IWH
Halle Institute for Economic Research
Member of the Leibniz Association

Discussion Papers

No. 3
March 2020

flexpaneldid
A Stata Toolbox for Causal Analysis with Varying Treatment Time and Duration

Eva Dettmann, Alexander Giebler, Antje Weyh

Authors

Eva Dettmann
Halle Institute for Economic Research (IWH) – Member of the Leibniz Association,
Centre for Evidence-based Policy Advice (IWH-CEP)
E-mail: eva.dettmann@iwh-halle.de
Tel +49 345 7753 855

Alexander Giebler
Halle Institute for Economic Research (IWH) – Member of the Leibniz Association,
Department of Structural Change and Productivity
E-mail: alexander.giebler@iwh-halle.de
Tel +49 345 7753 794

Antje Weyh
Institute for Employment Research of the Federal Employment Agency (IAB)
E-mail: antje.weyh@iab.de
Tel +49 371 9118 642

Editor

Halle Institute for Economic Research (IWH) – Member of the Leibniz Association

Address: Kleine Maerkerstrasse 8
D-06108 Halle (Saale), Germany
Postal Address: P.O. Box 11 03 61
D-06017 Halle (Saale), Germany

Tel +49 345 7753 60
Fax +49 345 7753 820

www.iwh-halle.de

ISSN 2194-2188

flexpaneldid
A Stata Toolbox for Causal Analysis with Varying Treatment Time and Duration*

Abstract

The paper presents a modification of the matching and difference-in-differences approach of Heckman et al. (1998) for the staggered treatment adoption design and a Stata tool that implements the approach. This flexible conditional difference-in-differences approach is particularly useful for causal analysis of treatments with varying start dates and varying treatment durations. Introducing more flexibility enables the user to consider individual treatment periods for the treated observations and thus circumventing problems arising in canonical difference-in-differences approaches. The open-source flexpaneldid toolbox for Stata implements the developed approach and allows comprehensive robustness checks and quality tests. The core of the paper gives comprehensive examples to explain the use of the commands and its options on the basis of a publicly accessible data set.

Keywords: causal inference, staggered treatment adoption, variation in treatment timing, effect heterogeneity, event study design, conditional difference-in-differences, matching

JEL classification: A11, D61, H20, Z0

* This is a completely revised version of: *Dettmann, Eva; Giebler, Alexander; Weyh, Antje:* flexpaneldid. A Stata Command for Causal Analysis with Varying Treatment Time and Duration. IWH Discussion Paper No. 5/2019. Halle (Saale) 2019.

1 Introduction

Difference-in-differences approaches have become a very popular research design for treatment effects estimation. For example, 20 percent of all empirical articles published by the American Economic Review between 2010 and 2012 have used this design (de Chaisemartin and D'Haultfoeuille 2019). In most empirical applications, the basic approach with two groups and two observation times is implemented, implicitly assuming constant treatment effects. In the context of improving access to data bases, however, researcher are able to collect more and better information on treated units and potential controls, with more than two times of observation. Now, not only treatments with fixed start dates can be described; more often we find individual treatments characterized by varying start dates and different treatment durations. In this case, a constant treatment effect over time is not plausible, and we find a growing number of studies dealing with potential biases and solutions for this problem.

One of the current strands of literature develops approaches within the framework of the staggered adoption design, where units that are treated once in the observation time are regarded as treated units from that date onwards. The approach we describe in this paper belongs to this group. To gain more flexibility we modify the conditional difference-in-differences approach of Heckman et al. (1998) in three ways. First, we include individual treatment time information from the panel into the matching process. Second we introduce a combined statistical distance function for matching. Third, we incorporate flexible observation durations into the difference-in-differences estimation. This flexible conditional DID approach ensures that varying treatment phases can be accounted for in an appropriate way and that the point in time an individual is compared to his 'statistical twin' can be exactly determined.

Our second contribution to the empirical literature is the development of an estimation tool that implements our approach and comprehensive robustness checks and quality tests. The Stata commands `flexpaneldid_preprocessing` and `flexpaneldid` are provided as an open-source toolbox, available at
https://cloud.iwh-halle.de/index.php/s/flexpaneldid.

The remainder of the paper is organized as follows. In section two, the special data structure and related challenges resulting from the treatment at different times and of different durations are explained in more detail. Section three gives an overview on empirical approaches dealing with this data structure when estimating causal effects and current enhancements in the econometric literature. In section four we introduce the flexible conditional difference-in-differences approach. Subject of section five and six is the presentation and explanation of the Stata tool consisting of two commands, `flexpaneldid_preprocessing` and `flexpaneldid`. In section five the syntax of the commands and some general instructions are given, in section six we present comprehensive examples for the use of the Stata tool.

2 Special characteristics of the data structure

The data structure we focus on is characterized by panel information on treated and non-treated observations, where treatment can start and end basically at every time. The described treatment structure is typical for policy interventions common in industrial and placed based policy (e. g. investment subsidies and R&D funding), labor market programs (e. g. support for start-ups or training vouchers) or research funding (e. g. EU funding for scientists). Besides, this type of treatment is typical in medicine, public finance, finance, economics of education or labor market research.

The flexibility in the treatment adoption implies some special characteristics that must be considered when estimating the treatment effect. Figure 1 illustrates the data structure with the so called *'staggered treatment adoption'* (Athey and Imbens 2018; Callaway and Sant'Anna 2019): an unbalanced panel data set of treated $(T_1, T_2, T_3, \ldots T_x)$ and non-treated units $(NT_1, NT_2, NT_3, \ldots NT_y)$ for the years 2004-2014 with varying dates of treatment application and individual durations from application to the start of the treatment and individual treatment durations. In this case, every time is a mix of individual pre-treatment, treatment and post-treatment phases, and we observe different 'sub-periods' for the treated units – in terms of the treatment start dates as well as

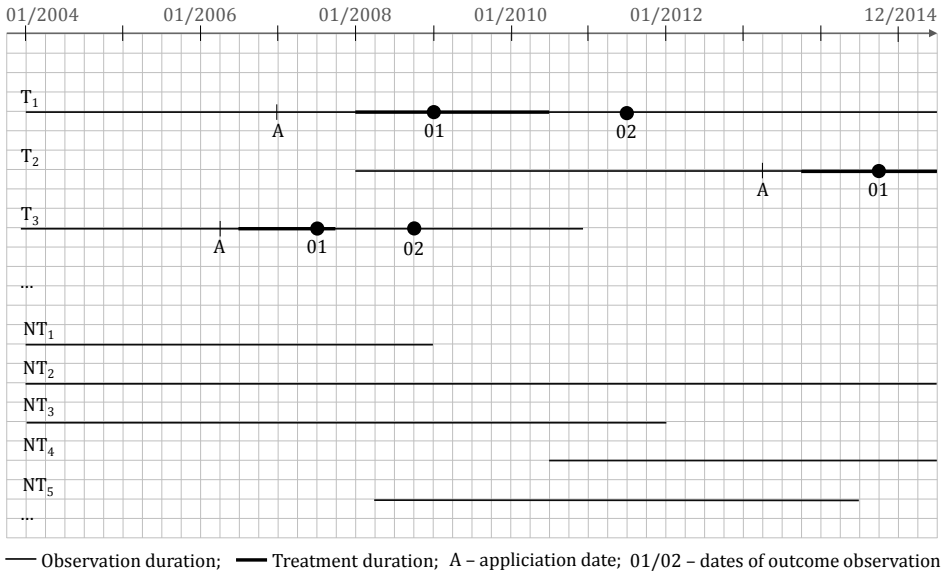

Figure 1: sketch of the typical data structure

As we know from Heckman et al. (1997, 1999), the economic environment influences the economic performance of persons or firms and should be considered when analyzing treatments effects. In a dynamic environment, this applies not only to the place but also

to the time of observation. The example in figure 1 can illustrate that: in the year 2008, the financial crisis 'arrived' in real economy. In the following recession, the economic environment strongly deteriorated – resulting in a worse economic performance of many firms and also persons. The subsequent upswing phase caused a significant performance improvement for firms and persons. If we would ignore this development in the economic environment and would compare e. g. a firm before to a firm after the crisis, we would compare them in different economic environments. This is referred to as *'calendar time effect'*. Also the treatment effect itself is influenced by the economic situation and can be heterogeneous over time (Bergemann et al. 2009).

Another phenomenon is referred to as *'dynamic treatment effect'* and means that the size of the effect may depend on the length of exposure to it (Callaway and Sant'Anna 2019). Observing e. g. the outcome development from application to one year later (from point A to point O1 in figure 1) or from application until one year after the treatment is finished (denoted by point O2) would mean that we estimate different treatment effects in this case. For example, Jacobson et al. (1993) argue that the earnings effect for displaced workers tend to be large immediately after displacement and get smaller over time.

We also find examples for *'selective treatment timing'* in the literature. In the presence of a positive temporary shock to a specific sector, firms in this sector might be more willing to invest and to apply for a subsidy at this time (Pellegrini and Centra 2006). Another example is an inter-temporal substitution of investments due to a restricted application time for investment subsidies that is observed e. g. in Bronzini and de Blasio (2006) observe.[1]

The *'Ashenfelter's dip'* (Ashenfelter 1978) or *'fallacy of alignment'* (Heckman et al. 1999) denotes the phenomenon that the anticipation of a treatment may lead to a temporary change in the behavior of the applicants. Examples for such behavioral changes are mainly discussed with regard to active labor market programs; e. g. unemployed persons may reduce their search effort for a new job when they anticipate their participation in a training program (Bergemann et al. 2009).

The above described special impacts on the treatment effect can only be considered by including the information on individual treatment times and 'sub-periods' into the estimation. In recent literature, we find some attempts to implement time information in existing models for causal inference. Another approach, the *flexible conditional DID* will be introduced in this paper.

3 Estimation approaches in empirical literature and current enhancements

In recent empirical literature we find a growing number of causal analyses of treatments in various economic fields that are based on data sets with a similar structure to the

[1] This phenomenon cannot be illustrated with the figure.

above described one. This literature review contains examples from labour market economics and health economics as well as evaluation studies of active labour market policies and place based policies. In this overview, we focus on examples for causal effects estimations on the basis of panel data with more than two observation times.

3.1 Canonical DID approaches

Following comprehensive overviews on different evaluation approaches (e. g. Abadie 2005; Blundell and Costa Dias 2009; Imbens and Wooldridge 2009) and controversial discussions in the literature,[2] the idea of a combined control for selection resulting from observable and unobservable heterogeneity found a widespread adoption in empirical literature.

One strand of the empirical literature is based on the (nonparametric) *'conditional DID'* introduced by Heckman et al. (1998), that combines a DID estimation and matching. Here the compared outcome changes are defined conditional on matched samples instead of the whole samples of treated and non-treated units. For example Bergemann et al. (2009) use a combination of kernel matching and DID to analyze causal effects of training in East Germany on transition from non-employment to employment. They allow for heterogeneous treatment effects resulting from calendar time by estimating different effects for distinct fixed time periods and discuss different ideas to control for a potential *Ashenfelter's dip*. The paper of Pellegrini and Centra (2006) evaluates the impact of investment subsidies on the performance of firms in manufacturing and firm services sectors in the Mezzogiorno region. Using a combined matching and DID estimator, the authors consider different treatment durations to capture dynamic treatment effects. The 'virtual project duration' for the controls are defined with the help of the start date of the projects and the average project duration of different auctions. Similarly, Bernini and Pellegrini (2011) analyze subsidies allocated to manufacturing firms in the southern Italian regions. They apply a canonical(kernel and stratification) matching and DID estimator to compare subsidized an rejected firms. Caliendo and Künn (2011) provide empirical evidence on the effectiveness of start-up subsidies for unemployed persons. A combination of matching of administrative data and subsequent DID of different outcomes reported in repeated computer assisted telephone interviews allow them to estimate long-term effects and consider effect heterogeneity among the interviewed persons.

Another very popular strategy incorporates the idea of the nonparametric DID estimator into a conventional panel regression model. The basis for such *'canonical DID models'* is that in the case of two analyzed groups and two time periods (which is referred to as the *'2x2 case'* (Goodman-Bacon 2019, e. g.)), the nonparametric DID estimator equals the the respective coefficient (of the interaction of the treatment group dummy and the post-treatment-period dummy) in the two-way fixed effects DID model[3] , if

[2] See the discussion of Dehejia and Wahba (1999, 2002); Dehejia (2005) and Smith and Todd (2005*a*,*b*) as the most prominent example.

[3] This term denotes a panel DID model with time and individual fixed effects (Athey and Imbens 2018).

the the common trend assumption is fulfilled (Angrist and Pischke 2009). Within this model framework, Autor (2003) assesses the contribution of unjust dismissal doctrine on employment outsourcing in the U.S. He includes 'lags' of the dependent variable to control for a potential *Ashenfelter's dip* and 'leads' to verify dynamics of the treatment effect over time. In a similar way, Neumark and Kolko (2010) evaluate the effectiveness of California's enterprise zone programs on employment with panel data. Ham et al. (2011) use a fixed effects DID model comparing treated Census tracts with their (in geographical sense) nearest neighbor tracts to measure the impact of different place-based policies on local labor markets. Bronzini and de Blasio (2006) estimate the causal effects of investment subsidies in Italy within a DID model including yearly post-treatment dummies to capture the time varying treatment effect. Marcus and Siedler (2015) analyze the effect of the late-night alcohol sales ban on alcohol-related hospitalization rates in Germany. They apply fixed-effects DID models with various control variables to estimate the mean effect of the ban as well as the development over time.
Heyman et al. (2007) use a combination of year-by-year Propensity Score Nearest Neighbor Matching and the canonical DID model to estimate the influence of foreign ownership on wages in Swedish firms. For a panel of Swedish manufacturing firms, Greenaway et al. (2005) perform a yearly Propensity Score caliper matching to create a pooled data set for the subsequent random effects DID model estimation of causal links between exports and firm performance. Freier et al. (2015) use a canonical DID model combined with entropy balancing of Hainmueller and Xu (2013) to estimate the effect of graduating from university with an honors degree on subsequent earnings in Germany.

3.2 Current developments in econometric literature

The use of canonical DID models in a panel data context rests on the presumption that the above mentioned equality of the DID estimator and the coefficient in the two-way fixed effects DID model can be generalized to more than two groups and/or more than two observation times. In the last few years, however, doubts in the appropriateness of this generalization and the implicite assumption of effect homogeneity are raised in the literature, especially when treatment time varies and the treatment effect is dynamic. In the following, we describe different decompositions of the canonical DID model that are the basis for the definition/derivation of the sources of bias in case of effect heterogeneity, and in some cases this decomposition is the starting point for the presentation of an advanced estimator. All of the approaches presented in the following include the timing of the treatment in one or the other way instead of characterizing the treatment by a binary variable (like in the 2x2 case). In so called *'staggered adoption designs'* (Athey and Imbens 2018; Callaway and Sant'Anna 2019), the treated units can then be categorized by groups (or cohorts) based on when they first receive treatment.[4] Some (more general) approaches also allow for changes in the treatment status over time.

[4]Such approaches are also regarded as *'stacked DID'* (Goodman-Bacon 2019) or (in the case of relative time definitions) as *'event studies'* (Abraham and Sun 2019). 'Relative time definition' means in that context that the absolute time is normalized such that the the observation period is measured with respect to treatment time.

Borusyak and Jaravel (2017) illustrate in an example very clearly the distinction between estimations in the 2x2 case and the staggered adoption design. They also show that, with canonical two-way FE models, the effect of early adopted treatments are underweighted in the estimated average treatment effect and demonstrate the consequences of different types of deviations from the assumed effect homogeneity for the estimated effects.
Exploiting the normal distribution assumption underlying the canonical DID model, Goodman-Bacon (2019) decomposes the estimator into a weighted average of all possible 2x2-two-way FE-DID models. The weight of each 2x2 model depends on the subsample size and the treatment timing.[5] Based on this decomposition, Goodman-Bacon (2019) illustrates the sources of the estimation bias in the canonical DID model: if the size of the effect is associated with the number of the treated units and/or the treatment timing, and in case of dynamic treatment effects. The decomposition also suggests that more flexible specifications (for example within an event-study framework) may be more robust.

Deshpande and Li (2017) use a (sligthly modified) two-way fixed effects DID model in an event study design to estimate the effect of closings of Social Security Administration field offices on the number of disability recipients. They exploit the variation in the timing of closure to compare (earlier) treated with later treated regions. In this design, the common trend assumption relaxes to the requirement that the timing of the closings must be as good as random rather than the closings themselves being random events. In a similar fashion, Fadlon and Nielsen (2017) construct counterfactuals to households affected by severe health shocks by using households that experience the same shock a few years later to estimate the effect on the spouses's labor supply. Their first step is to define the time of observation in relation to the year of the shock within a specific birth cohort. Households with individuals from the same cohort who experience the same shock some years later serve as controls. For these households, 'placebo shocks' are assigned to the data in order to create relative observation time in the same way as for the treatment group. In a second step, the treatment effect is estimated by a simple (non-parametric) dynamic DID estimator, i. e. a year-by-year comparison of both groups. A canonical DID model with household fixed effects is used additionally to estimate the mean effect.

Also in an event study design, Athey and Imbens (2018) show that the canonical DID is an unbiased estimator of a particular weighted average treatment effect consisting of partial effects of different adoption dates. They explicitly formulate two (rather strong) exclusion restrictions that must be fulfilled to be able to simplify the partial estimators to binary treatment indicators for every adoption time period: *no anticipation*, i. e. the current non-treatment outcome is not influenced by a future treatment. And *invariance to history*, i. e. for units that adopted the treatment earlier, the treatment outcome in the current period is not influenced by the treatment duration. Under these

[5] The weights combine the absolute size of the respective subsample, the relative size of the treatment and control group within the subsample and the timing of the treatment. Timing has two components, the time when the treatment is observed, and the 'time window' or the duration, in which the subsample is observed.

assumptions, the canonical DID model can be regarded as the weighted average of the effects of changes in the adoption dates: the effect for switching from never adopting to adopting in the first period, the one for switching from never adopting to adopting some time later, and the effect for changing from an earlier adoption date to adoption at the initial time period.
Abraham and Sun (2019) propose a cohort-specific treatment effect estimator (for the treated a specific number of periods after the initial treatment) where time is defined relative to the initial treatment. This *'interaction-weighted estimator'* is estimated using the linear two-way FE specification with interactions of relative time and cohort indicators and weights each cohort-specific estimator according to the sample share of the cohort in the respective time period. A similar idea is the nonparametric decomposition of the treatment effect in Callaway and Sant'Anna (2019). Here, *'group-time average treatment effects'* for groups of treated individuals are defined according to the time of the first treatment. Counterfactuals for the group members are found among non-treated units using propensity score matching, the group-specific effects are estimated using nonparametric DID. Depending on the treatment context (selective treatment timing, dynamic treatment effects, and calendar time effects), Callaway and Sant'Anna (2019) propose different aggregation procedures for the *group-time average treatment effects* that accounts for effect heterogeneity related to treatment timing.

In a more general setting, where units can switch between treatment and control position over time, de Chaisemartin and D'Haultfoeuille (2019) decompose the canonical DID model into different 2x2 comparisons. They show in a simple example, that in case of heterogeneous (or time-dependent) treatment effects, the weights of the partial estimators can become negative, if the control group is treated in the pre- and the post-treatment period.[6] Like in the studies of e. g. Goodman-Bacon (2019) and Borusyak and Jaravel (2017), the negative weights are the reason for the bias of the canonical DID model, if the treatment effect is heterogeneous. Similar to the idea of Abraham and Sun (2019), the average effect estimator is then expressed as a weighted average of partial individual estimators for different groups in different time periods. The suggested *'Wald-Time-Corrected estimator'* is a weighted average of comparisons of mean outcome developments between groups that switch from one treatment status to the other (from non-treatment to treatment, or vice versa) and groups that do not change the treatment status in the same time period.[7] Additionally, de Chaisemartin and D'Haultfoeuille (2019) propose test diagnostics for the presence of biased estimators. Within the same setting, Imai and Kim (2019*a*) establish the equivalence between matching and weighted two-way fixed effects estimators and decompose the canonical DID model into a weighted average of the estimators for unit fixed effects, time fixed effects, and pooled regression estimators - with negative weights for the pooled regression estimator. Based on this decomposition, they propose a (nonparametric) *'multi-period DID estimator'* that combines three sets of observations: the 'within-unit matched set' (which contains previous observations of the treated unit), the 'within-time matched

[6] Especially in periods with many treated groups and for groups that are treated for many periods, negative weights may be more likely.

[7] In the staggered adoption design, the estimator compares groups that switch from non-treatment to treatment with non-treated groups.

set' (which is defined as a group of control observations in the time of treatment), and the 'adjustment set' (which contains previous observations of the controls). The *multi-period DID estimator* is the average of the 2x2-DID estimators applied whenever there is a change from the control status to the treatment status.[8]

Imai et al. (2019) propose a matching-based DID estimator for time series cross sectional data. The estimator selects potential control observations for every treated unit at a specified time period using propensity score matching and weighting schemes. The first step is to align the treatment history for a specific time span via exact matching, thus creating matched sets for the treated units. The matched sets are then refined by caliper matching of the pre-treatment-outcome and additional covariates. The last step is the DID estimation as weighted average of individual differences, i. e. mutual comparisons of the outcome development of the treated and the development of the average outcome in the respective refined matched set. Imai et al. (2019) show that the proposed estimator is equivalent to a weighted linear two-way FE model under certain assumptions. As a measure for matching quality, they propose a check of the mean standardized difference between a treated and its matched control in each covariate at each pre-treatment time period, aggregated across all treated observations for each covariate and each time period.

In the next chapters 4 and 5, we present the nonparametric *'flexible conditional DID estimator'* and its implementation, the Stata toolbox `flexpaneldid`. This approach has similarities especially to the approaches of Callaway and Sant'Anna (2019) and Imai et al. (2019). The basic idea is also to combine matching and DID to find adequate controls for the treated units. Like Callaway and Sant'Anna (2019) we apply the staggered adoption framework and define time in relation to the treatment start. The *flexible conditional DID estimator* can be regarded as a special case of the *group-time average treatment effects* approach with the number of groups equal to the number of treated observations and respective group sizes of one. The single group-time estimators are summarized in a simple weighted average with respective group weights of one. Different from the mentioned approach, we select control observations individually for every treated unit and compare individual outcome developments - which is similar to the approach proposed by Imai et al. (2019). Different from both above mentioned approaches, we propose a statistical matching procedure that gives equal weights to each included covariate. This statistical distance function gives a 'pure' description of the similarities and disparities regarding the individual covariates, and the overall indicator reflects the comparability of the observations without covariate weights in favour of 'important' or particularly similar/dissimilar covariates. Since we consider the time information in the matching process, this approach is very flexible in the definition of treatment start and treatment duration.
Moreover, the definition of potential counterfactual events in the first step of the approach described in Fadlon and Nielsen (2017) is very similar to the data preprocessing of the *flexible conditional DID*.

[8] A similar model can be found in Imai and Kim (2019*b*) for unit-fixed effects DID models.

4 The flexible conditional DID

The aim of the approach is to consider problems associated with heterogeneous treatment effects in a panel data context (see section 2). In the *flexible conditional DID*, we incorporate the observation time information from the panel data into the matching process. The applied preprocessing decribed below will remove a potential calendar time effect.[9] Defining different observation periods for the outcome comparisons (i. e. estimating more than one treatment effect) may also consider a dynamic treatment effect. Moreover, the *flexible conditional DID* helps to account for behavioural changes like the *Ashenfelter's dip* in that it enables the user to consider expectations on the duration of the 'dip' and exactly determining the matching and the outcome observation time (in relation to the treatment start).

In the flexible conditional DID, the idea of the nonparametric conditional difference-in-differences approach introduced by Heckman et al. (1998) is transferred to the framework of the staggered adoption design. [10] In this combination of matching and DID, the *conditional independence assumption* for matching and the *common trend assumption* required for DID are replaced by the *'conditional parallel trend assumption'* (Callaway and Sant'Anna 2019), implying that unobservable individual characteristics must be invariant over time for units with the same observed characteristics. As for usual matching, the *common support condition* must be fulfilled (Callaway and Sant'Anna 2019). Additionally, the approach assumes no spillover effects (this corresponds to the *stable unit treatment value assumption* for matching), and that potential carryover effects do not influence the matching variables at the matching time (Imai et al. 2019). The last assumption is usual for the staggered adoption design and is referred to as *'irreversibility of treatment'* (Callaway and Sant'Anna 2019), i. e. if a unit receives a treatment, it is regarded as treated unit for all the following time periods.

The first step of the flexible conditional DID approach is an extensive data reorganization to incorporate the observation date of all matching variables and outcomes. Hence, we limit the set of potential partners for every treated unit to those observed just at the individual matching date, e. g. the treatment start. Then the matching algorithm selects one or more statistical twins among these pre-selected units.[11] In this step, we normalize the observation time of the matching variables and the outcomes such that they are measured with respect to the individual treatment start.

The second step is matching. Basically, each matching process that allows for (at least partial) exact matching is suitable. This exact matching option is required to consider the time information from the pre-selection process. As a novelty, we add a matching based on a combined statistical distance function.[12] This distance function

[9] Fadlon and Nielsen (2017) use a quite similar approach, and they state: 'By construction, this research design [...] mechanically nets out calendar [...] effects.'

[10] Ho et al. (2007) denote the matching process in this context as a nonparametric data preprocessing in the sense that it leads to more reliable causal effect estimates by reducing bias and variance.

[11] For example, when a firm receives investment subsidies in January 2007, we consider its characteristics in this month and will assign a firm which has similar characteristics in January 2007.

[12] In a previous simulation study, statistical distance functions proved to be superior to Mahalanobis metric and Propensity Score matching, especially in small samples. For more details see Dettmann

follows an idea of Kaufmann and Pape (1996) and can be described as the weighted average of scale-specific distance functions. It belongs to the group of linear-homogeneous aggregations (Opitz 1980). For our analysis, we consider the results of Dettmann et al. (2011) and combine the mean absolute difference for continuous and the generalized matching coefficient for categorical variables as they proved to be superior to other distance measures in the simulation analysis.[13]

When combining scale-specific distance functions, the functions usually have to be normalized and transformed (Diday and Simon 1976). In our case, the differences in the continuous variables are normalized by the maximal observed differences of the respective variables. The similarity information of the generalized matching coefficient is transformed into a distance measure. Weighting the functions by the respective number of variables, the distance function for a treated firm i and a non-treated firm j can be described as follows:

$$Dist_{ij} = \frac{1}{N}\left[N_m \cdot AD_{ij} + N_n \cdot (1 - GMC_{ij})\right]. \tag{1}$$

The terms $Dist_{ij}$, AD_{ij} and GMC_{ij} denote the aggregated distance function and the scale-specific distances, N is the total number of variables with $N = N_m + N_n$, where N_m is the number of continuous variables and N_n that of the categorical ones.

The mean difference of the continuous variables is calculated using the normalized absolute difference:

$$AD_{n,ij} = \frac{1}{N_m}\sum_{n=1}^{N_m}\frac{|x_{ni} - x_{nj}|}{diff_{max}(x_n)}$$

where $||$ denotes absolute values, and $diff_{max}(x_n)$ is the maximum observed difference of variable x_n.

The generalized matching coefficient GMC_{ij} can be defined as the share of covariates with equal values in all categorical variables:

$$GMC_{ij} = \frac{1}{N_n}\sum_{n=1}^{N_n} Q(x_{ni}, x_{nj}) \quad with \quad Q(x_{ni}, x_{nj}) = \begin{cases} 1 & \text{if } x_{ni} = x_{nj} \\ 0 & \text{else.} \end{cases}$$

As can be observed from the equation, using the generalized matching coefficient

et al. (2011). The reason for the better (in the sense of 'more similar') control groups compared to the results of the Mahalanobis metric may be seen in the consideration of the different scales of the matching variables in the presented approach. The weakness of Propensity Score matching is that estimating the score implies a weighting scheme of the variables according to their influence on the treatment probability, not on the outcome (Zhao 2004). This may result in quite different outcomes for units with identical scores, particularly in small samples (Fröhlich 2004). King and Nielsen (2016) and King and Zeng (2006) doubt the Propensity Score to be suitable for empirical studies, when the score itself has to be estimated.
Another alternative is the Coarsened exact matching procedure of Blackwell et al. (2009). This approach is implemented as option in the `flexpaneldid` command.

[13] Although the alignment of the history of outcome and covariates instead of considering their current values, as is proposed in Imai et al. (2019) is not explicitly described for our approach, such data can be included in the matching process as well.

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. [14]

Based on this matching process, the average treatment effect for the treated ATT is estimated. Within the framework of the conditional DID model, usually the mean outcome developments in the treated and the control group are compared. Different from the standard model, the flexible conditional DID compares individual differences in outcome development between the treated firms i and their respective controls j. The estimator is the mean of the individual comparisons.[15]

$$ATT = \frac{1}{I} \sum_{i=1}^{I} \left[(Y_{i,t_{0i}+\beta_i} - Y_{i,t_{0i}}) - (Y_{j,t_{0i}+\beta_i} - Y_{j,t_{0i}}) \right]. \tag{2}$$

As can be observed from equation 2, we include individual treatment start dates, denoted by index t_{0i}, and a flexible number of time units, e. g. months, $t_{0i} + \beta_i$, reflecting the individual duration from treatment start to outcome observation. Y denotes the outcome.[16]

To draw causal inference in the presence of non-random sampling we apply a t-test with corrected standard errors. The correction terms are implemented using the matching-based procedure of Abadie et al. (2004); Abadie and Imbens (2006, 2011). The number of matches we fix to two (like the default setting in the `teffects nnmatch` comand in Stata.)

Due to heterogeneous treatment durations, the observed periods may be heterogeneous among the treated individuals. The average treatment effect for the treated is thus a weighted average of different observation periods.[17]

With the `flexpaneldid` toolbox we provide an easy-to-apply implementation of the proposed matching and DID approach within the staggered adoption design. The command provides an option for the inclusion of the pre-treatment outcome development for a user-defined time period and also trends or developments of matching variables can be included, if they are defined in the data or prior to the use of `flexpaneldid`. Therefore, we want to make one important remark before presenting the command. It has been considered as common knowledge for the use of matching procedures that one should include those variables that influence the treatment assignment and the outcome

[14] Like in Imai et al. (2019), treated observations for whom we do not find suitable matches, are excluded from the sample to preserve internal validity.

[15] For simplicity of the notation, equation 2 denotes the case of a nearest neighbor matching; when more than one non-treated observations are selected as controls, the control outcomes are calculated as the average outcome over all selected controls. This is the case for radius matching and ties, which are are also implemented in the `flexpaneldid`.

[16] When more than one continuous covariate is used for matching, the resulting estimator will be biased without an adjustment. The same is true for categorial variables that are not exactly matched (Abadie and Imbens 2006, 2011). Following Abadie et al. (2004), we apply the regression-based bias correction to adjust the start and the end values of the outcome development.

[17] In case one wants to specify the observation period to be common for all treated units, the `flexpaneldid` command includes the option to define the observation period in relation to the treatment start (`'outcometimerelstart'`).

(Heckman et al. 1999). When choosing matching variables in the context of panel data, e. g. inclusion of outcome development and pre-treatment history of covariates as is proposed by Imai et al. (2019), the behaviour of the covariates over time should also be considered. Since matching is a sample selecting technique, especially the so called *'regression to the mean'* needs to be controlled for.[18]

5 The `flexpaneldid` toolbox

Under the following link (https://cloud.iwh-halle.de/index.php/s/flexpaneldid), we provide a Stata toolbox that implements the described flexible conditional DID approach and comprehensive quality and robustness checks. In the following, we describe the two commands `flexpaneldid_preprocessing` and `flexpaneldid` and explain the options that can be selected.

Before running the `flexpaneldid` toolbox one has to install or update the Stata ado-files `psmatch2, pstest` and `cem`, which are used in the `flexpaneldid` command.

5.1 Description

`Flexpaneldid` is a Stata toolbox for causal analysis of treatments with varying start dates and varying treatment durations within panel data with more than two observation times. It consists of two commands based on each other, `flexpaneldid_preprocessing` and `flexpaneldid`. In the `flexpaneldid_preprocessing`, the original data set is rearranged in that individual selection groups for every treated unit are created which contain all potential controls. The result of this preprocessing is a temporary dataset with information that are crucial for the use of the `flexpaneldid`.

Based on the temporary data set, the `flexpaneldid` estimates the average treatment effect for the treated. For this step, different matching approaches are available. Additionally, quality and robustness checks can be conducted.

The `flexpaneldid` toolbox contains many relative time definitions. The following graph illustrates the relationships between treatment start and related time definitions and treatment end and related time.

[18] The selection process may choose potential control units with extreme values relative to the group means in order to find partners for the treated. If those variable values vary over time, the matched units will 'regress back' toward the means of the groups from which they are selected (Daw and Hatfield 2018).

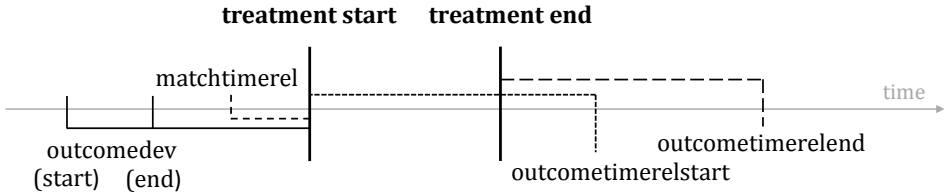

Figure 2: sketch of the relative time definitions in `flexpaneldid`

5.2 flexpaneldid_preprocessing

Syntax

```
flexpaneldid_preprocessing, id(varname) treatment(varname) time(varname)
    matchvars(varlist) matchtimerel(integer)
    [matchvarsexact(varlist) prepdataset(string) replace]
```

Inputs

`id(varname)` uniquely identifies objects in the panel dataset; the variable must be an integer or string.

`treatment(varname)` contains the variable defining the treatment; input must be in 0-1 format.
Important note: The variable must equal to one for the whole treatment phase. In case of repeated treatments for one unit (identified by a unique id), the repeated treatments are handled as one treatment phase.

`time(varname)` identifies the time information in the panel; input must be an integer indicating an absolute time, e. g. year, month, quarter.
Important note: If the data contain only information in 'date' format, this information must be converted into an integer.

`matchvars(varlist)` should contain all variables that may be used for matching.

`matchtimerel(integer)` is a relative time specification (in relation to the treatment start) that defines the time of matching; default = 0 (if no matching time is defined). In this case, *flexpaneldid_preprocessing* uses variable values observed at the treatment start.
Important note: The dimension of the parameter in parentheses depends on the dimension of time that is defined for `time`.
For example, `matchtimerel(-1)` means that the matching process is conducted one year before the treatment starts, if the dimension of the time variable is years.

Options

Option `matchvarsexact(`*varlist*`)` indicates those variables that are used for exact matching. Exact matching variables are applied already at the preprocessing step.

The `prepdataset(`*string*`)replace` option specifies the path where the data set containing the preprocessing result is stored. `replace` overwrites any existing data set with the new data. We highly recommend the use of this option.[19]

5.3 flexpaneldid

Syntax

```
flexpaneldid depvar, id(varname) treatment(varname) time(varname)
    prepdataset(string)
    (statmatching(con(varlist) cat(varlist) [ties | radius(float)]) |
    cemmatching(varname1 [(cutpoints1)] [varname2 [(cutpoints2)]...] [k2k])
    (outcometimerelstart(integer) | outcometimerelend(integer))
    [outcomedev(integer [integer]) test didmodel outcomemissing]
```

Before starting `flexpaneldid`, the user must reload the original data.

Inputs

depvar defines the analyzed outcome; the input must be numerical.

`id(`*varname*`)` uniquely identifies objects in the panel dataset; the variable must be an integer or string.

`treatment(`*varname*`)` contains the variable defining the treatment; input must be in 0-1 format.
Important note: The variable must equal to one for the whole treatment phase. In case of repeated treatments for one unit (identified by a unique id), the repeated treatments are handled as one treatment phase.

`time(`*varname*`)` identifies the time information in the panel; input must be an integer indicating an absolute time, e. g. year, month, quarter.
Important note: If the data contain only information in 'date' format, this information must be converted into an integer.

`prepdataset(`*string*`)` specifies the path where the preprocessing data set is stored. The information in this data is crucial for the use of the `flexpaneldid`.

[19] When the path is given, the `flexpaneldid` can repeatedly be applied without running the preprocessing again.

Options

One of the two options for the distance metric for matching must be selected:

`statmatching(`*`con(varlist) cat(varlist)`*`)` indicates that the statistical distance function according to equation 1 is used for matching. The variable names included in *`con(varlist)`* indicate the continuous matching variables, which must be numerical variables; *`cat(varlist)`* contains the categorical variables, which must be integers. The default matching algorithm is a nearest neighbor matching with replacement.[20] Alternatively, one can chose between the options `ties` and `radius(`*float*`)`. Option `ties` means that if more than one nontreated is the best partner for a treated observation, the counterfactual outcome is constructed using all nontreated with equal distance. Option `radius(`*float*`)` indicates that all nontreated within the defined radius are used to construct the counterfactual outcome. The defined radius must be a float number in the range between 0 and 1.

`cemmatching(`*`varname1 [(cutpoints1)] [varname2 [(cutpoints2)]...]`*`)` indicates that the *Coarsened Exact Matching* of Blackwell et al. (2009) will be executed.[21] Like in the `cem` command, including cutpoints for the matching variables is possible, either formats [*(#integer)* or *(numlist)*] are allowed. Using option `k2k` creates matched strata with equal numbers of treated and controls. See Blackwell et al. (2009) for more details on the `cem` command.

The `flexpaneldid` command enables the user to define the period of outcome development that should be compared between treated and controls. (The starting point of the observed development coincides to the start of the treatment.) One of both options must be selected:

`outcometimerelstart(`*`integer`*`)` is a relative time specification that defines the end of the outcome development in relation to the treatment start. In case of repeated treatments, the relative time refers to the start of the first treatment.
Important note: The dimension of the parameter in parentheses depends on the dimension that is defined for `time`.
For example, `outcometimerelstart(`*`3`*`)` means we observe the outcome development from the individual treatment start to three years after the start of the treatment, if the dimension of the time variable is years.

`outcometimerelend(`*`integer`*`)` is a relative time specification that defines the end of the outcome development in relation to the treatment end. In case of repeated treatments, the relative time refers to the end of the last treatment.
Important note: The dimension of the parameter in parentheses depends on the dimension that is defined for `time`.
For example, `outcometimerelend(`*`2`*`)` means that we compare the outcome development from the treatment start to two years after the treatment is finished, if the dimension of the time variable is years.

[20] This option refers to `psmatch2, neighbor(`*`1`*`) pscore(`*`statistical distance`*`)`.
[21] To ensure reproduceable results, a 'hard coded' seed value is set in the ado-file.

The command provides the possibility to consider the pre-treatment outcome in the matching process. Two options are available. `outcomedev(integer)` selects the level value of the outcome at a time defined in relation to the treatment start.
`outcomedev(integer integer)` defines an outcome development, the two integers give the start and the end of the development in relation to the treatment start.
Important note: The dimension of the parameter(s) in parentheses depend on the dimension that is defined for `time`. Both parameters are required to be *integer* ≤ 0.
For example, `outcomedev(-3 -1)` means the outcome development from three to one year(s) before the individual treatment starts, if the dimension of the time variable is years. `outcomedev(-3)` considers the outcome level three years before treatment starts as additional matching variable.

`test` executes quality tests after matching. The tests conducted in `pstest` and *quantile-quantile plots* are presented. Further test are presented depending on the matching process that is selected: For `cemmatching` additionally the overall imbalance measure L_1 and univariate imbalance measures described in Blackwell et al. (2009) are displayed. For `statmatching`, *KS-tests* for continuous variables and *chi-square tests* for the categorical variables are executed in addition.

`didmodel` is an option for robustness checks on the basis of a standard two-way DID model. The first model mimics the 2x2 case (two groups and two observation times) in a fixed effects model, namely the treatment start and the end of the defined period of outcome development. The second model is a canonical fixed effects DID model with standard errors allowing for intragroup correlations. The observation period is trimmed at the defined end of the outcome development.[22] *Important note*: Since both models are based on the assumption of homogeneous effects (see chapter 3 for more details), they should be used as robustness checks only, but not as standalone estimations.

Finally, every matching process is characterized by a tradeoff between sample size for the treatment effect estimation and definition of the best control group in terms of comparability. Default for the command is a check if the *depvar* is observable for the defined period of outcome development before matching is performed. (This reflects the preference for the sample size to be as big as possible). Using the `outcomemissing` option disables this check. It should only be used with rather big data sets when the user has a preference for the control group to be as similar as possible.

[22] Both models are estimated using the Stata command `xtreg`. The exact specification of the models is given in the output.

Stored information

Running `flexpaneldid` produces a new data set with the following variables that can be saved if required:

`id`	unique identifier
`treatment`	variable defining the treatment
`first_treatment`	variable defining time of the treatment start
`last_treatment`	variable defining the time of the treatment end
`nt_multi_select`	number of assignments for multiple assigned non-treated observations (for units assigned only once, this variable contains a missing)
`time`	time information
`depvar`	analyzed outcome
`panel_id`	observation identifier for internal use
`post_treat_dummy`	dummy indicating the the time after treatment
`post_treat_dummy_rel_time`	count variable for time periods after treatment

6 Detailed application examples

In this section, we present four comprehensive examples to illustrate the use of the `flexpaneldid` toolbox and the outputs resulting from the available options. In order to reproduce the data characteristics described in section 2 (see figure 1), we need a panel data set with more than two observation times, individual treatment starts and treatment durations, possible multiple treatments and differently scaled variables characterizing the observed units. We start with a publicly available data set, the 'patent' data provided by Wooldridge (2010)[23] and generate some additional variables: a fictive treatment variable and some categorically scaled variables by manipulating existing categorical variables and generating categorical variables from continuous ones. The result is a small example data set that exhibits a similar structure to the above described one and is provided at the following link: https://cloud.iwh-halle.de/index.php/s/flexpaneldid.[24] Under the same link, one can find the files `flexpaneldid_preprocessing.ado` and `flexpaneldid.ado` as well as the respective help files. Before starting work with the toolbox, one also has to install or update the Stata ado-files `psmatch2, pstest` and `cem`, which are used by the toolbox.

The example panel data set consists of yearly information on uniquely identified firms that are characterized by categorical and continuously scaled variables. The observation period is 1972 to 1981. The treatment can occur within the first five years of observation, the start and duration can vary among the treated firms. Also multiple treatments are possible. For the treated firms, the treatment variable equals one during the whole treatment period and zero before and after the treatment; for non-treated firms it is always zero. In case of multiple treatments, the treatment variable equals one from the start of the first treatment until the end of the last treatment. Suppose now we want to estimate the causal effect of a certain treatment on the number of patents at the firm

[23] The data set can be found at http://www.stata.com/data/jwooldridge/eacsap/patent.dta.
[24] The data generation is presented in the appendix.

level. The `flexpaneldid` toolbox offers different opportunities to do this.

6.1 Example: Preprocessing

First of all, we run `flexpaneldid_preprocessing`. Compulsory details are the individual identification of the observed units, `id(cusip)`, the definition of the treatment variable, `treatment(treatment)` and the variable identifying the time units in the panel, in our sample `time(year)`. Besides yearly data, also e.g. monthly or quarterly data can be the basis for estimations with the toolbox. The next necessary information are the matching variables. Here, all variables that will be potentially interesting for matching should be included, since the data preprocessing can be the basis for more than one run of the `flexpaneldid`. We also define the matching time in relation to the treatment start. For the example we choose `matchtimerel(-1)`, meaning that the (individual) matching time for treated firms is one year before the individual treatment starts.[25] Next, we define characteristics that are used for exact matching, in our example `matchvarsexact(sic_cat)`.[26] Although it is optional to save the preprocessing data, we recommend to use this option, because `flexpaneldid` can then repeatedly be applied without running the preprocessing again and again. Summing up all the information, the command looks like this:

```
flexpaneldid_preprocessing, id(cusip) treatment(treatment) time(year)
  matchvars(employ stckpr rnd sales return pats_cat rndstck_cat rndeflt_cat)
  matchtimerel(-1) matchvarsexact(sic_cat)
  prepdataset('preprocessed_data.dta') replace
```

After having submitted all the required information, we get the following output. The first part consists of a summary of all our submitted information and selected options:

```
*******************************************************************************
************************ flexpaneldid - preprocessing ************************
*******************************************************************************
-------------------------------------------------------------------------------
id:                     cusip
treatment:              treatment
time:                   year
matchvars:              employ stckpr rnd sales return pats_cat rndstck_cat rndeflt_cat
matchvarsexact:         sic_cat
match_time:             -1
prepdataset:            preprocessed_data.dta
-------------------------------------------------------------------------------
```

[25] Also the relative matching time refers to the variable indicating the time in the panel, `time(year)`.

[26] Depending on the number of treated observations, potential controls and the size of the original data set, the run time of the data preprocessing may be rather long. In this case, it is advantageous to define as many variables for exact matching as possible. This will reduce the size of the individual selection groups and thus, the run time.

The second part gives some details on the preprocessing. For each of the preprocessing steps, a dot is displayed to show the preprocessing progress. After the preprocessing is finished, we get a summary on the initial number of the treated observations (in our case 61), the number of observations that are dropped, because the preprocessing does not find any potential controls (in our case 0), the number of successfully assigned selection groups (61), and the average number of potential controls for the treated (or in other words, the mean size of the selection groups, here 45.7).

```
*******************************************************************************
*********************** Preprocessing *****************************************
*******************************************************************************
Preprocessing of 61 treated:
.................................................. 50
........... 61
*******************************************************************************
************************ Preprocessing - Summary ******************************
*******************************************************************************
Number of treated:                                    61
Number of treated dropped during preprocessing:        0
Number of treated after preprocessing:                61
Mean size of selection groups:                   45.7377
```

Based on the stored preprocessed data set, we can now use the `flexpaneldid` to estimate the treatment effect for the treated with different matching approaches. We are interested in the effect of a certain treatment on the number of patents.

6.2 Example: Nearest neighbor matching based on the statistical distance function including quality tests

Before we continue, we must reload the original data:

```
use example_data.dta, clear.
```

In the first example we want to run an estimation based on the flexible conditional DID approach described in section 4. Thus, we select option `statmatching` and distinguish between continuous and categorical matching variables: `statmatching(con(`*`employ stckpr rnd sales`*`) cat(`*`pats_cat rndstck_cat`*`))`. We will also include the pre-treatment development of the number of patents from two years to one year before treatment into the matching process, therefore we define `outcomedev(`*`-2 -1`*`)`.[27] The outcome we want to observe is the change in the number of patents from treatment start to three years afterwards. For the example we chose `outcometimerelstart(`*`3`*`)`. Furthermore, we want to get displayed the results of the quality tests for matching. Finally, we specify the path where the preprocessing data is stored. Summing up all the information, the command looks like this:

[27] It is important to note that the observation time of the pre-treatment outcome is defined relative to the treatment start and refers to the variable indicating the time in the panel, in our example `time(year)`.

```
flexpaneldid patents, id(cusip) treatment(treatment) time(year)
   statmatching(con(employ stckpr rnd sales) cat(pats_cat rndstck_cat))
   outcometimerelstart(3) outcomedev(-2 -1) test
   prepdataset('preprocessed_data.dta')
```

The output for the first example is as follows. At first, we get again a summary of our inputs. Second, we see a short summary for the executed matching procedure. In our example, for 47 out of the 61 treated units, the matching procedure finds a partner. In the matching process, 39 non-treated units are used as partners. That means, some of the non-treated units are used as partner for more than one treated, which is typical for the implemented nearest neighbor matching with replacement.

```
*******************************************************************************
************************ flexpaneldid *****************************************
*******************************************************************************
_______________________________________________________________________________

outcome:                patents
id:                     cusip
treatment:              treatment
time:                   year
outcome_time_start:     3
outcome_time_end:       .
outcome_dev:            -2 -1
cemmatching:
statmatcing:            , con(employ stckpr rnd sales) cat(pats_cat rndstck_cat)
test:                   test
outcomemissing:
didmodel:
_______________________________________________________________________________

*******************************************************************************
******************** Matching: STAT *******************************************
*******************************************************************************

*******************************************************************************
*************** flexpaneldid - Matching Summary *******************************
*******************************************************************************
                 |          NT           T
-----------------+------------------------------
All              |         165          61
Matched sample   |          39          47
```

As we selected the test option, the results of some quality checks for the matched groups are displayed. The tests are made at matching time, in our example one year before the treatment starts. First, the tests provided in the Stata command `pstest` by Leuven and Sianesi (2003) are conducted. For each of the matching variables, we find the means in the treated and in the control group, a measure for the standardized percentage difference – or bias – between the means in both groups, and a test if the means in the control group equal the ones in the treated group. Additionally we get an information on the similarity of the variances in the treated and the control group.

We would conclude that the means of all the matching variables are balanced, but the variances of `stckpr`, `sales` and `outcome_dev` are not.

```
*******************************************************************************
******************************** ps-test **************************************
*******************************************************************************
```

Variable	Mean Treated	Mean Control	%bias	t-test t	t-test p>|t|	V(T)/ V(C)
employ	26.197	20.337	8.6	0.42	0.677	1.13
stckpr	25.779	18.28	23.1	1.12	0.266	6.36*
rnd	38.116	31.329	4.6	0.23	0.822	1.30
sales	1201	1054.5	3.9	0.19	0.849	0.56*
pats_cat	1.9362	1.8936	3.2	0.15	0.879	0.98
rndstck_cat	3.4255	3.2979	5.2	0.25	0.802	1.05
outcome_dev	-1.617	-1.5957	-0.2	-0.01	0.991	1.81*

`* if variance ratio outside [0.56; 1.80]`

Ps R2	LR chi2	p>chi2	MeanBias	MedBias	B	R	%Var
0.037	4.77	0.688	7.0	4.6	44.2*	2.27*	43

`* if B>25%, R outside [0.5; 2]`

In case of matching based on the statistical distance function, additional scale-specific test statistics for the included variables are displayed. For all continuous variables, the results of a Kolmogorov-Smirnov test are presented, for the categorical variables, the results of chi-square tests are available. For reasons of space, we only present here the tests for the pre-treatment outcome development and the variables `employ` (continuous) and `pats_cat` (categorical) as an example. In the three displayed cases, the tests indicate no significant differences in the variable distributions between the treated and the control group. In case of the KS test, the corrected p-values of 0.28 for employment and 0.177 for the pre-treatment outcome development tell us that the variable distributions between the treated and the control group are not significantly different. In case of the χ^2 test for pats_cat, the p-value of 0.982 – and also a look at the respective number of observations in the categories – indicate balanced samples.

```
*******************************************************************************
******************************** KS-Test **************************************
*******************************************************************************
ksmirnov employ , by(treated)
Two-sample Kolmogorov-Smirnov test for equality of distribution functions
 Smaller group       D       P-value  Corrected
 ---------------------------------------------
 0:                   0.1915    0.178
 1:                  -0.0638    0.826
 Combined K-S:        0.1915    0.355      0.280
Note: Ties exist in combined dataset;
      there are 88 unique values out of 94 observations.

...
(Output for stckpr, rnd,  sales are omitted.)

ksmirnov outcome_dev , by(treated)
Two-sample Kolmogorov-Smirnov test for equality of distribution functions
 Smaller group       D       P-value  Corrected
 ---------------------------------------------
 0:                   0.0851    0.711
 1:                  -0.2128    0.119
 Combined K-S:        0.2128    0.238      0.177
Note: Ties exist in combined dataset;
      there are 27 unique values out of 94 observations.

*******************************************************************************
******************************** Chi2-Test ************************************
*******************************************************************************
tabulate pats_cat treated, chi2
  pats_cat |
      (at |
treatment |         treated
 time -1) |         0          1 |     Total
----------+----------------------+----------
        0 |         8          8 |        16
        1 |        14         12 |        26
        2 |         7          9 |        16
        3 |        11         11 |        22
        4 |         7          7 |        14
----------+----------------------+----------
    Total |        47         47 |        94
          Pearson chi2(4) =   0.4038   Pr = 0.982

...
(Output for rndstck_cat omitted.)
```

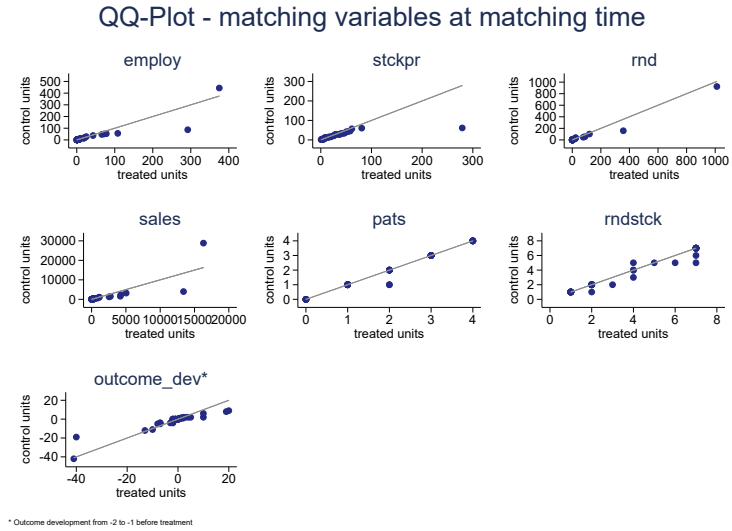

Figure 3: Quantile-quantile plots of the continuous matching variables, example 1

The quantile-quantile plots of the continuous matching variables give a graphical impression on the comparability of the matched groups. They compare the distributions in both groups by means of the plotted quantiles. The 45°-line represents identical distributions. From the figure 3 we see small deviations from the 45°-line for all displayed variables, mostly at the tails of the distributions.

In the last step, the estimation result for average treatment effect for the treated is displayed. In the display, all relevant information are summarized: the type of the estimator (in our example the nearest neighbor matching), the distance metric (in our example the statistical distance function), the number of the treated observations and unique controls included in the estimation (in the example 47 treated and 39 controls) and the mean number of matches per treated (in the example, not surprisingly, one). In our example, we observe a negative development of the number of patents for the period from the start of the treatment until three years afterward, both for the treated (-10.02) and the controls (-7.82). The mean difference in the patents development between treated and controls is -2.20.[28] To assess the statistical significance of this difference, we look at the p-value of the modified t-test for corrected standard errors.[29] The p-vaue of 0.7193 indicates that the difference is not significant.

[28] We apply the regression-based bias correction of Abadie and Imbens (2006, 2011) to adjust the start and the end values of the outcome development.

[29] The correction follows the matching-based procedure of Abadie et al. (2004); Abadie and Imbens (2006, 2011) with the number of matches fixed to two (like the default setting in the `teffects nnmatch` comand in Stata).

One important hint: as the data base contains only an arbitrarily defined treatment, the estimation results have no meaningful interpretation. We present them for illustrative purpose only.

```
********************************************************************************
*************** Conditional Diff-in-Diff *************************************
********************************************************************************
Average treatment effect for the treated
Estimator       : Nearest neighbor          No. of treated obs       =        47
Distance metric : Statistical DF            No. of unique controls   =        39
                                            Mean no. of matches      =         1
```

Outcome	mean Diff		DiD*	AI robust S.E.	z	P>|z|
	treated	controls				
patents	-10.0213	-7.8176	-2.2037	6.0938	-0.3616	0.7193

```
* Consistent bias-corrected estimator as proposed in Abadie & Imbens (2006,2011).
```

6.3 Example: Radius matching based on the statistical distance function including robustness checks

Now we want to use a random matching based on the statistical distance function – with slightly different matching variables. Also the second example we can base on the data set created in the preprocessing.[30]

Before we continue, we must again load the original data:

```
use example_data.dta, clear.
```

Besides some changes in the considered matching variables, we use some different options. Now, the observation time for the outcome development is defined in relation to the end of the treatment, `outcometimerelend(2)` denotes that we compare the outcome development from the treatment start to two years after treatment is finished. Also in this example, we take the pre-treatment outcome into account for matching, now with the level value two years before treatment starts, `outcomedev(-2)`. In the example we include the robustness checks based on the DID model, `didmodel`. Additionally, we chose the option `outcomemissing`.[31]. This may reduce the sample size, but will produce the best possible matches in that the approach selects potential partners considering only the matching variables (and not checking if the outcome is observable at the defined time).

[30] This makes apparent that it is advantageous to include all variables in the preprossesing that might be useful for further analyses.

[31] Although the example data set is not large, we chose the option for illustrative reasons.

```
flexpaneldid patents, id(cusip) treatment(treatment) time(year)
   statmatching(con(employ stckpr return) cat(rndeflt_cat rndstck_cat)radius(0.1))
   outcometimerelend(2) outcomedev(-2) didmodel outcomemissing
   prepdataset('preprocessed_data.dta')
```

The output is as follows. First, we again see all the given inputs. The second part contains the matching summary. Now, only 29 out of the 61 treated remain in the matched sample. The number of controls in the matched sample is with 66 larger than in the first example. This may be the result of the radius matching.

```
*******************************************************************************
************************ flexpaneldid *************************************
*******************************************************************************
-------------------------------------------------------------------------------
outcome:                patents
id:                     cusip
treatment:              treatment
time:                   year
outcome_time_start:     .
outcome_time_end:       2
outcome_dev:            -2
cemmatching:
statmatcing:            , con(employ stckpr return) cat(rndeflt_cat rndstck_cat) radius(0.1)
test:
outcomemissing:         outcomemissing
didmodel:               didmodel
-------------------------------------------------------------------------------

*******************************************************************************
********************* Matching: STAT ******************************************
*******************************************************************************

*******************************************************************************
***************** flexpaneldid - Matching Summary *****************************
*******************************************************************************
                  |          NT           T
------------------+--------------------------
All               |         165          61
Matched sample    |          66          29
```

Next, the estimation output is displayed. The header summarizes all relevant information. Different from the first example, the average number of controls is 4.2. Also the observed development of the number of patents for treated and controls strongly differs from the first example – due to the different matching estimator, the different selection strategy, and different matching variables.

```
********************************************************************************
*************** Conditional Diff-in-Diff *************************************
********************************************************************************
Average treatment effect for the treated
Estimator       : Radius                    No. of treated obs       =        29
Distance metric : Statistical DF            No. of unique controls   =        66
                                            Mean no. of matches      =     4.207
```

Outcome	mean Diff		DiD*	AI robust S.E.	z	P>|z|
	treated	controls				
patents	-1.3793	-0.7089	-0.6705	2.0218	-0.3316	0.7426

```
* Consistent bias-corrected estimator as proposed in Abadie & Imbens (2006,2011).
```

Since we opted for robustness checks, the results of two different specifications of the standard two-way fixed effects model are returned.[32] The first one is the model in the 2x2 case (meaning for two groups and two times). Here, the outcome development between the treatment start the specified end time is compared.[33] The coefficient of interest is -2.23706, meaning that the development of the number of patents among the treated is by 2.24 less than in the control group. From the p-value of 0.162 we would conclude that this difference is not statistically significant.

[32] For the estimation we use the `xtreg, fe` command in Stata, including a constant and time dummies, but no further covariates. The included time dummies are defined according to the dimension of the `time` identifier. In our case, the time units are years, and `flexpaneldid` defines year dummies for the regression.

[33] Since the treatment start and duration can vary, we observe more than two years in the estimation output.

```
*******************************************************************************
*** Fixed Effects Diff-in-Diff - robustness check of Conditional Diff-in-Diff **
*******************************************************************************

xtreg patents i.treated##i.post_treat_dummy_rel_time i.year if post_treat_dummy_rel_time == 2
| first_treatment == year, fe vce(cluster panel_id)
note: 1.treated omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =       190
Group variable: panel_id                        Number of groups   =        95

R-sq:  within  = 0.0413                         Obs per group: min =         2
       between = 0.1197                                        avg =       2.0
       overall = 0.0821                                        max =         2

                                                F(7,94)            =      1.00
corr(u_i, Xb)  = 0.2331                         Prob > F           =    0.4378

                                                   (Std. Err. adjusted for 95 clusters in panel_id)
```

patents	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
1.treated	0	(omitted)				
2.post_treat_dummy_rel_time	1.79453	3.458088	0.52	0.605	-5.071585	8.660644
treated#post_treat_dummy_rel_time						
1 2	-2.23706	1.58742	-1.41	0.162	-5.388919	.9147984
year						
75	-1.619814	1.685874	-0.96	0.339	-4.967156	1.727529
76	-1.774941	2.844999	-0.62	0.534	-7.423754	3.873872
77	.0137887	3.506279	0.00	0.997	-6.94801	6.975587
78	-3.732286	4.074362	-0.92	0.362	-11.82203	4.357456
79	-1.230354	3.352732	-0.37	0.714	-7.887283	5.426574
_cons	9.861463	1.039451	9.49	0.000	7.797608	11.92532
sigma_u	19.442247					
sigma_e	5.5396632					
rho	.92491141	(fraction of variance due to u_i)				

The second model estimates the mean treatment effect for the treated within the classical two-way FE model for the time from the earliest treatment start to two years after the latest treatment end (since we defined `outcometimerelend(2)`). The estimated treatment effect is −0.49, but also not significant – as is indicated by the p-value of 0.610.

```
*******************************************************************************
*************** Fixed Effects Diff-in-Diff - mean treatment effect ************
*******************************************************************************
xtreg patents i.treated##i.post_treat_dummy i.year if post_treat_dummy_rel_time <= 2,
fe vce(cluster panel_id)
note: 1.treated omitted because of collinearity
Fixed-effects (within) regression               Number of obs      =       637
Group variable: panel_id                        Number of groups   =        95
R-sq:  within  = 0.0215                         Obs per group: min =         5
       between = 0.0003                                        avg =       6.7
       overall = 0.0018                                        max =         8
                                                F(9,94)            =      1.31
corr(u_i, Xb)  = 0.0116                         Prob > F           =    0.2409
                                              (Std. Err. adjusted for 95 clusters in panel_id)
```

patents	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
1.treated	0	(omitted)				
1.post_treat_dummy	.3824173	.8686678	0.44	0.661	-1.342343	2.107178
treated#post_treat_dummy						
1 1	-.4891359	.9567283	-0.51	0.610	-2.388742	1.410471
year						
73	-.9263158	.5472448	-1.69	0.094	-2.012883	.1602516
74	-.9263158	.5824373	-1.59	0.115	-2.082759	.2301271
75	-1.103654	.625641	-1.76	0.081	-2.345879	.1385714
76	-1.743825	.8542492	-2.04	0.044	-3.439957	-.0476936
77	-1.985287	1.374455	-1.44	0.152	-4.714301	.7437261
78	-2.305253	1.261525	-1.83	0.071	-4.81004	.1995341
79	-1.795019	1.382494	-1.30	0.197	-4.539994	.9499549
_cons	10.7776	.4336799	24.85	0.000	9.916523	11.63869
sigma_u	20.43782					
sigma_e	4.4415796					
rho	.95490128	(fraction of variance due to u_i)				

6.4 Example: CEM matching with equal numbers of treated and controls in the strata including quality tests

Also the last example we base on the stored preprocessing data set. We want to use a CEM matching with equal numbers of treated and controls within a strata and test the quality of the matching results. Before we specify the `flexpaneldid` command, we

must reload the original data:

```
use example_data.dta, clear.
```

Besides the compulsory inputs (*depvar, id, treatment, time*), we now select `cemmatching`. In this option, we can manually choose strata for the matching variables – either as a fixed number of equally spaced strata or by explicitely defining the cutpoints, e. g. `cemmatching(employ (#5) stckpr (100 200 300) rnd sales pats_cat(#0) rndstck_cat(#0))`. If no stratification is defined (as is the case for the outcome development, *rnd* and *sales*), the default number of strata (13) is used. For the categorical variables, we must retain the number of values: `pats_cat (#0) rndstck_cat (#0)`. See Blackwell et al. (2009) for a more detailed description.

The following command lines sum up all our selections:

```
flexpaneldid patents, id(cusip) treatment(treatment) time(year)
   cemmatching(employ (#5) stckpr (100 200 300) rnd sales ...
    pats_cat(#0) rndstck_cat(#0) k2k)
   outcometimerelend(2) outcomedev(-2) outcomemissing test
   prepdataset('preprocessed_data.dta')
```

The first part of the output again gives a summary of the submitted details. For reasons of space we omit this display. The output for `cemmatching` differs from the `statmatching` output in that some details of the matching procedure are displayed: We find the stratification cutpoints for each variable, an alternative matching summary (containing the number of generated strata, the number of matched strata, the number of matched and unmatched treated and control units (in terms of observations, not uniquely identified units)) and the multivariate imbalance as an aggregated quality measure as well as the variable specific imbalances. Below, we find again the summary on the number of matched treated and controls (in terms of uniquely identified units like in the above described examples).

```
********************************************************************************
******************** Matching: CEM *********************************************
********************************************************************************
Cutpoints:
selection_group: (user)
  0
employ: (user)
                1
     +-----------+
  1  | .0849999785 |
  2  | 118.6432178 |
  3  | 237.2014355 |
  4  | 355.7596533 |
  5  | 474.3178711 |
     +-----------+
```

```
stckpr: (user)
         1
  ┌──────┐
1 │  100 │
2 │  200 │
3 │  300 │
  └──────┘

rnd: (sturges)
                  1
  ┌──────────────┐
 1 │            0 │
 2 │  84.31013997 │
 3 │  168.6202799 │
 4 │  252.9304199 │
 5 │  337.2405599 │
 6 │  421.5506999 │
 7 │  505.8608398 │
 8 │  590.1709798 │
 9 │  674.4811198 │
10 │  758.7912598 │
11 │  843.1013997 │
12 │  927.4115397 │
13 │   1011.72168 │
  └──────────────┘

sales: (sturges)
                  1
  ┌──────────────┐
 1 │  1.221999168 │
 2 │  2404.420947 │
 3 │  4807.619895 │

    ...

13 │  28839.60938 │
  └──────────────┘

pats_cat: (user)
  0
rndstck_cat: (user)
  0
outcome_dev: (sturges)
                  1
  ┌──────────────┐
 1 │            0 │
 2 │  70.41666667 │
 3 │  140.8333333 │

   ...

13 │          845 │
  └──────────────┘

Matching Summary:
-----------------
Number of strata: 948
Number of matched strata: 30
              0     1
      All  2163    47
  Matched    30    30
Unmatched  2133    17

Multivariate L1 distance: .63333333
```

```
Univariate imbalance:
                     L1      mean       min       25%       50%       75%       max
selection_group       0         0         0         0         0         0         0
         employ      .1  -.85613     -.174      .309      .747     2.175   -24.305
         stckpr      .1     2.795     -.875     1.125      1.75      3.25     27.25
            rnd  .06667  -1.8587  -.11696  -.02291   .02262    .6408   -46.112
          sales      .1  -14.339   -5.087   -7.169  -.51901    65.294    -7.374
       pats_cat       0         0         0         0         0         0         0
    rndstck_cat       0         0         0         0         0         0         0
    outcome_dev  .13333   .66667         0         1         0        -2        18

********************************************************************************
**************** flexpaneldid - Matching Summary ******************************
********************************************************************************
                 |          NT           T
-----------------+------------------------
All              |         165          61
Matched sample   |          26          30
```

Because we selected the `test` option, also the information of *pstest* and the quantile-quantile plots are displayed.

```
********************************************************************************
********************************* ps-test **************************************
********************************************************************************

-------------------------------------------------------------------------------
                         |        Mean                     |      t-test      |  V(T)/
Variable                 | Treated Control      %bias      |    t    p>|t|    |  V(C)
-------------------------+---------------------------------+------------------+-------
employ                   |  3.8005  4.6567     -11.3       |  -0.44  0.663    |  0.25*
stckpr                   |  12.761  9.9665      21.9       |   0.85  0.399    |  2.18*
rnd                      |  1.6609  3.5196     -25.3       |  -0.98  0.331    |  0.04*
sales                    |  147.55  161.89      -5.7       |  -0.22  0.826    |  0.60
pats_cat                 |     1.4     1.4       0.0       |   0.00  1.000    |  1.00
rndstck_cat              |  2.2667  2.2667      0.0        |  -0.00  1.000    |  1.00
outcome_dev              |  5.3667     4.7      8.7        |   0.34  0.737    |  2.26*
-------------------------------------------------------------------------------
* if variance ratio outside [0.48; 2.10]

-----------------------------------------------------------------------------
Ps R2   LR chi2   p>chi2   MeanBias   MedBias      B       R      %Var
-----------------------------------------------------------------------------
0.048      3.98    0.782      10.4       8.7      38.2*   0.08*     57
-----------------------------------------------------------------------------
* if B>25%, R outside [0.5; 2]
(30 observations deleted)
```

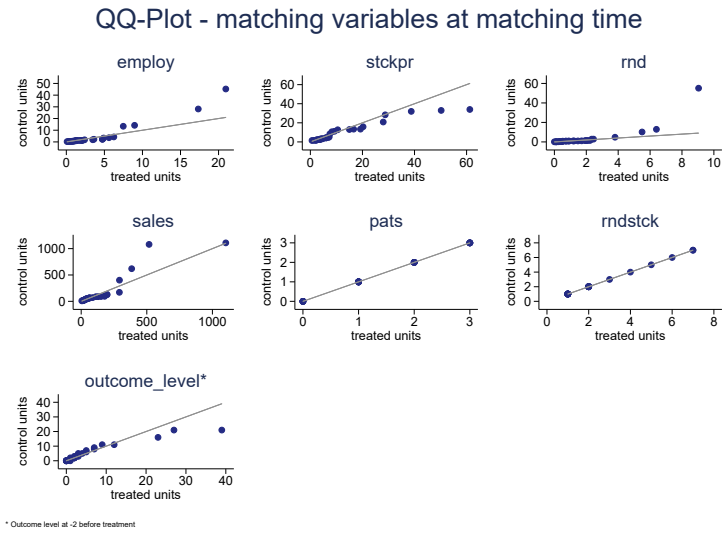

Figure 4: Quantile-quantile plots of the continuous matching variables, example 3

Also for the selected cemmatching, we get a summary of the estimation results. The output structure is comparable to the above described examples of the `statmatching` results. Different from the estimation approach applying the statistical distance function, the bias correction procedure of Abadie and Imbens (2006, 2011) is not applied.

```
*******************************************************************************
*************** Conditional Diff-in-Diff ************************************
*******************************************************************************
Average treatment effect for the treated
Estimator       : k2k                          No. of treated obs       =        30
Distance metric : CEM                          No. of unique controls   =        26
                                               Mean no. of matches      =         1
```

Outcome	mean Diff		DiD	S.E	z	P>|z|
	treated	controls				
patents	-1.6000	-0.7333	-0.8667	0.9634	-0.8996	0.3757

7 References

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

Abadie, A., Dukker, D., Leber-Herr, J. and Imbens, G. W. (2004), 'Implementing matching estimators for average treatment effects in Stata', *The Stata Journal* **4**(3), 290–311.

Abadie, A. and Imbens, G. W. (2006), 'Large Sample Properties of Matching Estimators for Average Treatment Effects', *Econometrica* **74**(1), 235–267.

Abadie, A. and Imbens, G. W. (2011), 'Bias-Corrected Matching Estimators for Average Treatment Effects', *Journal of Business and Economic Statistics* **29**(1), 1–11.

Abraham, S. and Sun, L. (2019), Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects, working paper, Massachusetts Institute of Technology (MIT).

Angrist, J. D. and Pischke, J.-S. (2009), *Mostly Harmless Econometrics*, Princeton University Press, Princeton, New Jersey.

Ashenfelter, O. (1978), 'Estimating the Effect of Training Programs on Earnings', *Review of Economics and Statistics* **60**(1), 47–57.

Athey, S. and Imbens, G. W. (2018), Design-based Analysis in Difference-In-Differences Settings with Staggerd Adoption, Working paper, Stanford University.

Autor, D. H. (2003), 'Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing.', *Journal of Labor Economics* **21**(1), 1–42.

Bergemann, A., Fitzenberger, B. and Speckesser, S. (2009), 'Evaluating the dynamic employment effects of training programs in East Germany using conditional difference-in-differences', *Journal of Applied Econometrics* **24**(5), 797–823.

Bernini, C. and Pellegrini, G. (2011), 'How are growth and productivity in private firms affected by public subsidy? Evidence from a regional policy', *Regional Science and Urban Economics* **41**, 253–265.

Blackwell, M., Iacus, s., King, G. and Porro, G. (2009), 'cem: C oarsened exact matching in stata', *The Stata Journal* **9**(4), 524–546.

Blundell, R. and Costa Dias, M. (2009), 'Alternative Approaches to Evaluation in Empirical Microeconomics', *The Journal of Human Resources* **44**(3), 565–640.

Borusyak, K. and Jaravel, X. (2017), Revisiting Event Study Designs, with an Application to the Estimation of the Marginal Propensity to Consume, working paper, Harvard University.

Bronzini, R. and de Blasio, G. (2006), 'Evaluating the impact of investment incentives: The case of Italy's Law 488/1992', *Journal of Urban Economics* **60**, 327–349.

Caliendo, M. and Künn, S. (2011), 'Start-up subsidies for the unemployed: Long-term evidence and effect heterogeneity', *Journal of Public Economics* **95**, 311–331.

Callaway, B. and Sant'Anna, P. H. C. (2019), Difference-in-Differences with Multiple Time Periods, Detu working paper, Temple University.

Daw, J. R. and Hatfield, L. A. (2018), 'Matching and Regression to the Mean in Difference-in-Differences Analysis', *Health Services Research* **53**(6), 4111–4117.

de Chaisemartin, C. and D'Haultfoeuille, X. (2019), Two-way fixed effects estimators with heterogeneous treatment effects, Working paper, CREST-ENSAE.

Dehejia, R. (2005), 'Practial Propensity Score Matching: A Reply to Smith and Todd', *Journal of Econometrics* **125**, 355–364.

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.

Deshpande, M. and Li, Y. (2017), Who is sreened out? application cost and the targeting of disability programs, working paper 23472, National Bureau of Economic Research (NBER).

Dettmann, E., Becker, C. and Schmeißer, C. (2011), 'Distance functions for matching in small samples', *Computational Statistics and Data Analysis* **55**(5), 1942–1960.

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

Fadlon, I. and Nielsen, T. H. (2017), Family labor supply responses to severe health shocks, working paper 21352, National Bureau of Economic Research (NBER).

Freier, R., Schumann, M. and Siedler, T. (2015), 'The earnings returns to graduating with honors — Evidence from law graduates', *Labour Economics* **34**, 39–50.

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

Goodman-Bacon, A. (2019), Difference-in-Differences with Variation in Treatment Timing, working paper 25018, National Bureau of Economic Research (NBER).

Greenaway, D., Gullstrand, J. and Kneller, R. (2005), 'Exporting May Not Always Boost Firm Productivity', *Review of World Economics* **141**(4), 561–582.

Hainmueller, J. and Xu, Y. (2013), 'ebalance: a Stata package for entropy balancing', *Journal of Statistical Software* **54**(7), 1–18.

Ham, J. C., Swenson, C., İmrohoroğlu, A. and Song, H. (2011), 'Government programs can improve local labor markets: Evidence from State Enterprise Zones, Federal Empowerment Zones and Federal Enterprise Community', *Journal of Public Economics* **95**, 779–797.

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., 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.

Heyman, F., Sjöholm, F. and Tingvall, P. G. (2007), 'Is there really a foreign ownership wage premium? Evidence from matched employer–employee data', *Journal of International Economics* **73**, 355–376.

Ho, D. E., Imai, K., King, G. and Stuart, E. A. (2007), 'Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference', *Political Analysis* **15**, 199–236.

Imai, K. and Kim, I. S. (2019*a*), On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data, Working paper, Harvard University.

Imai, K. and Kim, I. S. (2019*b*), 'When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?', *American Journal of Political Science* **63**(2), 467–490.

Imai, K., Kim, I. S. and Wang, E. (2019), Matching Methods for Causal Inference with Time-Series Cross-Sectional Data, Working paper, Harvard University.

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

Jacobson, L. S., LaLonde, R. J. and Sullivan, D. G. (1993), 'Earnings Losses of Displaced Workers', *The American Economic Review* **83**(4), 685–709.

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.

King, G. and Nielsen, R. (2016), Why Propensity Scores Should Not Be Used for Matching, Working paper, Harvard University.

King, G. and Zeng, L. (2006), 'The Dangers of Extreme Counterfactuals', *Political Analysis* **14**, 131–159.

Leuven, E. and Sianesi, B. (2003), PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing, Technical report, University of Oslo. access: 2016-05-06.
URL: *http://ideas.repec.org/c/boc/bocode/s432001.html*

Marcus, J. and Siedler, T. (2015), 'Reducing binge drinking? The effect of a ban on late-night off-premise alcohol sales on alcohol-related hospital stays in Germany', *Journal of Public Economics* **123**, 55–77.

Neumark, D. and Kolko, J. (2010), 'Do enterprise zones create jobs? Evidence from California's enterprise zone program', *Journal of Urban Economics* **68**, 1–19.

Opitz, O. (1980), *Numerische Taxonomie*, Fischer-Verlag, Stuttgart, New York.

Pellegrini, G. and Centra, M. (2006), Growth and efficiency in subsidized firms, *in* 'Workshop 'The Evaluation of Labour Market, Welfare and Firms Incentive Programmes", Istituto Veneto di Scienze, Lettere ed Arti, Venezia.

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

Smith, J. A. and Todd, P. E. (2005*b*), 'Does matching overcome LaLonde's critique of nonexperimental estimators? Rejoinder', *Journal of Econometrics* **125**(1-2), 365–75.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2nd edn, MIT Press, Cambridge, Massachusetts.

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.

Appendix

Generation of the example data set

We start with a publicly available data set with a similar structure to the one described in section 2 (see figure 1).

```
. use http://www.stata.com/data/jwooldridge/eacsap/patent.dta, clear
```

Based on this, we generate some additional variables. First, we add a fictive treatment variable that can occur within the first five years of the observation period. If treatment equals zero, this indicates a non-treated unit, $treatment = 1$ marks the duration of the fictive treatment of the treated units.

```
. set seed 13
. gen random = runiform()
. sort random
. gen treatment = 0
. replace treatment = 1 if random>=0.95 & year>=73 & year<=77
. sort cusip year
. replace treatment=1 if random<=0.5 & treatment[_n-1]==1 & year>=73 & year<=77
>  & cusip[_n-1]==cusip
. lab var treatment "treatment in 73 to 77"
. drop random
. order cusip year treatment
```

Additionally, we generate some categorical variables, i. e. we manipulate some of the existing categorical variables and generate new categorical variables from continuous ones.

```
. by cusip: egen merger_cat=max(merger)
. gen sic_cat=2000 if sic>=2000 & sic<2300
. replace sic_cat=2300 if sic>=2300 & sic<2600
. replace sic_cat=2600 if sic>=2600 & sic<2900
. replace sic_cat=2900 if sic>=2900 & sic<3200
. replace sic_cat=3200 if sic>=3200 & sic<3500
. replace sic_cat=3500 if sic>=3500 & sic<3800
. replace sic_cat=3800 if sic>=3800
. lab var sic_cat "sector categories"
. gen pats_cat=0 if patentsg==0
. replace pats_cat=1 if patentsg>=1 & patentsg<=3
. replace pats_cat=2 if patentsg>=4 & patentsg<=9
. replace pats_cat=3 if patentsg>=10 & patentsg<=50
. replace pats_cat=4 if patentsg>=51
. lab var pats_cat "patents categories"
. gen rndstck_cat=0 if rndstck==.
. replace rndstck_cat=1 if rndstck>0 & rndstck<=5
. replace rndstck_cat=2 if rndstck>5 & rndstck<=10
```

```
. replace rndstck_cat=3 if rndstck>10 & rndstck<=15
. replace rndstck_cat=4 if rndstck>15 & rndstck<=20
. replace rndstck_cat=5 if rndstck>20 & rndstck<=40
. replace rndstck_cat=6 if rndstck>40 & rndstck<=60
. replace rndstck_cat=7 if rndstck>60
. lab var rndstck_cat "RnDstock categories"
. gen rndeflt_cat=0 if rndeflt==0
. replace rndeflt_cat=1 if rndeflt>0 & rndeflt<=0.5
. replace rndeflt_cat=2 if rndeflt>0.5 & rndeflt<=1
. replace rndeflt_cat=3 if rndeflt>1 & rndeflt<=5
. replace rndeflt_cat=4 if rndeflt>5 & rndeflt<=10
. replace rndeflt_cat=5 if rndeflt>10
. lab var rndeflt_cat "RnDexpenditures categories"
```

The result is the dataset 'example_data.dta' provided at our homepage (https://cloud.iwh-halle.de/index.php/s/flexpaneldid).

Halle Institute for Economic Research –
Member of the Leibniz Association

Kleine Maerkerstrasse 8

D-06108 Halle (Saale), Germany

Postal Adress: P.O. Box 11 03 61

D-06017 Halle (Saale), Germany

Tel +49 345 7753 60

Fax +49 345 7753 820

www.iwh-halle.de

ISSN 2194-2188