Institut für Wirtschaftsforschung Halle

Halle Institute for Economic Research

Is there a Superior Distance Function for Matching in Small Samples?

Eva Dettmann

Claudia Becker

Christian Schmeißer

February 2010 No. 03

IWH-Diskussionspapiere

IWH-Discussion Papers

Is there a Superior Distance Function for Matching in Small Samples?

Eva Dettmann
Claudia Becker
Christian Schmeißer

February 2010 No. 03

Authors: *Eva Dettmann*
Halle Institute for Economic Research (IWH)
Department of Structural Economics
Phone: +49 345 77 53 862
Email: eva.dettmann@iwh-halle.de

Claudia Becker
Martin Luther University Halle-Wittenberg (MLU)
School of Law, Economics and Business, Chair of Statistics
Email: claudia.becker@wiwi.uni-halle.de

Christian Schmeißer
Halle Institute for Economic Research (IWH)
Department of Formal Methods and Databases
Email: christian.schmeisser@iwh-halle.de

Editor:

Halle Institute for Economic Research (IWH)
Prof Dr Dr h. c. Ulrich Blum (President), Dr Hubert Gabrisch (Head of Research)
The IWH is member of the Leibniz Association.

Address: Kleine Märkerstraße 8, 06108 Halle (Saale)
Postal Address: P.O. Box 11 03 61, 06017 Halle (Saale)
Phone: +49 345 7753 60
Fax: +49 345 7753 820
Internet: http://www.iwh-halle.de

Is there a Superior Distance Function for Matching in Small Samples?

Abstract*

The study contributes to the development of 'standards' for the application of matching algorithms in empirical evaluation studies. The focus is on the first step of the matching procedure, the choice of an appropriate distance function. Supplementary to most former studies, the simulation is strongly based on empirical evaluation situations. This reality orientation induces the focus on small samples. Furthermore, variables with different scale levels must be considered explicitly in the matching process. The choice of the analysed distance functions is determined by the results of former theoretical studies and recommendations in the empirical literature. Thus, in the simulation, two balancing scores (the propensity score and the index score) and the Mahalanobis distance are considered. Additionally, aggregated statistical distance functions not yet used for empirical evaluation are included. The matching outcomes are compared using non-parametrical scale-specific tests for identical distributions of the characteristics in the treatment and the control groups. The simulation results show that, in small samples, aggregated statistical distance functions are the better choice for summarising similarities in differently scaled variables compared to the commonly used measures.

Keywords: distance functions, matching, microeconometric evaluation, propensity score, simulation

JEL classification: C14, C15, C52

* The authors thank Prof. Heinz P. Galler for his very helpful comments and suggestions. Furthermore, we thank Heiner Dettmann and Wilfried Ehrenfeld for many long and fruitful discussions on the simulation design as well as their excellent contributions to the implementation.

Welches Distanzmaß sollte für Matching in kleinen Stichproben verwendet werden?

Zusammenfassung

Die Studie leistet einen Beitrag zur Entwicklung von „Standards“ für den Einsatz von Matchingverfahren in empirischen Evaluationsstudien. Der Fokus liegt dabei auf der Entscheidung für ein geeignetes Distanzmaß. Die strenge Orientierung der durchgeführten Simulation an realen Entscheidungssituationen stellt eine Ergänzung zu den meisten bisher bekannten Studien dar. Sie erklärt zum einen die Fokussierung auf kleine Stichproben, zum anderen die explizite Berücksichtigung unterschiedlich skalierter Variablen, die im Matchingprozess berücksichtigt werden müssen. Die Analyse umfasst diejenigen Distanzmaße, die in der theoretischen Literatur als vorteilhaft angesehen bzw. häufig in empirischen Studien eingesetzt werden: die Mahalanobisdistanz und Balancing Scores. Darüber hinaus werden zwei aus der Statistik bekannte – in Evaluationsstudien bisher allerdings nicht verwendete – aggregierte Distanzmaße untersucht. Die erzielten Matchingergebnisse werden anhand nichtparametrischer skalenspezifischer Tests auf Übereinstimmung der Merkmalsverteilungen bewertet. Die Ergebnisse zeigen, dass aggregierte Distanzmaße in kleinen Stichproben besser in der Lage sind, Ähnlichkeiten in unterschiedlich skalierten Merkmalen zusammenzufassen als die bisher gebräuchlichen Maße.

Schlagwörter: Distanzmaße, Matching, Mikroökonometrische Evaluation, Propensity Score, Simulation

JEL-Klassifikation: C14, C15, C52

1 Introduction

In many empirical investigations, the outcome of different treatments has to be compared. Often, it is not possible to control for potentially influential covariates. Instead, one tries to find objects in the various treatment groups that are highly similar with respect to the presumed influential variables. Mechanisms to find such objects are called matching algorithms. Well-known in metrical applications (see, e.g., Babor and Del Boca 2003, Cooney et al. 1991), over the last years such matching procedures have also been become a widely used tool in the evaluation literature of political decisions, particularly European labour market policy interventions (see, e.g., Bergemann et al. 2004, Hujer and Thomsen 2006, Lechner et al. 2004, Sianesi 2004). Besides the popularity in empirical studies, there is a broad discussion of matching properties and the performance of this approach compared to other evaluation techniques. Some of the most influential studies at this field of research are Angrist and Hahn (2004), Cochran and Rubin (1973), Dehejia and Wahba (2002), Heckman, Ichimura, Smith and Todd (1998), Heckman and Hotz (1989), and Rosenbaum and Rubin (1983).
Furthermore, many proposals to improve matching can be found, e.g. Abadie and Imbens (2002), Fröhlich (2004b), Heckman et al. (1997, 1998), Lechner (2001, 2004), or Lechner and Miquel (2005).

Another strand of literature is more concentrated on practical guidance for the choice of an appropriate matching method for empirical research. This includes on the one hand reviews on recent developments (Angrist and Krueger 1999, Heckman et al. 1999, Imbens and Wooldridge 2008, Rubin 2006) and general recommendations how to select a suitable matching procedure (see, e.g., Caliendo and Kopeinig 2005, Heckman and Robb 1985, Rosenbaum 1987).
On the other hand, simulations and sensitivity analyses to discover the behaviour of selected matching approaches in different research situations can be found in the recent literature. One influential example is the discussion between Dehejia and Wahba (1999, 2002) and Smith and Todd (2005) on the ability of matching to overcome the the shortcomings of non-experimental estimators expressed by LaLonde (1986). Other important studies include the comparison of different matching and regression approaches by Abadie and Imbens (2002), and the study of Augurzky (2000b) on the properties of propensity score matching. In a simulation study, Gu and Rosenbaum (1993) compare the performance of matching based on different

distance functions and assignment algorithms when the sample structure changes. Also in a simulation Fröhlich (2004a) compares assignment algorithms based on the propensity score. Altogether, the studies provide very different results, depending on the design of the analysis.

The present study contributes to this discussion of advances and drawbacks of matching approaches in certain empirical research situations. Unlike most former studies, this analysis is strictly based on typical decision situations in practice. This practical orientation requires a focus on small samples, because data collection is expensive, and hence, the data base for many research questions is of limited size.

Matching procedures consist of two basic elements: a distance measure to decide upon the similarity of objects, and an assignment algorithm to find adequate partners for the members of the treatment group. In empirical studies, particular attention must be paid to the effect of differently scaled variables on the matching result, because the set of matching variables will usually not consist of covariates of just one scale level. This requirement influences particularly the first element, i.e., the choice of the distance measure. Therefore, the central question investigated here is the following: Which distance function is suitable for identifying and summarising similarity information in small samples when the matching variables do not share the same level of scaling?

To answer this question, a simulation study is conducted that compares commonly applied distance functions, i.e., the propensity score, the index score and the Mahalanobis distance. Additionally, statistical distance functions not yet used for matching are introduced. These distance functions combine several scale-specific distance measures and are thus expected to better capture similarity information of differently scaled variables. In order to assess the ability of the analysed functions, different samples types are generated based on the model of a real German micro data set. The performance of the distance functions is evaluated using non-parametric scale-specific tests of the variable means and frequency distributions.

The paper is organised as follows. In the next section, the theory of matching and the associated basic assumptions are explained. Subsequently, the analysed distance functions (section 3) and the simulation design (section 4) are presented. Section 5 discusses possible performance measures as well as the results of the simulation in terms of the employed measures. The most important features of the study and the results are summarised in section 6.

2 The Matching Approach

In order to evaluate a treatment, the outcome of the treated individuals is compared to a non-treatment outcome. Of course, it is not possible in practice to observe both outcomes for one and the same individual. This problem is usually illustrated with the Model of Potential Outcomes:[1]

$$Y_{it} = D_i \cdot Y_{it}^T + (1 - D_i)\, Y_{it}^C. \tag{1}$$

The observed outcome Y_{it} of individual i at time t is the outcome in case of treatment Y_{it}^T or non-treatment Y_{it}^C. The dummy D_i indicates, whether the individual i is treated ($D_i = 1$) or not ($D_i = 0$).

In this sense, the fundamental evaluation problem is a problem of missing data. Therefore, average treatment effects instead of individual effects are analysed. The individuals exposed to the analysed treatment are pooled in a so-called treatment group, all other observed individuals are members of the non-treatment group. Out of the non-treatment group, some individuals are selected to be compared to the treatment group. They are summarised in the so-called control group. The treatment effect is then estimated by comparing the average outcomes of the treatment and the control group.[2] The central problem for every method estimating average treatment effects is to eliminate the so-called selection bias. This term denotes differences in the outcomes of the treated and the comparison individuals which would show up even if neither group was treated. To remove all influences on the outcome, except for the treatment effect, all observable and unobservable heterogeneities between the compared groups must be controlled for.[3]

[1] This approach is formulated by Rubin (1973a, 1973b, 1974, 1977) as a framework for the analysis of causal effects. It can be found under different synonyms in the literature, e.g., 'Switching Regression Model' (Heckman et al. 1999), 'Rubin Causal Model' (Imbens and Wooldridge 2008), or Roy-Rubin-Model (Hujer and Caliendo 2000).

[2] The most popular effect is the average treatment effect for the treated. Other effects are discussed in the literature, see, e.g., Heckman et al. (1999), Imbens (2004), or Imbens and Wooldridge (2008).

[3] Since in most evaluation studies in a social science context, no designed experiments can be performed – especially some random assignment to treatment and control group, like e.g. in clinical studies, is usually impossible here – this problem cannot be solved by building treatment and control group beforehand. Instead, the control group has to be constructed from the non-treatment group after the treatment has taken place.

Matching is based on the idea of finding 'statistical twins' to solve the selection problem. For each treated individual, one or more partners from the non-treatment group are assigned. The assignment process is solely based on observable characteristics. Thus, potential heterogeneity in unobservable factors cannot be removed with matching.[4]

The central assumption of matching states that the potential outcomes are equal for individuals with identical observed characteristics, irrespective of their assignment to the treatment or the control group. In other words: Given the matching variables, the potential treatment outcome and control outcome are independent of the assignment to treatment. This assumption is commonly called Conditional Independence Assumption (CIA) (Lechner 2001), Ignorable Treatment Assignment (Rosenbaum and Rubin 1983) or Unconfoundedness (Imbens 2004):

$$Y_t^T, Y_t^C \perp D|\mathbf{X}. \tag{2}$$

The vector of matching variables is denoted by $\mathbf{X}$, the assignment to treatment by D; Y_t^T and Y_t^C denote the potential outcomes for the treated and the controls, respectively, and $\perp$ means independence. This assumption requires that all variables relevant for the outcome as well as for the treatment are considered in the matching process.

A necessary condition to be able to estimate the treatment effect is that for all observed values of the matching variables, there exists at least one individual in each of the groups carrying this value:

$$0 < \Pr\left(D = 1|\mathbf{X}\right) < 1. \tag{3}$$

This condition is referred to as Overlap (Crump et al. 2009) or the Common Support Condition (Sianesi 2004). Individuals outside the Common Support cannot be considered for the estimation of average treatment effects and must be removed from the analysed sample.

[4] A commonly recommended solution for this drawback is to combine matching with the Difference-in-Differences approach, see, e.g., Smith and Todd (2005). In some cases, if the data base is rich and contains detailed information, it is also possible to construct indicator variables for unobservable factors, see Reinowski et al. (2005).

If both assumptions (2) and (3) are fulfilled, the treatment assignment is said to be strongly ignorable (Rosenbaum and Rubin 1983). Additionally, a causal interpretation of the estimated effect requires the independence of the individual treatment effects of influences due to the treatment of other individuals, i.e. the compliance of the so called Stable Unit Treatment Value Assumption (Fröhlich 2004b).
The basic requirement of the CIA (2) is to find one or more control(s) for every treated individual with similar values of the matching variables. The first idea one could have is to match on every single variable. This exact matching raises the potential problem of not finding partners which correspond with respect to all variables, particularly if many matching variables are considered (Black and Smith 2004). Reducing the dimension of $\mathbf{X}$ is not a feasible option, because in this case the compliance with CIA is not plausible (Heckman et al. 1997). So, the similarity information must be summarised in an appropriate way.
Derived from the CIA, two basic requirements can be stated for the choice of a similarity measure. Firstly, such measures have to correctly capture the similarities and differences in the observed variables that occur between the analysed individuals. Secondly, when summarising the information, each variable must be equally weighted in the total similarity measure.

3 Analysed Distance Functions

In this study, the most common distance functions in the literature are compared to some statistical functions not yet used in the evaluation context. The introduction of these new distance functions results from drawbacks of the commonly used functions, especially with respect to their suitability for capturing similarities of observations in differently scaled variables in small samples.

3.1 Balancing Scores

The class of balancing scores consists of all functions BS with the following feature (Rosenbaum and Rubin 1983):

$$\mathbf{X} \perp D | BS(\mathbf{X}) \,. \tag{4}$$

If the assignment to treatment D is according to the balancing score BS, the distribution of the variables $\mathbf{X}$ in the treatment and the control group do not differ from

each other. That means the value of a variable is independent of the assignment to treatment. In their seminal paper, Rosenbaum and Rubin (1983) prove that a balancing score fulfills the CIA, if the score consists of variables that satisfy the assumption

$$Y_t^T, Y_t^C \perp D | BS(\mathbf{X}). \tag{5}$$

The terms Y_t^T and Y_t^C denote the outcome in the treatment and non-treatment case, respectively. In the empirical literature, two balancing scores are common – the propensity score and the index score.

Propensity Score The most widely used score is the propensity score (PS). This one-dimensional distance function is defined as the probability of being in the treatment group, or the probability of participation: $PS(\mathbf{X}) = \Pr(D = 1|\mathbf{X})$ (see, e.g., Caliendo and Kopeinig 2005). Usually, this participation probability is not observable and thus, has to be estimated. Probit models are commonly used for this estimation.[5] In the context of a Probit model, the observable decision to take a treatment is assumed to be determined by an unobservable latent variable, the participation tendency:

$$D_i = \begin{cases} 1 & \text{if } IN_i > 0 \\ 0 & \text{else.} \end{cases} \tag{6}$$

Here, the term D_i denotes the individual participation decision, and IN_i the unobservable individual participation tendency. This normally distributed latent variable is assumed to be influenced by observable relevant characteristics, the matching variables:

$$IN_i = \beta \mathbf{X}_i + \varepsilon_i \quad \text{with} \quad \varepsilon_i \sim N(0, \sigma^2), \tag{7}$$

where $\mathbf{X}_i$ denotes the observed individual values of the relevant characteristics, β their influence on the participation tendency. Within the scope of the model, the PS is estimated using the standard normal distribution function Φ:

$$\widehat{PS}(\mathbf{X}_i) = \Phi(\beta \mathbf{X}_i). \tag{8}$$

[5] Another widely used estimation approach is the Logit model. For further information on Logit and Probit models see, e.g., Greene (2008).

The propensity score is not only prevalent in empirical studies. It is object of extensive theoretical research, too. Heckman, Ichimura and Todd (1998) show that neither the asymptotic bias nor the asymptotic variance of the estimated treatment effect is larger compared to that of exact matching. Furthermore, according to Angrist and Hahn (2004) the asymptotic efficiency of PS-matching is higher than that of exact matching. Hahn (1998) develops efficiency bounds for the variance of PS-based matching and points out that knowing the true score lowers the variance of the estimated effect.

Gu and Rosenbaum (1993) assess that the propensity score is superior to other distance functions in a matching procedure when the number of covariates is large. Augurzky (2000a) analyses the influence of specification errors in the PS-estimation model on the performance, i.e., bias and mean squared error (MSE), of the estimated treatment effect. Based on his results, he recommends using solely highly significant covariates to estimate the PS. A further interesting result of Fröhlich (2004a) and Zhao (2008) is that the failure to fulfill the linearity assumption implied by the PS estimation model has only a small influence on the performance of the estimator. Dehejia and Wahba (1999) propose an iterative approach to specify the model for PS-estimation. This approach is implemented in matching tools for standard software, e.g. STATA, and is now widely used in empirical literature.[6]

An estimation of the participation tendency mentioned above is also often applied in the empirical literature and is then called index score.

Index Score This score is derived from the PS estimation based on the Probit model:

$$\widehat{IN}_i = \beta \mathbf{X}_i. \tag{9}$$

Compared to the propensity score, with the linear estimator (9) differences between treated individuals and the control group in the tails of the score distributions become more apparent, i.e., individuals with PS close to zero or one can be better distinguished by means of the index score (Lechner 1998).

When using the scores in practice two potential problems should be considered. Since the participation probability or participation tendency is not observable, the corresponding score has to be estimated, and the balancing property of the scores

[6] For details of this iterative process see Becker and Ichino (2002).

proofed by Rosenbaum and Rubin (1983) only holds asymptotically. Furthermore, estimating one of the scores implies including and weighting the variables according to their influence on the participation, not on the outcome (Zhao 2004). This may result in quite different outcomes for persons with identical score values, particularly in small samples (Fröhlich 2004b). Therefore, in small samples the risk of biased treatment effect estimation is high, because the control group may consist of individuals that are not 'statistical twins' of the treated ones with respect to the outcome.

3.2 Statistical Distance Functions

Mahalanobis Distance The most common alternative distance measure in empirical literature is the Mahalanobis distance. The distance function of Mahalanobis (1936) is specifically for metrically scaled variables. The distance between two individuals is determined as the weighted sum of the variable-specific differences:

$$MD_{ij} = \left[(\mathbf{x_i}) - (\mathbf{x_j})\right]' \mathbf{Cov}^{-1} \left[(\mathbf{x_i}) - (\mathbf{x_j})\right]. \tag{10}$$

The terms $\mathbf{x_i}$ and $\mathbf{x_j}$ denote the vectors of the considered covariates of the treated individual i and the non-treated j. The covariance matrix is denoted by **Cov** and is defined as follows: $\mathbf{Cov} = \frac{1}{I+J-1} \sum (\mathbf{x_i} - \bar{\mathbf{x}}) (\mathbf{x_i} - \bar{\mathbf{x}})'$, where I and J are the numbers of treated and non-treated individuals and $\bar{\mathbf{x}}$ denotes the vector of the covariate means.

Using the covariance matrix to weight the inter-individual distances has the advantage that variances as well as potential correlations between the covariates are accounted for. The other side of the coin is that distances of covariates with outliers are understated (Gu and Rosenbaum 1993). Nevertheless, in a simulation study, Zhao (2004) shows that in small samples the Mahalanobis distance is superior to the propensity score in terms of the MSE of the estimated treatment effect and its components, variance and bias. A serious drawback for the application in empirical studies is the focus on variables of only one scale level, metrical variables. Thus, similarities in not metrically scaled variables cannot be appropriately considered. Gu and Rosenbaum (1993) point out that this is particularly problematic for nominal variables with rarely occurring values.

Following from the drawbacks of the commonly used distance functions described above it seems to be beneficial to consider an alternative way of aggregation: to construct distance functions that consist of different scale specific distance measures. When pooling information of differently scaled variables, the following two requirements have to be taken into account. Firstly, normalisation of the single distances is necessary to make sure that every distance information contributes to the overall distance to the same degree. One of the most often used normalisation strategies is to divide every single distance by the observed variable specific maximum distance (Diday and Simon 1976).[7] Secondly, a transformation of the scale specific information is often necessary to be able to interpret the aggregated function. The reason is that distance functions are usually calculated for metrically-scaled variables, whereas similarity measures are used for nominally-scaled variables. A common transformation for normalized measures in the literature is: $d_{n,ij} = 1 - s_{n,ij}$, where the distance between the treated individual i and non-treated individual j in variable x_n is denoted by $d_{n,ij}$, and the similarity by $s_{n,ij}$.
Two distance functions from statistical literature that meet these conditions seem to be applicable in empirical evaluation studies and are therefore presented next.

Mahalanobis Matching Distance Kaufmann and Pape (1996) propose a combination of different distance functions as weighted average of the scale-specific measures. The number of variables of each scale is used to weight the respective distance functions.
To summarise metrically- and nominally-scaled variables the aforementioned Mahalanobis distance can be combined with the Generalised matching Coefficient (GMC). This coefficient can be defined as the share of covariates with equal values in all nominally scaled variables:

$$gMC_{ij} = \frac{1}{no} \sum_{n=1}^{no} Q(x_{ni}, x_{nj}). \tag{11}$$

The Generalised matching Coefficient is denoted by gMC_{ij}, the number of nominal covariates by *no*. $Q(x_{ni}, x_{nj})$ is an indicator for the equality of individuals i and j in variable x_n:

$$Q(x_{ni}, x_{nj}) = \begin{cases} 1 & \text{if } x_{ni} = x_{nj} \\ 0 & \text{else.} \end{cases} \tag{12}$$

[7] For alternative normalisation strategies see, e.g., Wilson and Martinez (1997).

As can be observed from the equations, using the GMC allows for different numbers of possible values in the covariates. The variables with coincident values are equally weighted irrespective of the number of possible values.

When combining the scale-specific distance functions, the differences in metrically-scaled variables are normalised with the maximum difference. The similarity information from the GMC is transformed into a distance measure. Both functions are weighted by the number of metrically- and nominally-scaled variables, respectively:

$$MDMC_{ij} = \frac{1}{N}\left[me \cdot MD_{ij} + no \cdot (1 - gMC_{ij})\right]. \tag{13}$$

The terms $MDMC_{ij}$, MD_{ij} and gMC_{ij} stand for the Mahalanobis matching distance and the scale-specific distances, N denotes the total number of variables: $N = me + no$, where me is the number of metrically-scaled covariates and no that of the nominally-scaled ones.

Gower Distance Another possibility to summarise similarity information of differently scaled variables can be found in Gower (1971). Here, the weighted average of variable specific similarity coefficients is presented:

$$SG_{ij} = \frac{\sum_{n=1}^{N} w_n s_{n,ij}}{\sum_{n=1}^{N} pc_{n,ij}}. \tag{14}$$

The aggregated similarity measure is denoted by SG_{ij}, the similarity between the treated individual i and the non-treated one j in the variable x_n is denoted by $s_{n,ij}$. w_n is a variable specific weight, and $\sum pc_{n,ij}$ is the number of variables without missing values for the examined individuals i and j.

When using this similarity measure in empirical evaluation, the denominator of equation (14) simplifies to $\sum_{n=1}^{N} pc_{n,ij} = N$, because the values of the relevant matching variables must be observable for every individual. Furthermore, the weighting factor for every variable-specific similarity coefficient is set to one: $w_n = 1$, and the similarity coefficients are transformed into distance functions. The resulting distance coefficient DG_{ij} is called Gower distance in the following:

$$DG_{ij} = \frac{1}{N}\sum_{n=1}^{N} d_{n,ij}. \tag{15}$$

The term $d_{n,ij}$ stands for the distance between individuals i and j in variable x_n, N is the total number of covariates. How the specific distance $d_{n,ij}$ is determined depends on the scale of variable x_n:

$$d_{n,ij} = \begin{cases} \frac{|x_{ni}-x_{nj}|}{diff_{max_n}} & \text{if } x_n \text{ metrical} \\ 1 - Q_{n,ij} & \text{if } x_n \text{ nominal.} \end{cases} \tag{16}$$

For metrically-scaled variables the absolute difference $|x_{ni} - x_{nj}|$ is used. The maximum observed difference $diff_{max_n}$ in the variable x_n is employed to normalise this difference. Distances in nominally scaled variables are determined using the transformed GMC. If every variable is separately considered, this coefficient corresponds to the equality indicator (12).

In a simulation study the suitability of the aforementioned distance functions for matching in empirical studies, i.e., their ability to meet the requirements stated in section 2, is analysed.

4 Simulation

The purpose of this study is to contribute to the development of guidelines for the application of matching in empirical research. In order to mimic real decision situations as closely as possible, the study is focused on small samples. Following the literature, a sample consisting of 100 individuals is regarded as small in this study.[8] Furthermore, when working with real data the researcher is faced with variables of different scale levels that have to be considered for matching. Therefore, unlike most earlier studies, differently scaled variables are explicitly considered in the simulation.[9]

[8] In earlier simulation studies, see, e.g., Fröhlich (2004a) and Zhao (2004), samples of this size are defined as small. The only instance of a stricter definition of small is found in the study of Gu and Rosenbaum (1993), where they set this restriction to 50 individuals.

[9] The simulation is conducted using MATLAB 6.5.

4.1 Hypotheses

The characteristics of the described distance functions allow the following hypotheses regarding their ability to summarise similarity information from differently scaled variables in small samples:

- Compared to balancing scores, statistical distance functions should better be able to capture similarities and differences in the outcome determinants between individuals in small samples.
- Aggregated statistical distance functions should be superior to the Mahalanobis distance when metrically- as well as non-metrically scaled variables have to be considered for matching.

4.2 Simulation Setup

The question asked for this analysis is: Which distance function is most suitable for summarising the similarity and distance information of variables with different scale levels? To answer this question, a simulation study with varying variable structures for a fixed number of treated and non-treated individuals is conducted.

Sample Design The simulation study is based on samples that are representing a group of persons of treated or non-treated individuals with different observed characteristics. The choice of variables is geared to frequently used information in empirical labour market studies. Following this basic concept, variables with different scale levels are generated. The distributions of these variables are generated following the information of a real data set, the German microcensus.
The microcensus is a survey based on a representative one per cent sample of all households in Germany. The microcensus provides information about every person in each household as well as the family and household context. It is the basis for policy decisions in the Federal Republic and the EU as well as for the current labor market research.[10] The microcensus draws a realistic picture of the scale level of the information that usually have to be included in evaluation studies and is therefore

[10] Statistical information gained from the Census is the basis of decisions made on, for example, the resource allocation of the European Social Fund and the European Regional Development Fund. Furthermore, the data is used in the annual reports of The German Council of Economic Experts.

used as a model for the simulation samples. This study uses the 2004 sub-sample of persons aged 25 to 55 years that corresponds to a commonly analysed age group in labour market studies.

The information in the data contains different scale levels, e. g., the metrical variables age, duration of education, and income level. Dichotomous nominally-scaled variables include information on sex, nationality and employment status. A third group of information contains ordinal and nominally-scaled variables with more than two possible values, such as education level, family background or kind of occupation. This group serves as an orientation for the construction of polytomous nominal variables. Ordinal variables are not generated explicitly, since they are usually converted into dichotomous nominal variables for the identification of similarities.

For the simulation samples 5 metrical, 5 dichotomous and 5 polytomous nominal variables are generated. Normal distribution is assumed for all metrically-scaled variables. The means and standard deviations of the following characteristics in the microcensus underlie the random selections: age, number of children, duration of education, seniority and net income. The generated dichotomous variables are based on the binomial distribution and microcensus information on sex, marital status, German citizenship, working in the public service sector and residence in Eastern Germany. For generating polytomous nominal variables, the frequency of occurrence of the different values of education level, occupational qualification, kind of occupation, size of enterprise and sector in which respondents are employed is reproduced.
All the variables are drawn from univariate variable distributions.[11]

In addition, linear combinations of these variables are defined: the 'true' treatment effect, the treatment and the non-treatment income.[12] The income specification geared to income estimates in the Annual Report 2004/2005 of The German Council of Economic Experts.[13] The difference between both incomes is determined by an

[11] Thus, the correlation structure of the generated variables is not taken into account. This is a minor restriction that could affect the behaviour and performance of the analysed functions. However, analysing the joint distributions of the differently scaled variables goes beyond the scope of this study.

[12] The generation of additional variables follows the requirements of policy evaluation, e.g., evaluating the income effect of training for the participants.

[13] The determinants of income in this report are: age, number of children, duration of education, seniority, sex, German citizenship, residence in Eastern Germany, working in the public service sector and quadratic terms of the mentioned factors. See German Council of Economic Experts

individual treatment effect, which is also generated as a linear combination of several variables.[14]

The result of constructing a sample is an 'artificial' group of individuals having the characteristics of the persons in the German microcensus.

Simulation Design Every generated random sample consists of 100 treated and 1000 non-treated individuals that are characterised by the variables described above.[15]

In the simulation, four different sample types are generated. These types differ from each other in the strength and the nature of the deviation of the variables in the non-treatment subsample from that of the subsample of the treated individuals. For every sample type, 1000 random samples are generated, i.e., every simulation step consists of 1000 runs.

The variable structure in the subsamples of treated individuals is fixed over the simulation. Changes in the variable structure of the whole sample are achieved by combining the treatment subsamples with different non-treatment subsamples (see table 1). The analysis starts with random samples with almost identical distributions of all variables in both subsamples. The next steps are to generate dissimilarities solely in the distribution of the metrically-scaled characteristics and deviations solely in the nominally-scaled variables, respectively. The last step of the simulation is characterised by samples with dissimilar distributions in variables of all scale levels.

Table 1: **Simulation Design**

Distribution of ...		*Metrical Variables*	
		similar	dissimilar
Nominal Variables	similar	sample 1	sample 2
	dissimilar	sample 3	sample 4

(2004), box 30. The parameters of this estimation will be used to generate the incomes in the samples.

[14] A detailed description of the generated variables and their use for matching as well as for defining the treatment effect and the incomes can be found in table 7 in the Appendix.

[15] The matching result is influenced by the ratio of treated to non-treated persons: the assignment of adequate partners tends to be less difficult with a higher ratio (see, e.g., Fröhlich 2004a). The ratio of 1:10 persons is expected to be large enough to eliminate a potential negative influence of this ratio on the matching result.

Following Gu and Rosenbaum (1993), the variation between different samples is conducted by a shift of the mean variable value. The deviation term incorporates two components, the variable-specific variation as well as the desired amount of deviation.[16] For variables with different variability, a fixed amount of deviation has different influences of the resulting distributions: The effect of deviation for variables with higher variability will be smaller than that for less varying variables. Thus, the variable-specific variation is included in the deviation term.
Variations in metrically-scaled variables are generated by variable-specific deviations from the means in the treatment sample. For nominally scaled variables, the deviation term determines different frequencies of occurrence of the various possible values between treated and non-treated individuals. For each sample it is randomly decided whether a positive or a negative deviation term is generated.

The desired amount of deviation in similar non-treatment samples is one per cent of the variable-specific variance, for dissimilar variable distributions the value is set to 25 per cent.

5 Analysis of the Distance Functions

In previous studies and in the empirical literature, three distance functions are commonly recommended: the propensity score, the index score (see 3.1) and the Mahalanobis distance (see 3.2). These functions are compared in the simulation. The propensity score (8) and the index score (9) are estimated using a Probit model.[17] Additionally, two aggregated statistical distance functions not yet used for empirical evaluation, the Mahalanobis matching distance and the Gower distance (see 3.2) are included in the analysis.

[16] For metrically-scaled variables the standard deviation is used to define the variation. As no statistical deviation measure exists for nominally-scaled variables, the 'variance' is approximated by the deviation from the variable 'median'.
The value of the variable-specific desired amount of deviation is based on the test statistic of a goodness of fit test, i.e., the test statistic of a χ^2-homogeneity test for nominally-scaled variables and that of the t-test in the case of metrically-scaled variables.
Detailed information on the definition of the variations between the treatment and non-treatment subsamples can be found in table 7 in the Appendix.

[17] This study uses a tool developed by LeSage (1999) for the estimation.

All analysed distance functions are specified using the same variables. In the specification, metrically as well as dichotomous and polytomous nominally-scaled variables are included. Interaction terms and quadratic terms are not considered.

As is the case in every empirical study, the common support condition must be verified as true for every matching variable. The assumption is met if at least one non-treated individual with the same variable value as a treated one can be found, and vice versa. For metrical variables, the condition is regarded as fulfilled if the deviation does not exceed a range of 10 per cent of the variable-specific mean. For the nominal variables, only exact matches are accepted.
Objects with covariate values that are outside the common support are excluded from the matching process, and hence are not considered in the analysis of the distance functions.

Based on the information of inter-individual distances, partners are assigned to the treated persons. Only one assignment algorithm, an optimal nearest neighbor matching, is used for all assignments, because the analysis concentrates on distance functions. Optimal assignments are not very common in the evaluation literature, but compared to other procedures they have the advantage that – in terms of the defined criterion – they find the best solution (Rosenbaum 1989).

The Hungarian Algorithm is introduced to the literature by Kuhn (1955). Based on the work of König (1916) and Egervary (1931), he proposes a solution for the classical assignment problem. The aim of this optimisation procedure is the minimal cost full assignment of persons to jobs. This idea can be applied to the assignment of similar individuals.[18] The cost will be replaced by the individual distances between treated and non-treated individuals, the sum of these distances is the optimisation criterion.[19]

The combination of the distance functions described above and the assignment process results in various control groups. These control groups can be evaluated using different performance measures.

[18] The basic idea of this procedure is often explained using matrix notation, see, e.g., Bazaraa et al. (1990). An alternative description of the algorithm based on graph theory can be found in Reinowski et al. (2005).

[19] A tool developed by Borlin (1999) is used to implement this algorithm. Since the number of treated individuals differs from that of the non-treated ones, the distance matrix is adjusted prior to the calculation: In order to produce a quadratic matrix additional rows whose elements are large numbers, compared to the real distances in the data (i.e. 9999) are inserted.

5.1 Measuring the Performance

Measuring the performance of matching means to control for two possible sources of biased estimation results. Firstly, estimation bias occurs due to the insufficient balance of the variable distributions in both samples. Secondly, the loss of observations may induce biased results. While the first source of bias – a violation of the Conditional Independence Assumption – can lead to biased results in any situation, the loss of observations only influences the estimation results if heterogeneous treatment effects occur.

The loss of observations is mainly determined by the choice of an assignment algorithm. In this study, an optimal assignment is applied, so no loss of observations due to an insufficient assignment process is expected.[20] Thus, only a potential violation of CIA is inspected.

Because this assumption is not directly testable, there is no 'standard test' to evaluate matching results. To get an impression of the quality of the results, in many empirical studies descriptive analyses are conducted (see, e.g., Black and Smith 2004, Buscha et al. 2008).

Furthermore, various performance measures can be found in the literature. When the 'true' treatment effect is known – as is the case in simulation studies or sensitivity analyses with experimental data – a statistical efficiency measure, the mean squared error (MSE), is commonly applied (e.g., Abadie and Imbens 2002, Dehejia and Wahba 1999, LaLonde 1986). The MSE is composed of the bias and the variance of the estimated treatment effect. It provides information on the deviation of the estimated from the 'true' effect as well as the deviation of single estimates from their mean (see, e.g., Dekking et al. 2005).

The matching process, but not the estimation result itself, is examined by means of the Percent Bias Reduction.[21] The Bias before matching of a variable is the difference between the mean values in the treatment and the non-treatment sample, whereas the Bias after matching denotes the mean of the value differences between each treated individual and its control(s) (Ming and Rosenbaum 2000). The bias comparison shows how much of the initial deviation is removed in the matching pro-

[20] Furthermore, this aspect is more important for the evaluation of different assignment algorithms and thus, is not considered in the analysis of distance functions.

[21] Cochran (1968) introduced this performance measure to the literature. Subsequently, it has been widely used in empirical studies (e.g., Augurzky 2000b, Cochran and Rubin 1973, Gu and Rosenbaum 1993).

cess. Furthermore, the bias after matching is a source of information on the balance of the variable values between each treated individual and its control subsample.

A similar information can be gained by the comparison of the Standardised Difference of the variables before and after matching. The difference before matching is the ratio of the difference of the mean values in the treatment and the non-treatment sample and the pooled standard deviation of both subsamples. After matching the difference is determined using the mean value of the control group.[22] The comparison of both Standardised Differences also shows how much deviation is removed by matching. Additionally, the difference after matching can be seen as a proxy for the ability of the matching process to balance the variable distributions in the groups altogether. This performance measure is also commonly used in empirical studies (see, e.g., Rosenbaum and Rubin 1985, Sianesi 2004).

Unfortunately, no criterion for an acceptable size of the after-matching-values of the performance measures can be found (Smith and Todd 2005). Therefore, statistical tests of differences in the variable distributions of the treatment and the control group are conducted in some studies (e.g., Augurzky and Kluve 2004, Lechner 1999, Sianesi 2004).

Similarly, to assess the performance of the distance functions, non-parametric tests for related samples are used in this study. Because of the different scale levels of the considered variables, it is not possible to use one test for all covariates. Thus, scale-specific tests are applied: for metrically-scaled variables the Wilcoxon signed rank test (Sheskin 2004) and for dichotomous nominal variables the McNemar test (Sheskin 2004). As for polytomous nominal variables no test for related samples is available, the χ^2-test of homogeneity (Sheskin 2004) is used.

5.2 Results

In the following, the main results for each of the four simulated sample types, as described in table 1, are presented.[23]

[22] All other terms of the equation stay the same.

[23] See tables 8 to 12 in the Appendix for detailed results for each of the analysed distance functions. Additionally, the Bias before and after matching, as well as the Percent Bias Reduction is given in the tables. The information is regarded as supplementary and not used for the analysis, because the definition of the Bias as the difference of variable means is not adequate to detect discrepancies in the frequency distribution of nominal variables.

The results in the presented tables represent the average results of the 1000 random samples that are generated for each sample type. In the first column, the number of the matching variables is given: variables 1-5 are metrically-scaled, 6-9 dichotomous, and 10-12 polytomous nominally-scaled.[24] These variables are used to specify each of the distance functions. In the second column, the average variable values in the treatment group can be found. Within each sample type, they are identical for all distance functions. Between the different sample types, the averages may gradually differ from each other, because individuals may be excluded from the analysis if the common support condition is not fulfilled. The same applies to the averages in the non-treatment sample which are given in the third column.

In contrast, quite different average variable values can appear in the control groups resulting from applying the different distance functions. Such differences represent the different weighting schemes for the covariates that are associated with the analysed distance functions.

In columns 4-11, the average variable value in the control group as well as the results of the tests can be found for every distance function. The test results are reported as an average rejection rate over all random samples where all tests are performed on a nominal 5% signifikant level. The result of each single test is denoted by zero if the null hypothesis of equal means and frequency distributions cannot be rejected and one in the case of rejection, respectively. Thus, the lower the reported value, the more balanced are the variable distributions in the treatment and the control group – and the better is the distance function.

The analysis yields almost identical results for the propensity score and the index score.[25] Therefore, only the results for the propensity score are presented the following tables (columns 4-5).

5.2.1 Variable Distributions Similar

Table 2 contains the results for the initial sample type, similar distributions of all variables in the treatment and the non-treatment sample. The presented results indicate that the statistical distance functions are able to discover similarities in

[24] An overview of the matching variables is given in table 7 in the Appendix.

[25] Both scores differ slightly in the average rate of rejection of the tests in a few variables, but the average variable values are identical for all considered variables. See tables 8 and 9 in the Appendix.

differently scaled variables in most cases, whereas the analysed scores often fail to find suitable partners for the treated individuals. The differences in the quality of the matching results regarding the performance measures are conspicuous.

The comparison of the average variable means in the control groups generated by means of the PS and the index score with the initial values shows that in many samples non-treated individuals are selected whose characteristics, on average, differ more from those of the treated ones than before matching. The assignment of 'improper' non-treated individuals leads to relatively high rejection rates of the hypothesis of equal means and frequency distributions in the scale-specific tests. In more than half the samples, on average over all scale levels, the null hypothesis is rejected.

On the other hand, the average variable means in the control groups resulting from the application of statistical distance functions are altogether more similar to those of the treated individuals than the values in the initial non-treatment group. Based on the Mahalanobis and the Gower distance, the identified control groups on average show smaller deviations from the treated group than the initial sample of non-treated individuals.

Looking at the average variable means gives a first impression on the performance of the distance functions. However, they provide no information on the most important and interesting question for the empirical research: What distance function is able to identify the closest partners for the treated individuals and thus to generate the best possible control groups? This question can be answered by means of the test results.[26]

The average rejection rate of the hypothesis of equal mean values or frequency distributions over all scale levels is approximately 20 per cent in the case of the Mahalanobis distance, 25 per cent for the Gower distance, and for the Mahalanobis matching distance only approximately 10 per cent. The aim of matching, to balance the variable distributions in the treatment and the control group, is best achieved when the matching process is based on the Mahalanobis matching distance. Overall, this distance function seems to be the best choice for the identification of suitable partners for the treated individuals when the members of a small sample group are very similar.

[26] All tests are performed at the significance level $\alpha = 5\%$.

Table 2: **Simulation Results for Samples with Similar Distribution in both, Metrical and Nominal Variables**

Initial Data			Propensity Score		Mahalanobis Distance		Mahalanobis Matching		Gower Distance	
X[a]	T[b]	NT[b]	C[b]	Test[c]	C[b]	Test[c]	C[b]	Test[c]	C[b]	Test[c]
1	40.02	39.97	37.24	0.67	39.98	0.13	40.10	0.15	40.01	0.09
2	0.99	1.06	0.92	0.37	1.01	0.15	1.02	0.20	0.99	0.08
3	12.01	11.99	11.20	0.55	12.01	0.16	12.06	0.22	12.01	0.10
4	11.83	12.46	10.22	0.49	11.97	0.16	12.33	0.21	11.83	0.07
5	1300.00	1346.46	1137.44	0.52	1310.95	0.14	1346.94	0.21	1298.80	0.11
6	0.50	0.50	0.55	0.49	0.51	0.05	0.50	0.00	0.50	0.14
7	0.65	0.65	0.68	0.47	0.66	0.03	0.66	0.00	0.64	0.15
8	0.90	0.90	0.91	0.26	0.91	0.01	0.93	0.13	0.90	0.11
9	0.20	0.20	0.33	0.50	0.20	0.00	0.18	0.06	0.20	0.13
10	2.86	2.86	2.72	0.75	2.89	0.44	2.88	0.00	2.86	0.64
11	2.35	2.35	2.26	0.68	2.34	0.51	2.33	0.03	2.36	0.60
12	3.14	3.15	2.98	0.75	3.22	0.45	3.20	0.00	3.14	0.61

Notes:
Average results for 1000 samples.
Deviation of the means and frequencies of the variable values, resp.: 1 per cent of variable specific variance.
[a] Included variables. Scale level: 1-5 metrical, 6-9 dichotomous, 10-12 polytomous;
[b] Average value of the variable in the treatment sample (T), the non-treatment sample (NT) and the control group (C);
[c] Average rejection rate of the null hypothesis of equal means and frequency distributions; Significance level $\alpha = 5\%$;
Scale specific tests (metrical variables: Wilcoxon signed rank test, dichotomous: McNemar test, polytomous: χ^2-test).

A more detailed look at the test results reveals interesting variations of the distance functions in their ability to balance the variable distributions for the different scale levels. The rejection rate for polytomous variables is rather high when the matching process is based on the Mahalanobis distance and the Gower distance, i.e., about 45 and 60 per cent, respectively. The Mahalanobis matching distance is, however, a very good basis for balancing the distributions of polytomous variables. The 'weakness' of this distance function seems to be the balance of metrically-scaled variables, where the rejection rate is almost one fifth, similar to that of the Mahalanobis distance. The smallest rejection rate at this scale level appears for the Gower distance with 10 per cent. Regarding the balance of the distribution of dichotomous variables, all statistical distance functions seem to be applicable, although the rejection rate for the Mahalanobis distance is smallest.
As an intermediate result it can be stated that in the presence of similar samples the propensity score and the index score do not provide a good basis for identifying similar partners for the treated individuals. The statistical distance functions are better suited for this task. Between these three functions, variations in their abilities to balance the variable distributions depending on the scale level are detected. Overall, the best partners for the treated individuals (compared to the results of other distance functions analysed) can be identified by means of the Mahalanobis matching distance.

5.2.2 Variable Distributions Dissimilar with respect to one Scale Level

The following tables summarise the simulation results for the two sample types with dissimilarities in metrical or nominal variables. In table 3, the results for samples with dissimilar distributions of metrically-scaled and similar distributions of nominally-scaled variables are presented. The non-treatment samples are characterised by a deviation amount of 25 per cent of the variable-specific variance from the mean values in the treatment sample for metrically-scaled variables. For the nominally-scaled variables, the deviation in the frequency of occurrence of the variable values is 1 per cent of the variable-specific deviation from the variable median. Table 4 contains the results for the opposite case of similarly distributed metrical variables and dissimilar dichotomuos and polytomous variables.

Table 3: **Simulation Results for Samples with Dissimilar Metrical and Similar Nominal Variables**

Initial Data			Propensity Score		Mahalanobis Distance		Mahalanobis Matching		Gower Distance	
X[a]	T[b]	NT[b]	C[b]	Test[c]	C[b]	Test[c]	C[b]	Test[c]	C[b]	Test[c]
1	40.02	40.12	38.24	0.48	40.07	0.56	40.20	0.73	40.04	0.28
2	0.99	1.06	0.97	0.29	1.01	0.50	1.02	0.54	0.99	0.29
3	12.01	11.99	11.55	0.48	12.01	0.73	12.06	0.73	11.99	0.40
4	11.81	12.44	11.02	0.46	11.97	0.55	12.32	0.62	11.81	0.31
5	1298.03	1347.71	1224.51	0.49	1310.10	0.57	1346.13	0.69	1299.55	0.34
6	0.50	0.50	0.53	0.22	0.50	0.05	0.50	0.00	0.50	0.13
7	0.65	0.65	0.67	0.20	0.66	0.05	0.66	0.01	0.64	0.13
8	0.90	0.90	0.91	0.13	0.91	0.01	0.93	0.12	0.90	0.11
9	0.20	0.21	0.26	0.25	0.20	0.01	0.18	0.06	0.21	0.13
10	2.85	2.86	2.77	0.75	2.90	0.44	2.88	0.00	2.86	0.64
11	2.35	2.35	2.30	0.68	2.34	0.52	2.33	0.03	2.35	0.59
12	3.14	3.14	3.04	0.76	3.22	0.45	3.19	0.00	3.14	0.66

Notes:
Average results for 1000 samples.
Deviation of the means and frequencies of the variable values: similar variables 1 per cent of variable-specific variance, dissimilar variables 25 per cent.
[a] Included variables. Scale level: 1-5 metrical, 6-9 dichotomous, 10-12 polytomous;
[b] Average value of the variable in the treatment sample (T), the non-treatment sample (NT) and the control group (C);
[c] Average rejection rate of the null hypothesis of equal means and frequency distributions; Significance level $\alpha = 5\%$; Scale-specific tests (metrical variables: Wilcoxon signed rank test, dichotomous: McNemar test, polytomous: χ^2-test).

Table 4: **Simulation Results for Samples with Similar Metrical and Dissimilar Nominal Variables**

Initial Data			Propensity Score		Mahalanobis Distance		Mahalanobis Matching		Gower Distance	
X[a]	T[b]	NT[b]	C[b]	Test[c]	C[b]	Test[c]	C[b]	Test[c]	C[b]	Test[c]
1	40.01	40.01	37.31	0.61	40.03	0.18	39.96	0.24	40.02	0.10
2	1.00	0.98	1.04	0.60	0.97	0.34	0.95	0.43	0.98	0.23
3	12.01	11.99	11.26	0.55	12.00	0.37	11.93	0.46	12.01	0.23
4	11.87	11.68	11.44	0.37	11.62	0.17	11.52	0.19	11.70	0.12
5	1301.22	1288.11	1239.52	0.37	1285.29	0.17	1288.56	0.18	1289.68	0.11
6	0.50	0.50	0.55	0.38	0.50	0.05	0.50	0.00	0.49	0.14
7	0.65	0.65	0.67	0.36	0.66	0.03	0.66	0.01	0.65	0.13
8	0.90	0.90	0.90	0.23	0.91	0.01	0.93	0.19	0.90	0.12
9	0.20	0.20	0.29	0.40	0.20	0.01	0.18	0.09	0.20	0.12
10	2.85	2.93	2.60	0.84	2.93	0.47	2.89	0.00	2.93	0.69
11	2.35	2.45	2.11	0.88	2.33	0.48	2.34	0.05	2.44	0.81
12	3.14	3.16	2.96	0.80	3.22	0.46	3.20	0.00	3.15	0.64

Notes:
Average results for 1000 samples.
Deviation of the means and frequencies of the variable values: similar variables 1 per cent of variable-specific variance, dissimilar variables 25 per cent.
[a] Included variables. Scale level: 1-5 metrical, 6-9 dichotomous, 10-12 polytomous;
[b] Average value of the variable in the treatment sample (T), the non-treatment sample (NT) and the control group (C);
[c] Average rejection rate of the null hypothesis of equal means and frequency distributions; Significance level $\alpha = 5\%$; Scale-specific tests (metrical variables: Wilcoxon signed rank test, dichotomous: McNemar test, polytomous: χ^2-test).

The generated deviations of the variable values between the treated individuals and the non-treated ones are not visible in the average variable means. This is explained by the fact that, when generating the samples, positive and negative deviations are randomly drawn. Both cases are equally likely, i.e., the means over all samples do not differ from each other.

The test results for the two sample types draw a very similar picture of the performance of the analysed distance functions to that of the initial sample type. For the PS and index score distinctly higher rejection rates are again observed than for the statistical distance functions. In the samples with dissimilar metrical variables, the null hypothesis of equal mean values and frequency distributions is rejected on average in every second sample, compared to rejection rates below 40 per cent for the Mahalanobis distance, about 35 per cent for the Gower distance and 30 per cent for the Mahalanobis matching distance. The distinction is even clearer in the samples with dissimilar nominally-scaled variables. In case of the PS and the index score, the overall rejection rate is more than 50 per cent, whereas the rates for the statistical distance functions are much smaller (20, 30 and 15 per cent, respectively).

The results for the different scale levels show a pattern similar to that in the initial sample type. Balancing the distribution of polytomous variables is problematic with the Mahalanobis distance and – particularly – with the Gower distance. The average rejection rates are 50 and 65 per cent. In contrast, the difference of the distributions of the polytomous variables is not statistically significant between treated individuals and the control groups generated with the Mahalanobis matching distance; the average rejection rate is nearly 0 per cent in both sample types.

The opposite is true in the case of the metrically-scaled variables. At this scale level, the average rejection rates are smallest for the Gower distance, whereas the Mahalanobis distance and the Mahalanobis matching distance perform relatively poorly – particularly in the samples 3 and 4 (see table 1).

For dichotomous variables the Gower distance leads to a slightly worse balancing result than the Mahalanobis distance and the Mahalanobis matching distance.

The comparison of the results of both sample types reveals important differences in the abilities of the analysed distance functions. For samples with dissimilar metrical characteristics, a substantial deterioration of the performance of the Mahalanobis distance and the Mahalanobis matching distance compared to the results in the initial design is observed. The average rejection rate of the hypothesis of balanced metrical variables is about 65 per cent and is thus even higher than the rejection

rates for the PS and the index score. The Gower distance is, however, also able to balance the means of the metrical variables in the case of relatively dissimilar data – the average rejection rate here is about one third. Changes compared to the initial results are, on the other hand, very limited for the samples with dissimilar nominal variables.
Overall – for both sample types and scales of all levels – the 'ranking' of the analysed distance functions does not change, but the superiority of the Mahalanobis matching distance compared to the Gower distance and the Mahalanobis distance is less clear, especially in the sample type of dissimilar metrical variables. The statement concerning the performance of the propensity score and index score however, remains the same as for the initial sample case.

5.2.3 Variable Distributions Dissimilar with respect to both Scale Levels

Table 5 contains the results for the sample type with dissimilar distributions in both types (metrical and nominal) of variables. The deviation of the means and frequencies of the variable values is set to 25 per cent of the variable-specific variance. The results for the samples with dissimilar variables lead to a very similar assessment of the abilities of the analysed distance functions as the previous analysis steps. Especially striking are the contrasting results for the Mahalanobis matching distance and the Gower distance. The poor performance of the Mahalanobis matching distance regarding metrical variables – with average rejection rates of approximately 65 per cent – is contrasted by a perfect balance of the polytomous nominal variables – the rejection rates are at 0 per cent. The opposite is true for the Gower distance. Here, the average rejection rates in the case of metrical variables are about 35 per cent, and for polytomous variables almost 75 per cent. The perfomance of the Mahalanobis distance lies between the two aggregated functions within the scale levels, but is slightly worse overall.
An interesting point is the improvement of the performance of propensity score and index score relative to the statistical distance functions. The null hypotheses of equal means or frequency distributions is rejected, on average, in 45 per cent of the samples – compared to about 40, 35, and 30 per cent in the case of the Mahalanobis distance, Gower distance, and Mahalanobis matching distance, respectively.

Table 5: **Simulation Results for Samples with Dissimilar Metrical and Nominal Variables**

Initial Data			Propensity Score		Mahalanobis Distance		Mahalanobis Matching		Gower Distance	
X[a]	T[b]	NT[b]	C[b]	Test[c]	C[b]	Test[c]	C[b]	Test[c]	C[b]	Test[c]
1	40.01	39.94	38.69	0.45	39.99	0.59	40.04	0.71	40.00	0.30
2	0.99	1.08	0.96	0.24	1.02	0.53	1.04	0.54	0.99	0.31
3	12.00	12.01	11.62	0.45	12.02	0.75	12.08	0.74	12.00	0.45
4	11.81	12.54	11.08	0.41	12.03	0.55	12.43	0.63	11.85	0.35
5	1297.48	1352.24	1229.10	0.46	1313.38	0.61	1356.49	0.70	1302.10	0.34
6	0.50	0.50	0.54	0.24	0.51	0.05	0.50	0.00	0.50	0.17
7	0.65	0.65	0.68	0.22	0.66	0.06	0.66	0.01	0.65	0.14
8	0.90	0.90	0.92	0.17	0.92	0.02	0.93	0.14	0.90	0.13
9	0.20	0.20	0.27	0.27	0.20	0.02	0.18	0.07	0.20	0.15
10	2.85	2.92	2.75	0.82	2.92	0.47	2.89	0.00	2.92	0.73
11	2.35	2.49	2.21	0.89	2.38	0.64	2.35	0.00	2.49	0.86
12	3.14	3.16	3.07	0.76	3.22	0.48	3.21	0.00	3.15	0.66

Notes:
Average results for 1000 samples.
Deviation of the means and frequencies of the variable values, resp.: 25 per cent of variable-specific variance.
[a] Included variables. Scale level: 1-5 metrical, 6-9 dichotomous, 10-12 polytomous;
[b] Average value of the variable in the treatment sample (T), the non-treatment sample (NT) and the control group (C);
[c] Average rejection rate of the null hypothesis of equal means and frequency distributions; Significance level $\alpha = 5\%$;
Scale-specific tests (metrical variables: Wilcoxon signed rank test, dichotomous: McNemar test, polytomous: χ^2-test).

5.2.4 Summary of the Results

Summarising the results for all the sample types, there is a recurrent pattern in the quality of distance functions in terms of their abilities to balance the variable distributions, i.e. covariate means and frequency distributions of the values. Compared to the statistical distance functions, the propensity score and the index score are less able to identify and summarise similarities and differences in differently scaled variables. On average over all four sample types, the null hypothesis is rejected in about one half of the samples versus one third in the case of the Mahalanobis distance and Gower distance, and about one fifth in case of the Mahalanobis matching distance.

The results suggest that the use of parametric models to estimate distance functions and the implicit weighting of the included variables according to their impact on the participation probability is problematic in small samples. This confirms the reservations expressed in Fröhlich (2004b) about the use of PS in small samples.

The results are also consistent with the findings of Zhao (2004), who states that – in comparison with the PS – the Mahalanobis distance is the better distance function for balancing the variable distributions in small samples.[27] The observed differences between the results of the propensity score and the index score are – unlike in Augurzky (2000a) – very small. The higher suitability of the index score for distinguishing between individuals on the distribution tails cannot be observed in this study's results.

Table 6 presents a 'ranking' of the analysed distance functions that summarises the simulation results.

The ranking of the statistical distance functions varies if the detailed results for the different scale levels are considered. The Gower distance is not suitable for balancing the distributions of polytomous variables, the results for the dichotomous variables are also poor in comparison to those of Mahalanobis matching distance and Mahalanobis distance. The 'strength' of this distance function lies in capturing metrical variables.

Almost the opposite is true for the Mahalanobis matching distance. While the bal-

[27] A similar statement for a small number of covariates (N=5) can be found in Gu and Rosenbaum (1993), who also assessed the opposite for a large number of variables (N=20). The number of variables used in the present simulation study (N=12) should be 'small enough', so that the results do not contradict those of Gu and Rosenbaum (1993).

Table 6: **Ranking of the Analysed Distance Functions**

	Scale-specific Rank			Overall
Distance Function	metrical V.	dichotomous V.	polytomouos V.	Rank
Propensity Score	4	3	4	3
Index Score	4	3	4	3
Mahalanobis Distance	2	1	2	2
Gower Distance	1	2	3	2
Mahalanobis Matching D.	3	1	1	1

ance of nominal variables (dichotomous and particularly polytomous) after matching is very good in every sample type, the 'weakness' of this distance function becomes clear when metrical variables are not similarly distributed.
The performance of the Mahalanobis distance for the different scale levels lies in most cases between those of the two aggregated distance functions.

The hypothesis regarding the superiority of statistical distance functions in comparison to the scores is confirmed by the simulation results. It can be observed in the samples that the Mahalanobis distance and the aggregated distance functions are better able to capture similarities and differences in the determinants of the outcome than propensity score and index score.

Following the second hypothesis, an aggregated distance function should perform better than the Mahalanobis distance when differently scaled variables are considered. This hypothesis involves two aspects. It expresses the expectation that, compared to a specific distance for metrical variables, the 'correct' capture of the similarities or differences in nominal variables will result in better balanced frequency distributions of these variables after matching. Moreover, with an aggregated distance function a 'smoothing' of the remaining deviations over all variables should be achieved, because every single difference has got the same weight in the total distance.
The expectation of better balanced variable distributions of nominal variables can be confirmed for polytomous variables in case of the Mahalanobis matching distance. For dichotomous variables, however, no considerable differences between the Mahalanobis distance and the aggregated distance function are observed. The results of the Gower distance for nominally-scaled variables are worse than those of the Mahalanobis distance.

The smoothing property for the remaining variable deviations can only be observed for the Gower distance.

6 Conclusion

In the study, the suitability of different distance functions for capturing similarity information is analysed. With the focus on small samples and differently scaled variables this study contributs to the discussion about distance functions for matching. The selection of the distance functions is determined by former studies as well as theoretical considerations. Thus, two commonly used balancing scores (the propensity score and the index score) and three statistical distance functions are compared in this simulation study. Of the three distance functions, the Mahalanobis distance has featured in former studies as well. The other two aggregated distance functions have not yet been used for empirical evaluation.

From theoretical considerations it is expected that the included distance functions should differ in their applicability, i.e., aggregated statistical distance functions should be superior in summarising similarity information in empirical studies.

In the simulation, four sample types are generated that differ from each other in the strength and the nature of the deviation of the variable distributions in the subsamples of non-treated and treated. The performance of the distance functions is evaluated using scale-specific non-parametrical tests of differences in the variable distributions after matching. The simulation results confirm the expectation of superiority of aggregated statistical distance functions.

The selection of an appropriate distance function for empirical studies depends on the predominant scale level of the matching variables. When samples have many metrically-scaled variables, the Gower distance should be used to capture the similarity information of the matching covariates. When most of the matching variables are nominally scaled – particularly in the case of polytomous variables – the Mahalanobis matching distance is the best available distance function. In the case of dichotomous variables, the performance of the analysed functions do not differ to a great extent, even though the results of the Gower distance are slightly worse than those of both other functions.

The divergent results of the two aggregated distance functions regarding nominally- and metrically-scaled variables necessitates further research. The question is whether it is possible to combine the 'strengths' of both distance functions and

simultaneously to eliminate their 'weaknesses'.
Considering the elements of both distance functions, it appears that the Generalized matching Coefficient – as is used in the Mahalanobis matching distance – is a better function for capturing similarities in nominally-scaled variables than using a single indicator of correspondence for each covariate. For the identification of differences in metrically-scaled variables the normalised absolute differences – as in the Gower distance – seems to be a better alternative to the Mahalanobis distance. A subsequent study should analyse whether the strengths of both aggregated distance functions can actually be joined in an aggregated distance function consisting of the Generalized matching Coefficient for nominal variables and the normalised absolute differences for metrical variables.

References

Abadie, A. and Imbens, G. W. (2002). Simple and Bias-Corrected Matching Estimators for Average Treatment Effects, NBER Technical Working Paper T286, National Bureau of Economic Research, Cambridge.

Angrist, J. D. and Hahn, J. (2004). When to control for covariates? Panel asymptotics for estimates of treatment effects, *The Review od Economics and Statistics* 86(1): 58–72.

Angrist, J. D. and Krueger, A. B. (1999). Empirical Strategies in Labor Economics, *in* O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, Vol. 3, Elsevier Science B. V, Amsterdam, chapter 23, pp. 1277–1366.

Augurzky, B. (2000a). *Evaluation Strategies in Labor Economics - An Application to Post-secondary Education*, PhD thesis, Ruprecht-Karls-Universität, Heidelberg.

Augurzky, B. (2000b). Matching the Extremes. A sensitivity Analysis Based on Real Data, Discussion Paper No. 310, Ruprecht-Karls-Universität, Heidelberg.

Augurzky, B. and Kluve, J. (2004). Assessing the Performance of Matching Algorithms When Selection Is Strong, IZA Discussion Paper No. 1301, Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn.

Babor, T. and Del Boca, F. K. (eds) (2003). *Treatment matching in alcoholism*, Cambridge University Press.

Bazaraa, M. S., Jarvis, J. J. and Sherali, H. D. (1990). *Linear programming and network flows*, 2 edn, Wiley, New York.

Becker, S. O. and Ichino, A. (2002). Estimation of average treatment effects based on propensity scores, *The Stata Journal* 2(4): 358–377.

Bergemann, A., Fitzenberger, B. and Speckesser, S. (2004). Evaluating the Dynamic Employment Effects of Training Programs in East-Germany Using Conditional Difference-in-Difference, ZEW Discussion Paper No. 04-41, Zentrum für Europäische Wirtschaftsforschung GmbH (ZEW), Mannheim.

Black, D. A. and Smith, J. A. (2004). How robust is the evidence on the effects of college quality? Evidence from matching, *Journal of Econometrics* 121(1-2): 99–124.

Borlin, N. (1999). MATLAB function hungarian, http://www.mathworks.com/ matlabcentral/fileexchange/loadFile.do?objectId=94&objectType=file. May 2004.

Buscha, F., Maurel, A., Page, L. and Speckesser, S. (2008). The Effect of High School Employment on Educational Attainment: A Conditional Difference-in-Difference-Approach, IZA Discussion Paper No. 3696, Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn.

Caliendo, M. and Kopeinig, S. (2005). Some Practical Guidance for the Implementation of Propensity Score Matching, DIW Discussion Paper No. 485, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin.

Cochran, W. G. (1968). The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies, *Biometrics* 24(2): 295–313.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review., *Sakhya: The Indian Juornal of Statistics, Ser. A* 35(4): 417–446.

Cooney, N. L., Kadden, R. M., Litt, M. D. and Getter, H. (1991). Matching alcoholics to coping skills or interactional therapies: Two-year follow-up results, *Journal of Consulting and Clinical Psychology* 59(4): 598–601.

Crump, R. K., Hotz, J. V., Imbens, G. W. and Mitnik, O. (2009). Dealing with limited overlap in estimation of average treatment effects, *Biometrika* 96(1): 187–199.

Dehejia, R. H. and Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs, *Journal of the American Statistical Association* 94(448): 1053–1062.

Dehejia, R. H. and Wahba, S. (2002). Propensity Score-Matching Methods for Nonexperimental Causal Studies, *The Review of Economics and Statistics* 84(1): 151–161.

Dekking, F. M., Kraaikamp, C., Lopuhäa, H. P. and Meester, L. E. (2005). *A modern introduction to probability and statistics. Understanding why and how.*, Springer, London.

Diday, E. and Simon, J. (1976). Clustering Analysis, *in* K. S. Fu (ed.), *Digital Pattern Recognition*, Springer-Verlag, Berlin, chapter 3, pp. 47–94.

Egervary, E. (1931). On combinatorial properties of matrices, *Matematikai Lapok* 38: 16–28.

Fröhlich, M. (2004a). Finite Sample Properties of Propensity-Score Matching and Weighting Estimator, *The Review of Economics and Statistics* 86(1): 77–90.

Fröhlich, M. (2004b). Programme Evaluation with Multiple Treatments, *Journal of Economic Surveys* 18(2): 181–224.

German Council of Economic Experts (2004). *Erfolge im Ausland - Herausforderungen im Inland. Jahresgutachten 2004/05*, H. Heenemann, Berlin.

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties, *Biometrics* 27(4): 857–871.

Greene, W. H. (2008). *Econometric Analysis*, 6th edn, Prentice Hall, Upper Saddle River.

Gu, X. S. and Rosenbaum, P. R. (1993). Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms, *Journal of Computational and Graphical Statistics* 2(4): 405–420.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects, *Econometrica* 66(2): 315–331.

Heckman, J. J. and Hotz, J. V. (1989). Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training, *Journal of the American Statistical Association* 84(408): 862–880.

Heckman, J. J., Ichimura, H., Smith, J. A. and Todd, P. E. (1998). Characterizing Selection Bias Using Experimental Data, *Econometrica* 66(5): 1017–1098.

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): 605–654.

Heckman, J. J., Ichimura, H. and Todd, P. E. (1998). Matching As An Econometric Evaluation Estimator, *Review of Economic Studies* 65(2): 261–294.

Heckman, J. J., LaLonde, R. J. and Smith, J. A. (1999). The Economics and Econometrics of Active Labor Market Programs, *in* O. Ashenfelter and D. E. Card (eds), *Handbook of Labor Economics*, Vol. III, Elsevier Science B.V., Amsterdam, pp. 1865–2097.

Heckman, J. J. and Robb, R. (1985). Alternative Methods for Evaluating the Impact of Interventions, *in* J. J. Heckman and B. Singer (eds), *Longitudinal Analysis of Labor Market Data*, Cambridge University Press, Cambridge, pp. 156–245.

Hujer, R. and Caliendo, M. (2000). Evaluation of Active Labour Market Policy: Methodological Concepts and Empirical Estimates, IZA Discussion Paper No. 236, Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn.

Hujer, R. and Thomsen, S. L. (2006). How Do Employment Effects of Job Creation Schemes Differ with Respect to the Foregoing Unemployment Duration?, ZEW Discussion Paper No. 06-47, Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim.

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review, *The Review of Economics and Statistics* 86(1): 4–29.

Imbens, G. W. and Wooldridge, J. M. (2008). Recent Developments in the Econometrics of Program Evaluation, *Discussion Paper 3640*, Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn.

Kaufmann, H. and Pape, H. (1996). Clusteranalyse, *in* L. Fahrmeir, A. Hamerle and G. Tutz (eds), *Multivariate statistische Verfahren*, 2 edn, Verlag de Gruyter, Berlin, pp. 437–536.

König, D. (1916). Über Graphen und ihre Anwendung auf Determinantentheorie und Mengenlehre, *Mathematische Annalen* 77(4): 453–465.

Kuhn, H. W. (1955). The hungarian method for solving the assignment problem, *Naval Research Logistics Quarterly* 2: 83–97.

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data, *American Economic Review* 76(4): 604–620.

Lechner, M. (1998). *Training the East German Labour Force. Microeconometric Evaluations of Continuous Vocational Training after Unification*, Physica-Verlag, Heidelberg.

Lechner, M. (1999). Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification, *Journal of Business & Economic Statistics* 17(1): 74–90.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption, *in* M. Lechner and F. Pfeiffer (eds), *Econometric Evaluation of Labour Market Policies*, number 13 in *ZEW Economic Studies*, Physica/Springer-Verlag, Heidelberg, pp. 43–58.

Lechner, M. (2004). Sequential Matching Estimation of Dynamic Causal Models, IZA Discussion Paper No. 1042, Forschungsinstitut zur Zukunft der Arbeit (IZA).

Lechner, M. and Miquel, R. (2005). Identification of Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions, Discussion Paper No. 2005-17, Department of Economics, University of St. Gallen.

Lechner, M., Miquel, R. and Wunsch, C. (2004). Long-Run Effects of Public Sector Sponsored Training in West Germany, Discussion Paper No. 2004-19, Department of Economics, University of St. Gallen.

LeSage, J. P. (1999). Applied Econometrics using MATLAB, http://www.spatial-econometrics.com/html/mbook.pdf. May 2004.

Mahalanobis, P. C. (1936). On the generalized distance in statistics, for the classification problem, *Proceedings of the National Institute of Science India* II(1): 49–55.

Ming, K. and Rosenbaum, P. R. (2000). Substantial Gains in Bias Reduction from Matching with a Variable Number of Controls, *Biometrics* 56(1): 118–124.

Reinowski, E., Schultz, B. and Wiemers, J. (2005). Evaluation of Further Training Programmes with an Optimal Matching Algorithm, *Swiss Journal of Economics and Statistics* 141(4): 585–616.

Rosenbaum, P. R. (1987). The Role of a Second Control Group in an Observational Study, *Statistical Science* 2(3): 292–316.

Rosenbaum, P. R. (1989). Optimal Matching for Observational Studies, *Journal of the American Statistical Association* 84(408): 1024–1032.

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.

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): 33–39.

Rubin, D. B. (1973a). Matching to Remove Bias in Observational Studies, *Biometrics* 29(1): 159–183.

Rubin, D. B. (1973b). The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies, *Biometrics* 29(1): 185–203.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, *Journal of Educational Psychology* 66: 688–701.

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

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*, Cambridge University Press, Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore.

Sheskin, D. J. (2004). *Handbook of Parametric and Nonparametric Statistical Procedures*, 3 edn, Chapman and Hall, Boca Raton.

Sianesi, B. (2004). An Evaluation of the Swedish system of Active Labor Market Programs, *The Review of Economics and Statistics* 86(1): 133–155.

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

Wilson, D. R. and Martinez, T. R. (1997). Improved Heterogeneous Distance Functions, *Journal of Artificial Intelligence Research* 6: 1–34.

Zhao, Z. (2004). Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence, *The Review of Economics and Statistics* 86(1): 91–107.

Zhao, Z. (2008). Sensititvity of propensity score methods to the specifications, *Economics Letters* 98: 309–319.

A Additional Information on the Simulation

Table 7: **Definition of the Samples**

	Generated Variables				Application for ...		
	Characteristics[a]	Mean	Std.dev.[b]	# of values[c]	Matching	Effect[d]	Outcome[d]
x_1	Age	40.00	8.00		x	x	x
x_2	# of children	0.70	1.00		x		x
x_3	Duration of education	12.00	2.50		x		x
x_4	Seniority	10.00	9.00		x		x
x_5	Net income	1200.00	800.00		x		
x_6	Sex	0.50		2	x		x
x_7	Marital status	0.64		2	x		x
x_8	German citizenship	0.91		2	x	x	x
x_9	Public service sector	0.17		2			x
x_{10}	Eastern Germany	0.15		2	x		x
x_{11}	Sector			3		x	
x_{12}	Education level			4	x		
x_{13}	Qualification			4	x	x	
x_{14}	Kind of occupation			4	x	x	
x_{15}	Size of enterprise			4			x
x_{16}	Quadrat. term	$\left(x_1^2\right)$					x
x_{17}	Quadrat. term	$\left(x_4^2\right)$					x
x_{18}	Interaction term	$(x_4 * x_{14})$				x	

Modification of Variables in Non-treatment Sample compared to the Treated Individuals

$x_1 - x_5$: $\bar{x}_{NT} = \bar{x}_T \pm (\breve{P}_t + k * \sigma)$

$x_6 - x_{10}$: $z_1^{NT} = \breve{P}_{\chi^2} * z_1^T \pm (1 + k)$

$z_0^{NT} = z^{NT} - z_1^{NT}$

x_{11} : $z_3^{NT} = \breve{P}_{\chi^2} * z_3^T \pm (k * s_M)$

$z_1^{NT} = \frac{z_3^{NT} * z_1^T}{z_3^T}$; $z_2^{NT} = z^{NT} - z_1^{NT} - z_3^{NT}$

$x_{12} - x_{15}$: $z_4^{NT} = \breve{P}_{\chi^2} * z_4^T \pm (k * s_M)$

$z_1^{NT} = \frac{z_4^{NT} * z_1^T}{z_4^T}$; $z_2^{NT} = z_2^T$; $z_3^{NT} = z^{NT} - z_1^{NT} - z_2^{NT} - z_4^{NT}$

Notes:

[a] Characteristics in the microcensus;
[b] Standard deviation of the characteristics in the microcensus;
[c] Number of possible variable values;
[d] Application of the variable to define the treatment effect and the outcome in the case of participation and of non-participation.

Scale level of the variables: x_1–x_5 metrical (normally distributed); x_6–x_{10} dichotomous; x_{11}–x_{15} polytomous.

$\bar{x}_T$, $\bar{x}_{NT}$ – Variable mean in the treatment and non-treatment samples, respectively;
z_v^T, z_v^{NT} – Frequency of occurrence of a variable value v in the treatment and non-treatment sample;
$\breve{P}_t$, $\breve{P}_{\chi^2}$ – Constant term, orientated on the test statistics of the goodness-of-fit tests, $\alpha = 5\%$ (normally distributed variables t-test, dichotomous and polytomous variables χ^2-test of homogeneity);
σ – Standard deviation of normal distribution;
s_M – Absolute deviation from the variable median;
k – Desired amount of deviation (similarity: 1 per cent of variable-specific variance, dissimilarity: 25 per cent).

Table 8: **Analysis of the Propensity Score**

	Average Values				Bias		
Variables[a]	T[b]	NT[b]	C[b]	Test Results[c]	before M.	after M.	Reduction
Similar Metrical and Nominal Variables							
1	40.02	39.97	37.24	0.67	1.18	2.86	-142.71
2	0.99	1.06	0.92	0.37	0.10	0.15	-50.69
3	12.01	11.99	11.20	0.55	0.52	0.93	-80.45
4	11.83	12.46	10.22	0.49	1.15	2.03	-77.09
5	1300.00	1346.46	1137.44	0.52	120.36	214.30	-78.05
6	0.50	0.50	0.55	0.49	0.10	0.16	-65.79
7	0.65	0.65	0.68	0.47	0.09	0.14	-54.75
8	0.90	0.90	0.91	0.26	0.06	0.06	-8.42
9	0.20	0.20	0.33	0.50	0.08	0.17	-111.90
10	2.86	2.86	2.72	0.75	0.04	0.18	-406.03
11	2.35	2.35	2.26	0.68	0.03	0.14	-399.67
12	3.14	3.15	2.98	0.75	0.04	0.21	-436.52
Dissimilar Metrical and Similar Nominal Variables							
1	40.02	40.12	38.24	0.48	2.42	2.27	6.11
2	0.99	1.06	0.97	0.29	0.20	0.14	32.77
3	12.01	11.99	11.55	0.48	1.08	0.74	30.85
4	11.81	12.44	11.02	0.46	2.30	1.84	19.94
5	1298.03	1347.71	1224.51	0.49	240.39	195.02	18.87
6	0.50	0.50	0.53	0.22	0.10	0.10	-0.64
7	0.65	0.65	0.67	0.20	0.09	0.09	3.32
8	0.90	0.90	0.91	0.13	0.06	0.05	18.12
9	0.20	0.21	0.26	0.25	0.08	0.09	-19.13
10	2.85	2.86	2.77	0.75	0.04	0.18	-389.44
11	2.35	2.35	2.30	0.68	0.03	0.14	-382.39
12	3.14	3.14	3.04	0.76	0.04	0.21	-423.65
Similar Metrical and Dissimilar Nominal Variables							
1	40.01	40.01	37.31	0.61	1.28	2.86	-123.24
2	1.00	0.98	1.04	0.60	0.18	0.22	-20.96
3	12.01	11.99	11.26	0.55	0.73	0.88	-20.96
4	11.87	11.68	11.44	0.37	1.13	1.63	-44.20
5	1301.22	1288.11	1239.52	0.37	115.23	170.71	-48.15
6	0.50	0.50	0.55	0.38	0.10	0.13	-34.37
7	0.65	0.65	0.67	0.36	0.09	0.12	-28.08
8	0.90	0.90	0.90	0.23	0.06	0.06	-3.80
9	0.20	0.20	0.29	0.40	0.08	0.13	-65.20
10	2.85	2.93	2.60	0.84	0.08	0.28	-253.02
11	2.35	2.45	2.11	0.88	0.10	0.25	-141.83
12	3.14	3.16	2.96	0.80	0.04	0.23	-470.86
Dissimilar Metrical and Nominal Variables							
1	40.01	39.94	38.69	0.45	2.42	2.06	15.15
2	0.99	1.08	0.96	0.24	0.21	0.13	38.80
3	12.00	12.01	11.62	0.45	1.08	0.70	34.76
4	11.81	12.54	11.08	0.41	2.31	1.68	27.23
5	1297.48	1352.24	1229.10	0.46	244.86	183.59	25.02
6	0.50	0.50	0.54	0.24	0.10	0.10	4.08
7	0.65	0.65	0.68	0.22	0.10	0.09	9.71
8	0.90	0.90	0.92	0.17	0.06	0.05	15.51
9	0.20	0.20	0.27	0.27	0.08	0.10	-21.14
10	2.85	2.92	2.75	0.82	0.07	0.18	-172.85
11	2.35	2.49	2.21	0.89	0.14	0.17	-21.26
12	3.14	3.16	3.07	0.76	0.04	0.19	-343.52

Notes:
Average results for 1000 samples.
Deviation of the means and frequencies of variable values:
similar variables 1 per cent of variable-specific variance, dissimilar variables 25 per cent.

[a] Scale level of the included variables: 1-5 metrical, 6-9 dichotomous, 10-12 polytomous;
[b] Average value of the variable in the subsamples (T – treated, NT – non-treated, C – control group);
[c] Average rejection rate of the null hypothesis of equal means and frequency distributions;
Scale-specific tests: Wilcoxon signed rank test for metrical variables, McNemar test for dichotomous variables, χ^2-test for polytomous variables, resp.; Significance level $\alpha = 5\%$.

Table 9: **Analysis of the Index Score**

	Average Values				Bias		
Variables[a]	T[b]	NT[b]	C[b]	Test Results[c]	before M.	after M.	Reduction
Similar Metrical and Nominal Variables							
1	40.02	39.97	37.24	0.66	1.18	2.86	-142.71
2	0.99	1.06	0.92	0.37	0.10	0.15	-50.69
3	12.01	11.99	11.20	0.54	0.52	0.93	-80.45
4	11.83	12.46	10.22	0.50	1.15	2.03	-77.09
5	1300.00	1346.46	1137.44	0.51	120.36	214.30	-78.05
6	0.50	0.50	0.55	0.50	0.10	0.16	-65.79
7	0.65	0.65	0.68	0.47	0.09	0.14	-54.75
8	0.90	0.90	0.91	0.26	0.06	0.06	-8.42
9	0.20	0.20	0.33	0.49	0.08	0.17	-111.90
10	2.86	2.86	2.72	0.75	0.04	0.18	-406.03
11	2.35	2.35	2.26	0.68	0.03	0.14	-399.67
12	3.14	3.15	2.98	0.75	0.04	0.21	-436.52
Dissimilar Metrical and Similar Nominal Variables							
1	40.02	40.12	38.24	0.47	2.42	2.27	6.11
2	0.99	1.06	0.97	0.29	0.20	0.14	32.77
3	12.01	11.99	11.55	0.47	1.08	0.74	30.85
4	11.81	12.44	11.02	0.45	2.30	1.84	19.94
5	1298.03	1347.71	1224.51	0.49	240.39	195.02	18.87
6	0.50	0.50	0.53	0.22	0.10	0.10	-0.64
7	0.65	0.65	0.67	0.20	0.09	0.09	3.32
8	0.90	0.90	0.91	0.13	0.06	0.05	18.12
9	0.20	0.21	0.26	0.26	0.08	0.09	-19.13
10	2.85	2.86	2.77	0.75	0.04	0.18	-389.44
11	2.35	2.35	2.30	0.68	0.03	0.14	-382.39
12	3.14	3.14	3.04	0.76	0.04	0.21	-423.65
Similar Metrical and Dissimilar Nominal Variables							
1	40.01	40.01	37.31	0.61	1.28	2.86	-123.24
2	1.00	0.98	1.04	0.61	0.18	0.22	-20.96
3	12.01	11.99	11.26	0.54	0.73	0.88	-20.96
4	11.87	11.68	11.44	0.36	1.13	1.63	-44.20
5	1301.22	1288.11	1239.52	0.37	115.23	170.71	-48.15
6	0.50	0.50	0.55	0.38	0.10	0.13	-34.37
7	0.65	0.65	0.67	0.35	0.09	0.12	-28.08
8	0.90	0.90	0.90	0.23	0.06	0.06	-3.80
9	0.20	0.20	0.29	0.40	0.08	0.13	-65.20
10	2.85	2.93	2.60	0.84	0.08	0.28	-253.02
11	2.35	2.45	2.11	0.88	0.10	0.25	-141.83
12	3.14	3.16	2.96	0.80	0.04	0.23	-470.86
Dissimilar Metrical and Nominal Variables							
1	40.01	39.94	38.69	0.44	2.42	2.06	15.15
2	0.99	1.08	0.96	0.23	0.21	0.13	38.80
3	12.00	12.01	11.62	0.44	1.08	0.70	34.76
4	11.81	12.54	11.08	0.40	2.31	1.68	27.23
5	1297.48	1352.24	1229.10	0.45	244.86	183.59	25.02
6	0.50	0.50	0.54	0.23	0.10	0.10	4.08
7	0.65	0.65	0.68	0.22	0.10	0.09	9.71
8	0.90	0.90	0.92	0.18	0.06	0.05	15.51
9	0.20	0.20	0.27	0.27	0.08	0.10	-21.14
10	2.85	2.92	2.75	0.82	0.07	0.18	-172.85
11	2.35	2.49	2.21	0.89	0.14	0.17	-21.26
12	3.14	3.16	3.07	0.76	0.04	0.19	-343.52

Notes: see table 8.

Table 10: **Analysis of the Mahalanobis Distance**

	Average Values				Bias		
Variables[a]	T[b]	NT[b]	C[b]	Test Results[c]	before M.	after M.	Reduction
Similar Metrical and Nominal Variables							
1	40.02	39.97	39.98	0.13	1.18	0.51	57.07
2	0.99	1.06	1.01	0.15	0.10	0.05	51.56
3	12.01	11.99	12.01	0.16	0.52	0.21	58.70
4	11.83	12.46	11.97	0.16	1.15	0.51	55.32
5	1300.00	1346.46	1310.95	0.14	120.36	50.79	57.80
6	0.50	0.50	0.51	0.05	0.10	0.02	76.18
7	0.65	0.65	0.66	0.03	0.09	0.02	76.49
8	0.90	0.90	0.91	0.01	0.06	0.02	69.91
9	0.20	0.20	0.20	0.00	0.08	0.02	80.37
10	2.86	2.86	2.89	0.44	0.04	0.06	-75.65
11	2.35	2.35	2.34	0.51	0.03	0.04	-50.29
12	3.14	3.15	3.22	0.45	0.04	0.10	-144.10
Dissimilar Metrical and Similar Nominal Variables							
1	40.02	40.12	40.07	0.56	2.42	1.01	58.14
2	0.99	1.06	1.01	0.50	0.20	0.09	54.93
3	12.01	11.99	12.01	0.73	1.08	0.45	58.45
4	11.81	12.44	11.97	0.55	2.30	0.99	56.69
5	1298.03	1347.71	1310.10	0.57	240.39	101.80	57.65
6	0.50	0.50	0.50	0.05	0.10	0.02	74.74
7	0.65	0.65	0.66	0.05	0.09	0.02	74.69
8	0.90	0.90	0.91	0.01	0.06	0.02	66.87
9	0.20	0.21	0.20	0.01	0.08	0.02	77.76
10	2.85	2.86	2.90	0.44	0.04	0.07	-82.78
11	2.35	2.35	2.34	0.52	0.03	0.05	-65.58
12	3.14	3.14	3.22	0.45	0.04	0.09	-131.87
Similar Metrical and Dissimilar Nominal Variables							
1	40.01	40.01	40.03	0.18	1.28	0.59	54.25
2	1.00	0.98	0.97	0.34	0.18	0.07	59.22
3	12.01	11.99	12.00	0.37	0.73	0.30	59.07
4	11.87	11.68	11.62	0.17	1.13	0.56	50.22
5	1301.22	1288.11	1285.29	0.17	115.23	55.01	52.26
6	0.50	0.50	0.50	0.05	0.10	0.02	75.68
7	0.65	0.65	0.66	0.03	0.09	0.02	75.96
8	0.90	0.90	0.91	0.01	0.06	0.02	67.90
9	0.20	0.20	0.20	0.01	0.08	0.02	77.92
10	2.85	2.93	2.93	0.47	0.08	0.08	-2.10
11	2.35	2.45	2.33	0.48	0.10	0.06	41.17
12	3.14	3.16	3.22	0.46	0.04	0.10	-138.49
Dissimilar Metrical and Nominal Variables							
1	40.01	39.94	39.99	0.59	2.42	1.03	57.37
2	0.99	1.08	1.02	0.53	0.21	0.10	54.81
3	12.00	12.01	12.02	0.75	1.08	0.45	58.18
4	11.81	12.54	12.03	0.55	2.31	0.98	57.45
5	1297.48	1352.24	1313.38	0.61	244.86	105.67	56.84
6	0.50	0.50	0.51	0.05	0.10	0.03	74.03
7	0.65	0.65	0.66	0.06	0.10	0.03	73.55
8	0.90	0.90	0.92	0.02	0.06	0.02	65.81
9	0.20	0.20	0.20	0.02	0.08	0.02	77.47
10	2.85	2.92	2.92	0.47	0.07	0.08	-17.82
11	2.35	2.49	2.38	0.64	0.14	0.06	58.47
12	3.14	3.16	3.22	0.48	0.04	0.10	-128.08

Notes: see table 8.

Table 11: **Analysis of the Mahalanobis Matching Distance**

	Average Values				Bias		
Variables[a]	T[b]	NT[b]	C[b]	Test Results[c]	before M.	after M.	Reduction
Similar Metrical and Nominal Variables							
1	40.02	39.97	40.10	0.15	1.18	1.00	15.21
2	0.99	1.06	1.02	0.20	0.10	0.07	27.71
3	12.01	11.99	12.06	0.22	0.52	0.38	26.96
4	11.83	12.46	12.33	0.21	1.15	1.00	12.37
5	1300.00	1346.46	1346.94	0.21	120.36	125.44	-4.22
6	0.50	0.50	0.50	0.00	0.10	0.01	89.95
7	0.65	0.65	0.66	0.00	0.09	0.01	87.31
8	0.90	0.90	0.93	0.13	0.06	0.03	48.80
9	0.20	0.20	0.18	0.06	0.08	0.02	71.44
10	2.86	2.86	2.88	0.00	0.04	0.03	4.43
11	2.35	2.35	2.33	0.03	0.03	0.04	-26.20
12	3.14	3.15	3.20	0.00	0.04	0.06	-42.00
Dissimilar Metrical and Similar Nominal Variables							
1	40.02	40.12	40.20	0.73	2.42	2.05	15.38
2	0.99	1.06	1.02	0.54	0.20	0.14	30.57
3	12.01	11.99	12.06	0.73	1.08	0.76	29.48
4	11.81	12.44	12.32	0.62	2.30	1.94	15.40
5	1298.03	1347.71	1346.13	0.69	240.39	238.98	0.59
6	0.50	0.50	0.50	0.00	0.10	0.01	89.71
7	0.65	0.65	0.66	0.01	0.09	0.01	87.39
8	0.90	0.90	0.93	0.12	0.06	0.03	50.20
9	0.20	0.21	0.18	0.06	0.08	0.02	72.78
10	2.85	2.86	2.88	0.00	0.04	0.03	3.25
11	2.35	2.35	2.33	0.03	0.03	0.04	-33.75
12	3.14	3.14	3.19	0.00	0.04	0.05	-33.40
Similar Metrical and Dissimilar Nominal Variables							
1	40.01	40.01	39.96	0.24	1.28	1.14	11.12
2	1.00	0.98	0.95	0.43	0.18	0.12	32.88
3	12.01	11.99	11.93	0.46	0.73	0.53	26.29
4	11.87	11.68	11.52	0.19	1.13	1.07	4.88
5	1301.22	1288.11	1288.56	0.18	115.23	118.65	-2.97
6	0.50	0.50	0.50	0.00	0.10	0.01	89.53
7	0.65	0.65	0.66	0.01	0.09	0.01	85.69
8	0.90	0.90	0.93	0.19	0.06	0.03	44.26
9	0.20	0.20	0.18	0.09	0.08	0.02	69.58
10	2.85	2.93	2.89	0.00	0.08	0.04	45.83
11	2.35	2.45	2.34	0.05	0.10	0.03	67.77
12	3.14	3.16	3.20	0.00	0.04	0.06	-52.13
Dissimilar Metrical and Nominal Variables							
1	40.01	39.94	40.04	0.71	2.42	2.05	15.41
2	0.99	1.08	1.04	0.54	0.21	0.14	31.74
3	12.00	12.01	12.08	0.74	1.08	0.78	27.77
4	11.81	12.54	12.43	0.63	2.31	2.00	13.29
5	1297.48	1352.24	1356.49	0.70	244.86	243.20	0.68
6	0.50	0.50	0.50	0.00	0.10	0.01	88.57
7	0.65	0.65	0.66	0.01	0.10	0.01	85.71
8	0.90	0.90	0.93	0.14	0.06	0.03	49.49
9	0.20	0.20	0.18	0.07	0.08	0.02	70.77
10	2.85	2.92	2.89	0.00	0.07	0.04	32.94
11	2.35	2.49	2.35	0.00	0.14	0.03	78.53
12	3.14	3.16	3.21	0.00	0.04	0.06	-54.01

Notes: see table 8.

Table 12: **Analysis of the Gower Distance**

	Average Values				Bias		
Variables[a]	T[b]	NT[b]	C[b]	Test Results[c]	before M.	after M.	Reduction
Similar Metrical and Nominal Variables							
1	40.02	39.97	40.01	0.09	1.18	0.23	80.47
2	0.99	1.06	0.99	0.08	0.10	0.02	78.23
3	12.01	11.99	12.01	0.10	0.52	0.11	78.35
4	11.83	12.46	11.83	0.07	1.15	0.23	79.51
5	1300.00	1346.46	1298.80	0.11	120.36	24.94	79.28
6	0.50	0.50	0.50	0.14	0.10	0.10	0.59
7	0.65	0.65	0.64	0.15	0.09	0.09	-0.79
8	0.90	0.90	0.90	0.11	0.06	0.06	-2.04
9	0.20	0.20	0.20	0.13	0.08	0.08	-2.69
10	2.86	2.86	2.86	0.64	0.04	0.08	-135.03
11	2.35	2.35	2.36	0.60	0.03	0.07	-166.17
12	3.14	3.15	3.14	0.61	0.04	0.09	-122.06
Dissimilar Metrical and Similar Nominal Variables							
1	40.02	40.12	40.04	0.28	2.42	0.40	83.58
2	0.99	1.06	0.99	0.29	0.20	0.04	79.91
3	12.01	11.99	11.99	0.40	1.08	0.22	79.52
4	11.81	12.44	11.81	0.31	2.30	0.46	79.95
5	1298.03	1347.71	1299.55	0.34	240.39	45.68	81.00
6	0.50	0.50	0.50	0.13	0.10	0.10	0.96
7	0.65	0.65	0.64	0.13	0.09	0.09	0.49
8	0.90	0.90	0.90	0.11	0.06	0.06	-2.21
9	0.20	0.21	0.21	0.13	0.08	0.08	-0.82
10	2.85	2.86	2.86	0.64	0.04	0.08	-123.32
11	2.35	2.35	2.35	0.59	0.03	0.08	-176.37
12	3.14	3.14	3.14	0.66	0.04	0.09	-123.86
Similar Metrical and Dissimilar Nominal Variables							
1	40.01	40.01	40.02	0.10	1.28	0.22	82.85
2	1.00	0.98	0.98	0.23	0.18	0.04	80.07
3	12.01	11.99	12.01	0.23	0.73	0.15	79.38
4	11.87	11.68	11.70	0.12	1.13	0.28	75.23
5	1301.22	1288.11	1289.68	0.11	115.23	25.57	77.81
6	0.50	0.50	0.49	0.14	0.10	0.10	-2.35
7	0.65	0.65	0.65	0.13	0.09	0.09	0.28
8	0.90	0.90	0.90	0.12	0.06	0.06	-2.70
9	0.20	0.20	0.20	0.12	0.08	0.08	1.63
10	2.85	2.93	2.93	0.69	0.08	0.10	-31.96
11	2.35	2.45	2.44	0.81	0.10	0.12	-22.56
12	3.14	3.16	3.15	0.64	0.04	0.09	-120.63
Dissimilar Metrical and Nominal Variables							
1	40.01	39.94	40.00	0.30	2.42	0.42	82.88
2	0.99	1.08	0.99	0.31	0.21	0.04	80.67
3	12.00	12.01	12.00	0.45	1.08	0.22	79.33
4	11.81	12.54	11.85	0.35	2.31	0.46	80.09
5	1297.48	1352.24	1302.10	0.34	244.86	45.33	81.49
6	0.50	0.50	0.50	0.17	0.10	0.10	1.57
7	0.65	0.65	0.65	0.14	0.10	0.10	0.59
8	0.90	0.90	0.90	0.13	0.06	0.06	-0.49
9	0.20	0.20	0.20	0.15	0.08	0.08	-2.62
10	2.85	2.92	2.92	0.73	0.07	0.10	-47.25
11	2.35	2.49	2.49	0.86	0.14	0.15	-6.43
12	3.14	3.16	3.15	0.66	0.04	0.09	-112.25

Notes: see table 8.