to the evaluators—after the resolution of intricate problems of data protection. These data allowed a detailed description of the participants and their outcomes plus a comparative analysis of the effects of different treatments they received. However, these data did not provide a basis for an assessment of the causal effect of the program since data on nonparticipants were lacking. What the program design did offer, however, was a quasi-experimental design with participating and nonparticipating regions, though the selection was not random. Therefore, in our attempt to appraise the net effect of the actions undertaken on the regional level, we used an experimental control group comparison approach on the level of the job center districts.[2] In addition to sheer necessity because of the lack of individual data, we saw this approach justified by the ambition of the program to produce effects beyond identifiable participants. If regional networks evolve between the labor market stakeholders, e. g., job centers and companies, and if the program changes attitudes towards the elderly, then this should lead to more entries into employment not only by participants of the program, but by people aged fifty or older in general. This potential regional impact of the program was measured by two labor market indicators available for all districts, participating or not, namely the changes in employment levels of the target population and their entry rates into employment.

[2] For a similar approach in other scientific fields see Girma and Paton (2006); Millimet and List (2004); and Park, Wang, and Wu (2002).

The evaluation consisted not only of the analysis of the effects on the regional labor market presented in depth here, but also the analysis of the economic context of regions, the strategies of the local actors to integrate the older long-term unemployed, the actions that were undertaken to activate the participants, how and if regional networks evolved, and the efficiency of the program with regard to the integration costs per person. Data concerning the participants' experiences with the program were gathered in telephone interviews as well as in group discussions in the field. Also, local actors from job centers, educational institutions, and companies were interviewed to gather insights into the problems and demands regarding the integration of the elderly in the labor market.

Causal Effects, Observational Studies, and the Propensity Score

The fundamental evaluative problem in this study is that one cannot observe what would have happened to a unit that receives a treatment *if it wasn't* subject to the treatment (also called the counterfactual outcome).

A solution to address this problem lies in *experimental studies* where units are randomly assigned to an experimental-group that receives a special treatment and to a control-group that does not (or is treated with placebo). Due to the randomized assignment, characteristics that might influence the treatment tend to be equally distributed in the experimental and the control-groups. Thus, the (average) causal effect of the treatment can be observed directly: it is the difference between the outcomes in the experimental and the control-group. But for various reasons, truly experimental designs are rare in social policy. The exclusion of otherwise eligible individuals from a program from which they would hypothetically benefit, only in order to prove the hypothesis that it was this program and nothing else that helped others who did receive services, is widely perceived as

questionable on ethical grounds. In our case, we did have a discretionary assignment of treated districts but the selection was far from random. It was first a self-selection of about 180 out of 450 potential applicants, and then it was a selection of the 62 best concepts out the 180 applications. The problem that arises from nonrandom assignment is that one cannot directly compare treated units to untreated ones to appraise the (average) treatment effect, because characteristics that might influence the outcomes despite the treatment are not randomly distributed. The distribution of the characteristics is then likely to be biased. The solution to account for the bias lies in observing all characteristics that might influence the outcome and to construe an experimental group and a control group that are *balanced*, which means they are equal in the distribution of the confounding variables. The design is then called an *observational study* (Rosenbaum, 2001). The crux with this design is the identification of all relevant confounding variables that might influence the treatment itself or the selection into treatment (and thus have to be observed *before* the treatment begins). Unlike in experimental studies where one can assume that all confounding variables are balanced between the experimental and the control groups, there might be a *hidden bias* between the groups if there are unobserved covariates that influence the outcome despite the treatment. In other words, we have to ensure by reasoning and prior knowledge that the outcome of the treatment is conditionally independent from all confounding variables; this is often referred to as the *conditional independence assumption* (CIA) (Gangl & DiPrete, 2004) or as *strongly ignorable treatment assignment* (Rosenbaum & Rubin, 1983).

The control group can be drawn from all untreated units that could have been chosen for the treatment. The experimental group is drawn from the treated units and, depending on the chosen method to construe the groups, must not include all treated units (see below). Further all program regions are referred to as *treated*, the nonprogram regions are referred to as *untreated,* and I will refer to the constructed groups for the identification of the treatment effect as *experimental* and *control* groups.

The construction of the experimental and the control groups is done by statistical matching. The goal of the matching is to find an untreated region for every treated region that is most similar compared to all the other untreated regions regarding the confounding variables (further on called *covariates*). As Rosenbaum and Rubin (1983) show, it is not necessary to compare values of each covariate but to use the propensity score for the matching procedure if one is interested in the average treatment effect on the treated (ATT). The propensity score is the probability for the assignment to a treatment given the covariates and can easily be compared between treated and untreated units. The match for a treated region is then the untreated region with the minimum difference of the propensity score compared to all other untreated regions. As the propensity score is unknown in observational studies, it has to be estimated from the covariates, which can be done with a (multivariate) regression model with the assignment to treatment being the dependent variable. The estimated coefficients are then used to calculate the propensity score. It is common to use logit regression for the estimation of the coefficients used for propensity score calculation (Rosenbaum & Rubin, 1985).

The final matching can be done with different methods. In the example below, single-neighbor matching and caliper matching is used.[3] When performing single-neighbor-matching, the units with the smallest difference in the propensity score are matches no matter how big the difference is. With caliper matching, the procedure is much the same, but the matches have to be within a predefined maximum distance from each other.

[3] See Gangl and DiPrete (2004) for an overview of different matching methods.

Furthermore, the stable unit-treatment value assumption (SUTVA) has to be justified. This assumption states that the units, whether they are in the experimental or control group, must not affect each other regarding the outcome.[4] In our case, this assumption would be violated if people from the target group living in a nonexperimental region would find a new job in a program region, or vice versa.[5]

Outcome: Change in Stock of Employment and Mean Entry Rate

A common concept for the measurement of the employment level at two points in time is the age-specific *labor turnover rate* (Cramer & Koller, 1988; Erlinghagen & Knuth, 2002; Organisation for Economic Co-operation and Development, 1996). For the calculation of this rate, one needs to know the exits from and entries into jobs in a given period of time. Information on the amount of exits from jobs is currently not available for Germany. As a result of this limitation, we have chosen two employment indicators for the targeted age group in order to evaluate the success of the program on the regional level. The first indicator is the *change in stock of employment* subject to social insurance contribution at two points in time: at the end of 2005 (before the start of the program) and at the end of the first quarter of 2007 (the most recent available data before the end of the evaluation). The stock of employment in German register data is (currently) counted on the person level, which means that even if one person has two jobs, only one is counted. The second indicator is the *mean entry rate into employment* during the observed five quarters. The entries into jobs in German register data are (currently) counted casewise, no matter if more than one entry is from one person. Both indicators were calculated from official data of the Federal Agency for Work. Before assessing the impact of the program, we will compare the values of the indicators for program regions to *all* other regions.

At the end of the fourth quarter of 2005, there was a stock of 24,418,048 employees aged fifteen to sixty-five in Germany, excluding apprentices. 6,137,334 or 23.7 percent of these were aged fifty to sixty-five. By the end of the first quarter of 2007, the stock of employees in this age group had grown by 6% in Germany as a whole. During the observed time frame, the growth in stock was lower (5.5%) in the program regions than in regions not participating in the program (6.3%, see Table 1). This could not be expected, as the program regions were not selected because of difficult labor integration problems to begin with, but because their concepts were considered most promising in the ideas competition.

The changes in stock are not only influenced by labor demand, but also by retirement and demographics. A growth in stock can be effected by an increase in hiring, a slow-down of separations, or simply by larger cohorts "ageing into" the target group while being employed. However, the employment register does not provide exit data on the small territorial scale of the job centers. We therefore calculated the average entry rate as an indicator for the dynamics of the regional labor market. These are calculated as the geometric mean of entries in employment in Quarter Q_t divided by stock of employment in Quarter Q_{t-1} over the five observed quarters.[6]

As shown in Table 2, the mean entry rate for people aged fifty to sixty-five was slightly higher in program regions compared with all other regions (0.0714 versus 0.0707). It seems that in the regions with a "pact for

[4] Hujer, Blien, Caliendo, & Zeiss (2002) stress that regarding evaluation on individual level, SUTVA is always violated because of "the immense amounts spent on ALMP [active labor market policy] in Germany and the large scale of the programs, spill-over effects on nonparticipants are very likely" (p. 2).

[5] This aspect can be practically neglected in analysis for Germany. See footnote 8 for an explanation.

[6] The geometric mean was chosen because the arithmetic mean overestimates growth rates averaged over time.

employment," the labor market was or became more dynamic for the target group.

The pivotal question then is, can we assume a causal relationship between the existence of a "pact for employment" and the measured values of the indicators for the success of the program?

Table 1
Stock in Employment Subject to Social Security Contribution at the End of Fourth Quarter 2005 and the First Quarter 2007, Without Apprentices

	End of 4th quarter 2005		End of 1st quarter 2007		
Unit	**Absolute**	**Ratio**	**Absolute**	**Ratio**	**Alteration**
15 to 65					
Germany	24,418,048	100%	24,941,332	100%	+2.1%
Non-program regions	16,828,464	100%	17,177,804	100%	+2.1%
Pacts of employment	7,589,586	100%	7,763,528	100%	+2.3%
50 to 65					
Germany	5,787,510	23.%	6,137,334	24.6%	+6.0%
Non-program regions	3,990,077	23.7%	4,241,338	24.7%	+6.3%
Pacts of employment	1,797,433	23.7%	1,896,996	24.4%	+5.5%

Table 2
Mean Entry Rate, First Quarter 2006 to First Quarter 2007, Without Apprentice

Unit	**Mean entry rate**
Aged 15 to 65	
Germany	0.1189
Non-program regions	0.1187
Pacts of employment	0.1196
Aged 50 to 65	
Germany	0.0708
Non-program regions	0.0707
Pacts of employment	0.0714

Causal Analysis of the Impact of the "Pacts for Employment"

As we showed in the previous section, the change in stock of employment and the mean entry rate in the program regions developed differently from the regions without this special treatment, though in different directions. We will now check out if a causal relationship between the existence of a pact for employment and the dependent variables can be assumed. Since one cannot observe what would have happened to the regions with a treatment in the absence of it, we have to assess the effect of the pacts for employment by comparing this

experimental group to a control group. The difference between the mean outcome of the experimental group and the mean outcome of the control group is then the effect of the program on the experimental group, given that the conditional independence assumption holds (average treatment effect on the treated, ATT). The composition of the experimental and control groups is achieved with a statistical matching procedure that compares the propensity scores of treated and untreated regions.[7] The matching procedure can be done applying different techniques. Because of the small number of units available in our case, we used single-neighbor-matching with repetition. In this approach, for every treated unit, a suitable untreated unit is identified under the condition that an untreated unit can be the match for more than one treated unit. The outcome of that unit is then weighted with the number of occurrences as a match.

When performing single-neighbor-matching, there is no upper limit for the maximum distance of the propensity scores of two regions. The only criterion for the matches is that the difference of the propensity score should be at a minimum compared with all other possible matches. It might therefore be suitable to define an upper limit for the maximal distance between treated and untreated regions yielding more comparable experimental and control groups. In the following, the upper limit is called *caliper*. However, the drawback of using a caliper is that there may be no comparable untreated unit for a treated within the predefined caliper. If the experimental group determined under such a restriction is not identical to the group of the treated units, then the calculated average treatment effect is valid only for the experimental group, but not for the whole group of the treated. As no formal rules exist for the choice of the caliper, only reasoning about the object of investigation and trial of different calipers can help (Cochran & Rubin, 1973). With regard to arriving at conclusions for as many units under consideration as possible, the caliper should be generous in order to have only few units without matches; whereas with regard to the reliability of results, the caliper should be small. We will therefore present matching models with different calipers and without a caliper. The calipers are chosen as a multiple of the standard deviation of the propensity scores (Gangl & DiPrete, 2004).

The matching has to be done with covariates measured *before* the treatment begins. Most of the regional data are only available on an annual basis, with December 31 as the date of reference. The program started in September, 2005, but as we know from other components of our evaluation, there was little integration into the labor market during those first months of the program in late 2005. Therefore, we used data for December 31, 2005 for the estimation of the coefficients of the independent variables, and the indicators for the success of the program are observed from January 1, 2006, to March 31, 2007.

One problem that could arise when observing the regional impacts of a program is that of spillover effects, e. g. a person from program region A may find a job in nonprogram region Z, thus contributing to the outcome indicator of Z instead of A. However, the willingness of (long-term) unemployed to move for a new job is generally very low in Germany,[8] so that such spillover effects can be practically neglected so that the stable unit treatment value assumption is not violated.

Various covariates might influence the chosen outcome indicators and therefore have to be accounted for when matching nontreated to treated regions. Roughly speaking, these represent aspects of labor market supply,

[7] The calculations were carried out with STATA and PSMATCH2 (Leuven & Sianesi, 2003).

[8] Only 11 percent of surveyed unemployed in Germany would definitely move for a new job, including short-term unemployed (Brixy & Christensen, 2002), and there is no loss in unemployment compensation or basic income if one does not move for a new job.

demand, and environment. As an indicator of the economic power of a region, we chose the *regional gross domestic product per citizen* in 2005.[9] The past economic development was measured as the *mean change in the gross domestic product* from 2003 to 2005. Both measures have influence on the employment growth (Elhorst, 2003). The *population density* is a proxy variable for distances in the regions, which is a factor of supply and demand (Arntz & Wilke, 2008; van der Laan, 1992). From other parts of the evaluation, it is known that one of the barriers for integration of the target group into the labor market is limited mobility of the target group due to missing or lost driver's licenses, lack of transportation, or the inability to use public transport. In order to account for the different regional forms of organization and structure of decision making, we include a variable called *distinguishing between rural districts and cities*. In rural districts, usually more stakeholders are part of decision-making processes leading to more complex negotiations. The employment policy of the companies towards people aged fifty to sixty-five is reflected by the *regional age-specific employment rate*. Since age-specific employment rates may be influenced by regional variance in demographics, the *share of the fifty to sixty-five age group in the regional population* is included in the regression model. From analyzing individual data of program participants, we know that people from the target group find new jobs mostly in small and medium-sized businesses (SME). It is therefore expected that the regional proportion of SME's positively influences the probability of reintegration into work for the target group. The regional importance of SMEs is indicated by the proportion of jobs in SMEs in the region. What is also known from analyzing program participants' data is that service companies often employ people of the target group. Therefore, the *share of service companies in the region* is included in the regression model. The size of the service sector of a region is given as the proportion of jobs in service companies.

We estimate models for Germany as a whole as well as separate models for East and West Germany because of the remaining economic differences between the "old" and the "new" part of the country. In the estimation for the whole of Germany, we include a *variable indicating whether a region is located in east or in west part of Germany*.

Although we tried to build the model as completely as possible, there are some factors that we could not account for due to missing data. For example, we could control for the number of people living in a region aged fifty to sixty-five, but we could not control for the quality of the supply that might result in structural unemployment because of a mismatch of supply and demand in the region. A further shortcoming of the model is the missing information on money spent per person in the target group on active labor market policy, but we assume that more money is spent in the treated regions than in regions without treatment.[10]

Unlike in the descriptive comparisons presented in Tables 1 and 2, all job center districts that initially applied to take part in the program "Perspektive 50plus" but were not awarded are excluded from the possible control group because information is lacking if they implemented their concepts partially or completely anyhow. If this should have happened, and if such a region were included in the control group as a result of the matching procedure, this would result in a biased estimate. As a result, 145 job center districts are available to be included in the control group, as

[9] At the time of the analysis 2006 data was not available.

[10] On the macro level, it would be possible to come across this hidden bias by involving data on the amount of money spent by the job centers and the local branches of the Federal Agency for Work on active labor market policy in the estimations of the propensity score. However, data for the 178 local Agencies for Work cannot be broken down to the smaller 450 job center districts, so we could not include this factor in our estimations.

compared with 92 job center districts treated within one of the 62 regional pacts for employment. The pacts and potential candidates for the control group are shown in Figure 1.

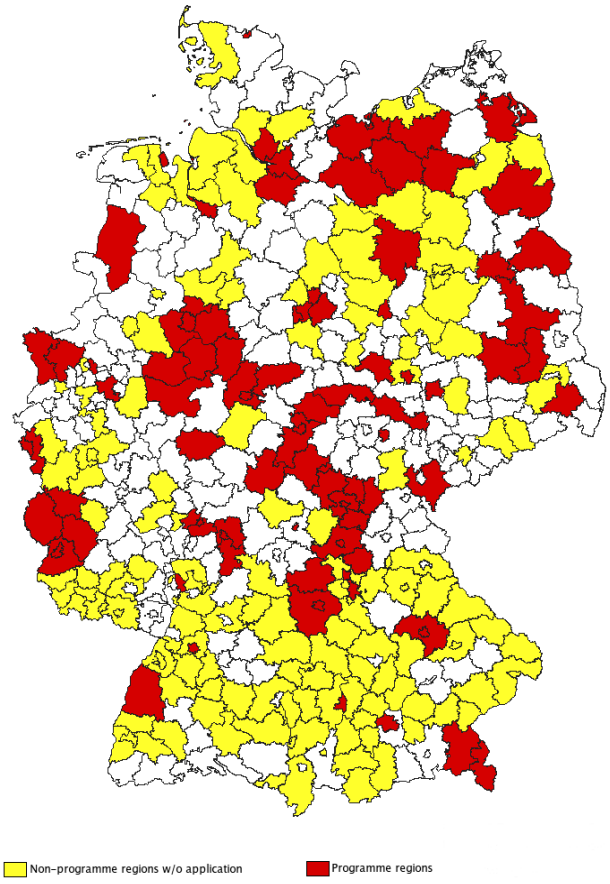

Figure 1. Program Regions and Nonprogram Regions With and Without Application

The matching procedure is repeated four times. In the first run, no caliper is applied and therefore there is a match in the control group for every unit in the experimental group. The other three runs are done with different calipers, defined respectively as 0.5, 0.25, and 0.05 times the standard deviation of the propensity score. By repeating the matching procedure with different calipers, the similarity of the two groups with regard to the control variables can be optimized against as full as possible inclusion of the treated group.

To assess the harmonization between experimental and matched control group, the standardized bias suggested by Rosenbaum and Rubin (1985) is calculated.[11] A bias of zero percent is optimal, meaning the groups are perfectly balanced. Before matching there is 83.5 percent bias in the mean propensity score between the groups of the treated and the untreated. After the matching *without a caliper*, there is a reduction in bias of 93.3 percent, resulting in a 5.6 percent bias, thus indicating considerable harmonization between the experimental and the control groups compared with unmatched groups of treated and untreated regions. Whereas the unmatched groups have a significantly different mean propensity score, the difference after the matching is no longer significant (see Table 3).

When the matching is done with a caliper of 0.5 times the standard deviation of the propensity score (0.0952), the bias is reduced by 97.1 percent, resulting in a 2.4 percent bias between the experimental and control groups.[12] In order to achieve this lower bias, four treated regions were excluded from the experimental group because there is no match for the regions within the range defined by the caliper. A further reduction of the bias to 0.8 percent is achieved by an even smaller caliper of 0.25 times the standard deviation of the propensity scores (0.0476), which increases the treated regions without a match to eight. An even further reduction of the caliper to 0.05 times the standard deviation of the estimated propensity

[11] The (standardized) percent bias is defined "as the difference of sample means in the treated and matched control subsamples as a percentage of the square root of the average of sample variances in both groups" for each covariate X (Caliendo, 2006, p. 78). The percent reduction of the bias "is $100(1-b_M/b_I)$, where b_I and b_M are the treated versus control differences in covariate means initially and after matching, respectively" (Rosenbaum & Rubin, 1985, p. 36).

[12] The mean of the average change of the gross domestic product of the last three years is significantly different between experimental and control groups at the 10 percent level.

scores (0.00952) can reduce the bias only a little more (0.7 percent), but fourteen treated regions are now without a match, resulting in only seventy-eight of the ninety-two program regions in the experimental group.

Breaking the repeated matching down between the two parts of the country, it can be observed that in West Germany, 62 job center districts were taking part in the program while 123 regions did not apply, and therefore are candidates for the control group. Without matching, the bias between treated and untreated groups is 73.6 percent, with a significant difference in the mean propensity scores (see Table 4). After constructing experimental and control groups without applying a caliper, the bias is reduced by 93.7 percent to 4.6 percent. The mean comparison test shows no significant differences between experimental and control groups. With a caliper of 0.5 times the standard deviation of the propensity scores (0.0837), the bias is reduced by 99.3 percent to 0.5 percent, and three program regions are without a match. A reduction by 99.7 percent is yielded with a caliper of 0.25 times the standard deviation of the propensity scores (0.0419), with four treated regions being without a match. A further reduction of the caliper to 0.05 times the standard deviation (0.00837) of the propensity scores results in a negative bias between the two groups.

In East Germany there are thirty program regions and only twenty-two untreated regions that did not apply to participate in the program. Because the untreated regions are allowed to be a match for more than one treated region, this is not a problem a priori. Again, there is a significant difference in the propensity scores of treated and untreated regions before the matching (see Table 5). Even though the bias reduction between experimental and control group is not as high as for the whole of Germany and for West Germany (77.6% to 15.5%), the differences in the propensity scores are not significant after the matching without a caliper. When a caliper of 0.5 times the standard deviation of the propensity scores is introduced (0.08614), five treated regions are without a match and the bias is reduced to 4.1 percent. With a caliper of 0.25 times the standard deviation of the propensity scores and six regions not included in the experimental group, the bias is reduced to 3.0 percent. With a caliper of 0.05 times the standard deviation of the propensity scores, the bias is reduced to 0.7 percent but sixteen program regions are without a match, which is more than half of the program regions in East Germany.

Table 3
Comparison of Means and Bias Reduction of Propensity Scores, Germany

		Mean(PS)				t-Test	
Sample	**| No match |**	**Experimental group**	**Control group**	**Bias in %**	**Bias reduction in %**	**t**	**P > |t|**
Without Matching		0.48299	0.33060	83.5		6.51	0.000
	Matched						
Without caliper	0	0.48299	0.47282	5.6	93.3	0.34	0.733
0.5*S.D.(PS)	4	0.46430	0.45987	2.4	97.1	0.15	0.878
0.25*S.D.(PS)	8	0.44712	0.44570	0.8	99.1	0.05	0.960
0.05*S.D.(PS)	14	0.43120	0.43085	0.2	99.8	0.01	0.990

Table 4
Comparison of Means and Bias Reduction of Propensity Scores, West Germany

		Mean(PS)				t-Test	
Sample	**| No match |**	**Experimental group**	**Control group**	**Bias in %**	**Bias reduction in %**	**t**	**P > |t|**
Without Matching		0.42046	0.29424	73.6		5.17	0.000
	Matched						
Without caliper	0	0.42046	0.41253	4.6	93.7	0.22	0.827
0.5*S.D.(PS)	3	0.39698	0.39610	0.5	99.3	0.03	0.979
0.25*S.D.(PS)	4	0.38981	0.39025	-0.3	99.7	-0.01	0.990
0.05*S.D.(PS)	14	0.34166	0.34240	-0.4	99.4	-0.02	0.981

Table 5
Comparison of Means and Bias Reduction of Propensity Scores, East Germany

		Mean(PS)				t-Test	
Sample	**| No match |**	**Experimental group**	**Control group**	**Bias in %**	**Bias reduction in %**	**t**	**P > |t|**
Without matching		0.62416	0.51145	70.2		2.44	0.018
	Matched						
Without caliper	0	0.62416	0.59895	15.7	77.6	0.59	0.557
0.5*S.D.(PS)	5	0.55871	0.55218	4.1	94.2	0.20	0.839
0.25*S.D.(PS)	6	0.54529	0.54049	3.0	95.7	0.17	0.863
0.05*S.D.(PS)	16	0.52611	0.52493	0.7	98.9	0.06	0.950

Results: Average Treatment Effect on Program Regions

After constructing different pairs of experimental and control groups by applying different calipers, we calculate the average treatment effect on the treated, which is now the difference between the groups in the two indicators described above: the change in stock of employment and the mean entry rate during the five observed quarters. For the whole of Germany, we calculate the effect for each of the four matching models introduced above. For the separate West and East models, we only use the first three matching models because in both cases, the respective fourth model was no improvement to the other three.

As we already highlighted in the description of the indicators, the change in stock of employment subject to social security contribution of people aged fifty to sixty-five in Germany as a whole was slightly higher in the regions without a pact for employment. This holds also true for the different versions of experimental and control groups. However, testing the coefficients shows that none of these differences is significant. In other words, our models suggest that there is no impact of the program on the regional level at all (see Table 6).[13] Likewise, with regard to the alternative indicator of outcome, the mean entry rates for persons aged fifty to sixty-five show no significant differences between the experimental and the control groups, no matter which version of the model is taken into account.

When estimating the average effect of the pacts of employment stratified for West and East Germany separately, the results are nearly identical to the results for Germany as a whole. The coefficients point in the same direction as in the description, but they are not significant (see Tables 7 and 8). The only exception is the effect on changes in the stock of employment in the target group in East Germany, which is negative when the experimental and control groups are constructed without applying a caliper, but positive when the different calipers are introduced (albeit all three coefficients are not statistically significant).

[13] As the propensity score for the regions is estimated, the estimation of standard errors is done with a bootstrapping with 1000 replications (Gangl & DiPrete, 2004). As some authors doubt in the application of bootstrapping to matching procedures (Abadie & Imbens, 2006) we repeated all calculations without bootstrapping. Even with bigger standard errors the results point in the same direction.

Table 6
Estimates of Average Treatment Effect on Treated Districts, Germany

	Coefficient	**Std. err.†**	**z**	**P > |z|**
Outcome 1: Change in stock				
Without caliper	-0.0252330	0.0052733	-0.48	0.632
0,5*S.D.(PS)	-0.0014749	0.0052579	-0.28	0.779
0,25*S.D.(PS)	-0.0002063	0.0051447	-0.04	0.968
0,05*S.D.(PS)	-0.0016186	0.0055008	-0.29	0.769
Outcome 2: Mean entry rate				
Without Caliper	0.0022213	0.0036029	0.62	0.538
0,5*S.D.(PS)	0.0015045	0.0033918	0.44	0.657
0,25*S.D.(PS)	0.0009337	0.0034367	0.27	0.786
0,05*S.D.(PS)	0.0006279	0.0036060	0.17	0.862

† Estimated standard error with bootstrapping (1,000 replications)

Table 7
Estimates of Average Treatment Effects on Treated Districts, West-Germany

	Coefficient	**Std. err. †**	**z**	**P > |z|**
Outcome 1: Change in stock				
Without caliper	-0.0042490	0.0057956	-0.73	0.464
0,5*S.D.(PS)	-0.0042872	0.0058559	-0.73	0.464
0,25*S.D.(PS)	-0.0046013	0.0062162	-0.74	0.459
Outcome 2: Mean entry rate				
Without caliper	0.0021885	0.0032283	0.68	0.498
0,5*S.D.(PS)	0.0025485	0.0032405	0.79	0.432
0,25*S.D.(PS)	0.0026369	0.0031763	0.83	0.406

† Estimated standard error with bootstrapping (1,000 replications)

Table 8
Estimates of Average Treatment Effect on Treated Districts, East Germany

	Coefficient	**Std. err.†**	**z**	**P > |z|**
Outcome 1: Change in stock				
Without caliper	-0.0031170	0.0114011	-0.27	0.785
0,5*S.D.(PS)	0.0049271	0.0108133	0.46	0.649
0,25*S.D.(PS)	0.0072418	0.0120796	0.60	0.549
Outcome 2: Mean entry rate				
without Caliper	0.0049526	0.0040507	1.22	0.221
0,5*S.D.(PS)	0.0048396	0.0039143	1.24	0.216
0,25*S.D.(PS)	0.0053982	0.0042334	1.28	0.202

† Estimated standard error with bootstrapping (1,000 replications)

Conclusions and Discussion

To sum up, no significant average effect of the program on regional labor markets for people aged fifty to sixty-five could be established with regard to the indicators chosen. This is not to say that the pacts for employment may not have had a positive impact on the individual level. If chances to be hired were improved for the participants, this was possibly offset at the aggregate regional level due to substitution within the targeted age group. Furthermore, without individual data and a time horizon for the evaluation that would allow following these individuals in the employment register, it cannot be refuted that integrations effected within the framework of the program may be more sustainable than integrations without the program. In any case, however, the net quantity of integrations effected in the program regions compared with nonprogram regions was not large enough to show up as a significant difference on the regional level, and the hope for effects beyond the individuals treated directly was probably vain.

The module of our evaluation presented here shares the fate of not coming up with a significant result with many other evaluations of labor market programs (Bundesministerium für Arbeit und Soziales, 2006; Kaltenborn, Knerr & Schiwarow, 2006). Nevertheless, it could be demonstrated that a quantitative causal analysis is technically feasible for Germany's counties and independent cities even in the absence of data on treated and nontreated individuals, as long as the program is not implemented countrywide. Such an approach might even be preferable if the impact of the program under consideration cannot be observed at all—or only insufficiently—on the individual level. However, problems might arise if the diversity between participating and nonparticipating regions is very high (as would be the case if only regions with certain negative labor market indicators are eligible) or if either the number of participating regions or the number of nonparticipating regions is very small.

On the methodological side, further investigation should be done about how different matching strategies (e. g., mahalanobis matching and kernel density matching [Gangl & DiPrete, 2004]) influence the construction of the experimental and control groups on the regional level and if the groups remain constant. To estimate the effects of the program not on the individual level but on the regional level was due to the sheer necessity of the constraints associated with the lack of individual data.

Evaluators always have to deal with constraints in data. Our approach shows that when individual data are not available, it is always worth it to think about possibilities for assessing impact on aggregated regional or—in more general terms—macro level.

The methods and calculations presented here give insight only on the *effect* of the program on regional labor markets, what we estimated is the *effect of causes*. Another question arises if one wants to know what the reasons for the effects were or why there was no significant impact of the program. Then one wants to know about the causes of effects. It is our strong belief that evaluation should always examine both—the effects and the causes. An observational study as present here is never sufficient to get the big picture of an objective. In other terms, it doesn't open the black box; it only sees the program as the input and the effects as the output. Of course, the client and the public want to know if the taxes spent for the program had positive effects with regard to the programs goals. But an evaluation should also *open* the black box and see what is inside. Otherwise, there will be no learning effects for future programs and much value of a program's evaluation is lost.

Acknowledgements

The author would like to thank Marcel Erlinghagen, Matthias Knuth, and four anonymous reviewers for helpful discussion and comments.

References

Abadie, A. & Imbens, G. W. (2006). On the failure of the bootstrap for matching estimators. *National Bureau of Economic Research Technical Working Paper 325*. Retrieved December 21, 2008 from http://www.nber.org/papers/t0325

Arntz, M., & Wilke, R. A. (2008). Unemployment duration in Germany: Individual and regional determinants of local job finding, migration and subsidized employment. *Regional Studies,* 1-19, Retrieved December 11, 2008, from http://www.informaworld.com/10.1080/00343400701654145

Brixy, U., & Christensen, B. (2002). Wie viel würden Arbeitslose für einen Arbeitsplatz in Kauf nehmen?, *IAB Kurzbericht*, *25*. Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung.

Bundesministerium für Arbeit und Soziales (2006). *Die Wirksamkeit moderner Dienstleistungen am Arbeitsmarkt. Bericht 2006 des Bundesministeriums für Arbeit und Soziales zur Wirkung der Umsetzung der Vorschläge der Kommission Moderne Dienstleistungen am Arbeitsmarkt (ohne Grundsicherung für Arbeitsuchende). Kurzfassung der Ergebnisse.* Retreived December 21, 2008, from http://www.bmas.de/coremedia/generator/3042/property=pdf/hartz__bericht__kurzfassung.pdf

Caliendo, M. (2006). *Microeconometric evaluation of labor market policies*. Berlin: Springer.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya A, 35*, 417-446.

Cramer, U. & Koller, M. (1988). Gewinne und Verluste von Arbeitsplätzen in Betrieben – der "Job-Turnover"-Ansatz, *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung 1988-3*, Nürnberg: Institut für Arbenitsmarkt- und Berufsforschung

Elhorst, J. P. (2003). The mystery of regional unemployment differentials: Theoretical and empirical explanations, *Journal of Economic Surveys*, *17*(5), 709-748.

Erlinghagen, M. & Knuth, M. (2002). *Auf der Suche nach dem "Turbo-Arbeitsmarkt," Graue Reihe 2002-03*. Gelsenkirchen: Institut Arbeit und Technik.

Gangl, M. & DiPrete, T. A. (2004). Kausalanalyse durch Matchingverfahren. In A. Dieckmann, *Methoden der Sozialforschung*,

Wiesbaden (pp. 396-420). VS Verlag für Sozialwissenschaften.

Girma, S., & Paton, D. (2006). Matching estimates of the impact of over-the-counter emergency birth control on teenage pregnancy, *Health Economics*, *15*(9), 1021–1032.

Hujer, R., Blien, U., Caliendo, M., & Zeiss, C. (2002). Macroeconometric evaluation of active labour market policies in Germany: A dynamic panel approach using regional data. *Discussion Paper No. 916*. Bonn: Institute for the Study of Labour.

Kaltenborn, B., Knerr, P., & Schiwarov, J. (2006). Hartz: Arbeitsmarktreformen auf dem Prüfstand. *Blickpunkt Arbeit und Wirtschaft, 3*. Retrieved December, 17, 2008, from http://www.wipol.de/download/blickpunkt200603.pdf

Leuven, E. & Sianesi, B. (2003), *PSMATCH2: STATA module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing, Version 1.2.3*. Retrieved December, 11, 2008, from http://ideas.repec.org/c/boc/bocode/s432001.html

Millimet, D. L., & List, J. A. (2004). The case of the missing pollution haven hypothesis, *Journal of Regulatory Economics*, *26*(3), 239–262.

Organisation for Economic Co-operation and Development. (1996). *Employment outlook 1996*. Paris: Author.

Park, A., Wang, S., & Wu, G. (2002). Regional poverty targeting in China. *Journal of Public Economics*, *86*, 123-153

Rosenbaum, P. R. (2001). *Observational studies*. New York: Springer.

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

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*, 33-38.

van der Laan, L. (1992). Structural determinants of spatial labour markets: A case study of The Netherlands. *Regional Studies, 26*(5), 485-498.