Gabriel M. Ahlfeldt

Weights to address non-parallel trends in panel difference-in-differences models

Article (Accepted version) (Refereed)

Gabriel M. Ahlfeldt♣

Weights to address non-parallel trends in panel difference-in-differences models*

Abstract: Causal inference using the difference-in-differences (DD) method relies on the untestable assumption of parallel counterfactual trends across units that are assigned to different treatments. To facilitate the application of the method in settings where the parallel-trends assumption is seemingly violated, I suggest weighting observations such that the conditional correlations between treatments and pre-treatment outcome trends are minimised, i.e. weighted trends are parallel. I evaluate the performance of a weighted parallel trends (WPT) DD estimator in a Monte Carlo study and provide an application to a case-study context in which a benchmark estimate exists. The WPT DD approach can be applied in settings with multiple continuous treatment variables as well as to estimating time-varying treatment effects.

Keywords: Access, difference-in-differences, land price, matching, Monte Carlo, noise, weights
Version: April, 2018
JEL: R12, R14, R41, N73, N74

1 Introduction

Difference-in-differences (DD) analysis (Ashenfelter and Card, 1985) has become a standard programme evaluation technique due to its potential to control for unobserved individual effects and time effects. The effect of a programme, a policy, or another exogenous event usually referred to as a treatment, on an outcome is identified from a comparison of subjects with different exposures to a programme (first difference) before and after (second difference) the programme is implemented. The key identifying assumption in this comparison is that of parallel counterfactual trends. In the simplest case of a binary treatment (either treated or not treated), the treated and non-treated (control) subjects are assumed to follow the same outcome trend in the absence of a treatment. This

♣ London School of Economics and Political Sciences (LSE) & Centre for Economic Policy Research (CEPR). g.ahlfeldt@lse.ac.uk, www.ahlfeldt.com

* I thank seminar participants of the CESIfo Economic Studies Conference "On the Use of Geocoded Data in Economic Research" for comments and suggestions. Ina Blind, Matz Dahlberg, and Gustav Engstrom (the editors) and four anonymous referees have provided much appreciated feedback. I also thank seminar and conference participants in Berlin (Humboldt), Barcelona (UEA), Heidelberg, Miami (UEA), Orléans, Toulouse (SEA), Vaduz (VfS), Wuerzburg, and Zurich (KOF) and especially Kristoffer Möller, Jos van Ommeren, Michael Pflüger, and Sevrin Waights who have commented on closely related work. Kristoffer Möller and Sevrin Waights provided outstanding help with the compilation of the data set. Sascha Möbius and Neele Reimann-Phillip also provided excellent research assistance. The German Science Foundation (DFG project code NI 938/2-1) and the Fritz-Thyssen-Foundation (project code AZ.10.10.2.070) are acknowledged for financial support. The usual disclaimers apply.

assumption, however, is not only ambitious it is also not testable because the counterfactual cannot be directly observed. Arguably, the closest approximation of the counterfactual trend is the "pre-trend", the trend observed before the effects of a treatment can be anticipated.

This paper is a companion paper to Ahlfeldt et al. (2016) who estimate the land price capitalisation effects of the metro rail system in Berlin today and a century ago, using different identification strategies. At one stage of the analysis, the authors conduct a DD-based intervention study of the land price effects of Line A, the first electrified metro line in Germany, which opened in 1902 in Berlin. They are interested in separately identifying the presumably positive effects of being close to a station and the presumably negative effects that originate from the noise of the elevated rail line. They wish to identify the temporal structure of the treatment effects, i.e. treatment effects before the opening due to anticipation effects and temporal trends in the treatment effects after the opening due to a presumably non-instantaneous adjustment to a new equilibrium. Thus, they face the challenge of conducting an intervention study in a setting with multiple continuous treatments whose effects are continuous in space and time.

In this empirical setting, they encounter that the treatments (noise and proximity to a station) are strongly correlated with pre-trends in land prices, creating an identification problem due to the likely violation of the parallel-trends assumption. A weighted DD is suggested to address the problem. The idea of the estimator is to use time-invariant weights to reweight land parcels in such a way that the correlation between noise and station distance on the one hand (the treatment variables) and the pre-treatment trend in land price (the pre-trend) on the other, is minimised. It is argued that if this objective is achieved in a targeted pre-treatment period, as well as in other (non-targeted) pre-treatment periods that have not been used in the weights construction (an overidentification test), the correlation will likely also be mitigated in the remaining (post-treatments) periods. As a result, the parallel-trend assumption required for the identification of the treatment effects is more plausible.

Since in this approach the key assumption is that (counterfactual) outcome trends across subjects exposed to different treatment intensities are parallel, conditional on weighting, I refer to the estimator as weighted-parallel-trends (WPT) DD in the remainder of this paper. I complement the Ahlfeldt et al. (2016) case study application of WPT DD in two respects. First, I introduce the identification problem in the context of a Monte Carlo study, in which I evaluate the WPT DD performance. Second, I expand on the case study by comparing OLS DD and WPT DD and exploring the sensitivity of the WPT DD results to changes in covariates, objective functions, and algorithms used

in the parallel trends weights construction. The Monte Carlo results suggest that the WPT DD has the potential to reduce OLS bias to the extent that the objective of minimising treatment-trend correlations (over a pre-treatment period) is achieved. In the case study application, the WPT DD provides results that are more plausible than the OLS results given theoretical expectations and existing evidence. In particular, the noise effects estimated by WPT DD, unlike those from OLS DD, are close to estimates that exploit plausibly exogenous variation from a spatial discontinuity in noise at a tunnel entrance (Ahlfeldt et al., 2016). Reassuringly, different implementations of the WPT DD yield similar results.

This paper contributes to a growing literature concerned with violations of the non-parallel trends assumption in DD where the idea of using weights in DD is not new. In a closely related theoretical paper, Abadie (2005) discusses how a semi-parametric DD estimator can be used to address the identification problem that arises if selection into treatment is correlated with individual trends. Heckman et al. (1998), Heckman et al. (1997), Smith and Todd (2005), Kline (2011), and Hainmueller (2012) also discuss estimators that rely on the reweighting of observations to improve balancing conditions. Such weighting approaches differ from alternative approaches to controlling for time-varying confounding factors that rely on time-differencing (Lee, 2015), controlling for treatment-trend interactions (Ahlfeldt, Moeller, et al., 2017), or interactive fixed effects (Bai, 2003; Gobillon and Magnac, 2016). Other relevant papers devoted to studying and relaxing the standard DD identification restrictions include, Meyer (1995), Angrist and Krueger (1999), Blundell and Macurdy (1999), Besley and Case (2000), Blundell et al. (2004), and Athey and Imbens (2006).[1] In terms of the approach to evaluating the WPT DD, this paper is related to a literature using Monte Carlo simulations to establish small sample properties of estimators (see e.g. Conley and Taber, 2011). The approach to benchmarking DD results against more local estimates that exploit plausibly exogenous variation is similar to Kline (2011) and Ahlfeldt, Koutroumpis, et al. (2017).

In general, the causal inference literature has by now made considerable progress in developing estimators with well-developed properties that solve the fundamental identification problem conditional on some restrictions. There are also several user-written programmes that help applied researchers in taking these estimators to the data (e.g. Gomez, 2015; Guardabascio and Ventura, 2013; Hainmueller et al., 2011; Hainmueller and Xu, 2013; Leuven and Sianesi., 2003). However,

[1] For a recent survey of the literature see, Lee (2016).

empirical challenges are manifold, so sometimes the standard tools may not fit the needs in a specific case study.

The Line A studied by Ahlfeldt et al. (2016) serves as an example. Since the empirical problem the authors face is that of a correlation between a treatment and a time trend in the outcome, a natural consideration would be to control for the confounding trend. The generalised DD suggested by Lee (2015) addresses the presence of confounding trends by means of time differencing. Triple differences (DD in first time difference) allow arbitrary linear trends to be controlled for, while quadruple differences accommodate quadratic trends, etc. In the present case, a judgement on the appropriate order of the trend, however, is difficult because even the direction of the treatment-trend correlation changes already from the first period to the second. For the same reason, the inclusion of parametric treatment-trend interactions is ambitious. The interactive fixed effects estimator uses an algorithm to identify interactions of to-be-identified individual effects (a factor loading) and time effects (a factor) to control for confounding trends. However, in the presence of a treatment effect that potentially builds up over much of the treatment period, the separate identification of time-varying treatment effects on the one hand, and factor loadings that interact with time effects on the other, represents a tough challenge.

Empirical approaches that rely on reweighting represent a seemingly attractive alternative. Examples include the inverse probability weighting (Hernán et al., 2001) and the special case of entropy balancing (Hainmueller, 2012), the propensity score matching (Rosenbaum and Rubin, 1983), or the synthetic control method (Abadie and Gardeazabal, 2003). The main problem with the application of these tools to the present case is that they serve the purpose of evaluating singular treatments and not multiple correlated treatments. Moreover, it is not straightforward to apply them to settings with continuous treatment variables and intervention study designs with time-varying treatment effects (as e.g. in Autor, 2003).

The focus of the literature on singular binary treatments suggests that the challenges faced by Ahlfeldt et al. (2016) are quite specific to their case study. However, countervailing externalities can be expected not only for infrastructures such as highways or airports, but also public facilities such as police and fire stations. Multiple correlated treatments can also occur in completely different contexts, e.g. if individuals are supported by multiple social programmes (job training, education programmes, or housing improvements).

The purpose of this paper is to illustrate how in a relatively simple and transparent approach, time-invariant weights that minimise the treatment-trend correlation in the pre-period help to address non-parallel trends in a setting where the literature, as far as I know, does not offer off-the-shelf solutions. I am largely agnostic about the computation of the weights used in WPT DD, an aspect that is often at the centre of methodological contributions. At this stage, I do not wish to promote any of the algorithms used in this paper (or in the companion paper). It is easy enough to evaluate whether an algorithm achieves its purpose of minimising the treatment-trend correlation in a given period. And it is also straightforward to overidentify the result using another period that has not been used in the weights construction. The researcher can also easily check the selectivity of the weighted sample to avoid an external validity problem if heterogeneity in the treatment effect is expected. As long as these tests are passed, it is, in principle, up to the researcher's creativity to develop a weighting strategy that works in an empirical setting. Of course, this does not preclude making use of existing computational procedures. As an example, Hainmueller (2012) notes that entropy balancing weights could be paired with standard estimators.

The remainder of the paper is split into the Monte Carlo study in section 2 and the case-study application in section 3. The final section 4 concludes.

2 Monte Carlo study

To illustrate the empirical context and shed some light on the properties of the WPT DD, it is useful to apply the WPT DD to a setting in which the true effect to be estimated is known. Therefore, I conduct a motivating simulation study. A replication directory is available at the Harvard Dataverse.[2]

2.1 Data-generating process

Consider a data-generating process (DGP) of the following form:

$$y_{it} = \sum_n \alpha_n P_t D_i^n + \mu_i + \varphi_t + \omega_i f(t) + \epsilon_{it}, \quad (1)$$

where y_{it} is an outcome observed for a unit i = 1,...,I at period t = 1,...,T, D_i^n is one of n = 1,...,N treatment variables, and $P_t = P(t \geq z)$ is an indicator variable taking the value of one for all peri-

[2] Published at : https://doi.org/10.7910/DVN/7FJEE1

ods $t \geq z$ in which the treatment is in place, and zero otherwise. α_n is one of N difference-in-differences parameters to be estimated. In the application introduced in Section 3, y_{it} corresponds to parcel land prices observed in different years, $D_i^{n=1}$ and $D_i^{n=2}$ correspond to distance from the nearest station and exposure to noise, and P_t indicates all years when Line A is in operation. μ_i and φ_t are individual and time fixed effects and, critical for the point being made, there is an interaction of an unobserved individual trend effect ω_i (a factor loading) with a time trend $f(t)$ (a factor), with $f'(t) \neq 0$.

For the $\omega_i f(t)$ interaction to represent a threat to identification of α_n, ω_i must be correlated with D_i^n. I create an imperfect correlation between ω_i and D_i^n by choosing the following functional form:

$$\omega_i = \vartheta_i + \frac{1}{N} H_i \sum\nolimits_n D_i^n, \tag{2}$$

where H_i is a to-be-specified variable that is not observed by the researcher and ϑ_i is a random variable. From an estimation point of view, the problem is that the difference between the marginal effect of D_i^n in the before and in the after period, which is identified by the standard panel DD regression omitting a control for $\omega_i f(t)$, is contaminated by the general time trend:

$$E\left(\frac{\partial y_{i,t\geq z}}{\partial D_i^n} - \frac{\partial y_{i,t<z}}{\partial D_i^n}\right) = \alpha_n + \frac{1}{N} E(H_i)\left(\bar{f}^{,t\geq z} - \bar{f}^{,t<z}\right) \tag{3}$$

where a bar indicates the mean value within the pre-treatment ($t < z$) or the post-treatment ($t \geq z$) period. This is an instance of the well-known non-parallel trends problem. Weighted estimators are generally thought to remove bias due to non-random sampling (Shaun and Ian, 2011). To motivate the WPT approach, I build on this idea and first generate a population j = 1,...,J in which H_j has a zero mean so that the DD estimator would identify α_n. Then, I draw the non-random sample i = 1,...,I using sampling fractions F_j that are correlated with H_j, which introduces the estimation problem. The task that I delegate to the algorithms introduced in section 2.3 is then simply to identify the sampling weights $S_i = F_i^{-1}$ in the observed sample i = 1,...,I that remove the sampling bias. If the correct set of weights is identified, the correlation between the pre-treatment outcome trend $\Delta y_{j,pre}$ and D_i^n is eliminated. If the latent variable H_i is time-invariant as assumed here, the correlation between the outcome trend $\Delta y_{j,post}$ in the post-treatment period and D_i^n is also eliminated, and so is the estimation problem.

To ensure that the algorithms described in section 2.3 stand a chance of identifying the correct sampling weights $S_i = F_i^{-1}$, I define the sampling fractions in the population as $F_j = 1/\sum_m r_m h_j^m$

and the latent variable in equation (2) as $H_j = -M/2 + \sum_m h_j^m$, where h_j^m is one of m = 1,…,M variables that correspond to some locational attributes such as distance from the CBD in the application. As discussed in the next section, the attributed variables h_j^m are constructed so to ensure that $E(H_j) = 0$. With this definition, the sampling bias and the trend interaction that give rise to the estimation problem are a function of the same variables. The algorithms introduced in section 2.3 will exploit this feature to approximate S_i.

2.2 Simulations

I draw individual μ_i and time effects φ_t and the independent component in the trend effect ϑ_i from independent uniform (0,1) distributions once and for all (i.e. they are the same in all experiments). For the time trend, I use polynomial specifications of the orders o=1,2,3, i.e., $f(t) = \sum_o \Gamma_o t^o$ (see Figure A1 in the appendix for an illustration of the functional forms chosen). I set $\alpha_n = 1, I = 1000$, $T = 10$, and $z = 0.5T$. In line with the empirical application in section 3, I set $N = 2$ and M = 3. The $N = 2$ treatment variables $D^{n=1}$ and $D^{n=2}$ and the M = 3 attribute variables $h^{m=1}$, $h^{m=2}$, and $h^{m=3}$ are drawn from independent uniform (0,1) distributions at the beginning of each experiment. In each experiment, I first generate the data for the data universe ($j = 1, .., J$) and then draw a sample (i = 1,…,I) with the probability of an observation being sampled of $F_j = 1/\sum_m r_m h_j^m$, where r_1, r_2, and r_3 are scalars drawn from an independent uniform (0,1) distribution in each experiment. Finally, I add ϵ_{jt}, which is drawn from a normal distribution with a mean of zero and standard deviation of 0.1 in each experiment.

Once the data for an experiment have been generated, I estimate equation (1) using an OLS DD and different versions of WPT DD, in each case omitting any control for $\omega_i f(t)$. Because in the Monte Carlo setting, I have control over the sample fractions, it seems natural to use the sampling weights $S_i = F_i^{-1}$ in one set of experiments. By construction, this definition removes the non-parallel trends problem in the reweighted sample, so that it serves as a sanity check for the Monte Carlo setting. In reality, sampling fractions will not be observed by the researcher, so an approximation of the sampling weights needs to be inferred from the observed data. I discuss two approaches to approximation these weights next.

2.3 Algorithms

There are various potential approaches to finding a suitable set of weights to be used in a WPT DD, and the relative performance is likely context-dependent. Basically, any algorithm that succeeds in finding weights that minimise treatment-trend correlations in the pre-treatment period could be

deemed suitable. In section 3.3.1 I discuss how the validity of a set of weights can be evaluated in an empirical application. While I take advantage of the Monte Carlo setting to evaluate the average performance of two different approaches to approximating S_i, I do not claim that any of them are theoretically superior to any other.

2.3.1 Grid search

In the first approach, I begin by defining the set of potential weights as $\hat{S}_i = \sum_m q_m h_i^m$. The empirical task then is to find the vector of parameters $Q = (q_{m=1}, q_{m=2}, q_{m=3})$. For this purpose, I conduct a grid search over the parameter space defined by $q_1 = 0, 0.1, 0.2, \dots, 1$, $q_2 = 0, 0.1, 0.2, \dots, 1$, $q_3 = 0, 0.1, 0.2, \dots, 1$. I note that I have defined h_i^m so that non-negativity of $\hat{S}_i$ is ensured for any Q. In a case-study application, a transformation of observed variables may be required, such as the Gaussian transformation introduced in section 3.2.2. In each iteration of the grid search, I recover the marginal effect of each treatment variable D_i^n on the outcome trend from a regression of the following form:

$$\Delta\dot{y}_{\imath 1} = c_Q^0 + \sum\nolimits_n c_Q^n \dot{D}_i^n + \aleph_{Qi} \tag{4}$$

where Δy_{j1} is the change in the outcome from the first to the second period and the point accent indicates normalisation by standard deviation, so that c_Q^n is the marginal effect of treatment variable $\dot{D}_j^n$ in units of standard deviations. Targeting the change from the first to the second period (instead of the change over the entire pre-treatment period) is in line with the application in section 3, where the non-targeted periods before the treatment are used to over-identify the WPT weights.

In each regression Q, observations are weighted by a set of weights $\hat{S}_i(Q)$. I select the parameter combination that minimises the "additive" objective $B^A = \sum_n \left(\hat{c}_Q^n\right)^2$. As already discussed in 2.1, I expect this approach to perform well in the Monte Carlo setting because the weights S_i I search for are the inverse of the sample fractions $F_j = 1/\sum_m r_m h_j^m$.

2.3.2 Iterative approach

In the second approach, I am agnostic about the structure of the sampling fractions F_j. I assume that the researcher observes h_i^m. However, the researcher does not know how these variables relate to the sampling fractions. To identify a suitable set of weights $\hat{S}_i$, I follow an iterative process in which I start from an initial set of weights $\hat{S}_i^{s=1} = 1$ in iteration s=1. Each iteration s begins with an estimation of the marginal effect c_s^n of each treatment on the outcome trend over the first period.

$$\Delta\dot{y}_{i1} = c_s^0 + \sum\nolimits_n c_s^n \dot{D}_i^n + \Psi_{si} \tag{10}$$

Next, I run an augmented regression in which I allow for heterogeneity in the marginal effects by adding $\dot{D}_i^n \times h_i^m$ interaction terms.

$$\Delta\dot{y}_{i1} = b_s^0 + \sum\nolimits_n b_s^n \dot{D}_i^n + \sum\nolimits_m \sum\nolimits_n b_s^{n,m} \dot{D}_i^n \times h_i^m + \beth_{si}, \tag{11}$$

where b_s^n and $b_s^{n,m}$ are parameters of interest to be estimated. The estimated marginal effect of a treatment is $\hat{\pi}_i^n = \partial\Delta\dot{y}_{i1}/\partial\dot{D}_i^n = \hat{b}_s^n + \sum_m \sum_n \hat{b}_s^{n,m} h_i^m$. I acknowledge that a distribution of marginal effects could be estimated non-parametrically without requiring h_i^m, e.g. by means of locally weighted regressions (see e.g. McMillen, 1996). To reduce the computational requirements in the Monte Carlo, I opt for a parametric approach.

The algorithm then follows a simple tree structure. If, for any treatment n, $\hat{c}_s^n < 0$, I create $\hat{S}_i^{s=2,n} = g(\hat{S}_i^{s=1}, \hat{\pi}_i^n)$, a set of weights that positively depends on the initial weights and the marginal effect of the treatment, thus the first-order conditions satisfy $g_S > 0$ and $g_\pi > 0$. Likewise, if $\hat{c}_s^n > 0$, I create weights $\hat{S}_i^{s=2,n} = g(\hat{S}_i^{s=1}, \hat{\pi}_i^n)$ that positively depend on the initial weights and negatively on the marginal effect, i.e. $g_S > 0$ and $g_\pi < 0$. For the next iteration, I then create a new set of weights $\hat{S}_i^{s=2} = \sum_n h(|\hat{c}_s^n|)\hat{S}_j^{s=2,n}$, where $h'>0$, i.e., weights are adjusted more strongly for treatments with a larger absolute $\hat{c}_s^n$. The intuition is that the new weights vector $\hat{S}_i^{s=2}$, compared to $\hat{S}_i^{s=1}$, attaches greater weights to observations where the marginal effect is positive if the average marginal effect is negative, and vice versa. This ensures that the algorithm generally converges towards a set of weights that minimise $|\hat{c}_s^n|$. The iterations are repeated until an objective is achieved. Here, I define as the objective that the largest of the standardised partial correlation is below a threshold value v, i.e. I minimise $B^M = \max\left(|\widehat{c^1}|, |\widehat{c_s^2}|\right)$ and stop the algorithm if in iteration s where $B_s^M < v$. I refer to this objective function as the "min-max" objective function.

The algorithm generally achieves the objective quickly, but to keep the Monte Carlo speedy, I set a maximum number of iterations after which I proceed to the next Monte Carlo experiment. I note that this naïve algorithm does not necessarily converge to a global minimum in the objective function. To prevent it from drifting off into an undesirable local minimum (given the global objective), I build in a loop that sets back $\hat{S}_i^s$ to a weighted combination of the current $\hat{S}_i^s$, the previous $\hat{S}_i^{s-1}$ and the "best" (in terms of lowest B^M) weights found across previous iterations if B_s^M exceeds the smallest B^M achieved in past iterations by a sufficiently large margin. I will spare the reader the

detailed functional forms and thresholds that I use in the algorithm. I refer the interested reader to the replication directory.

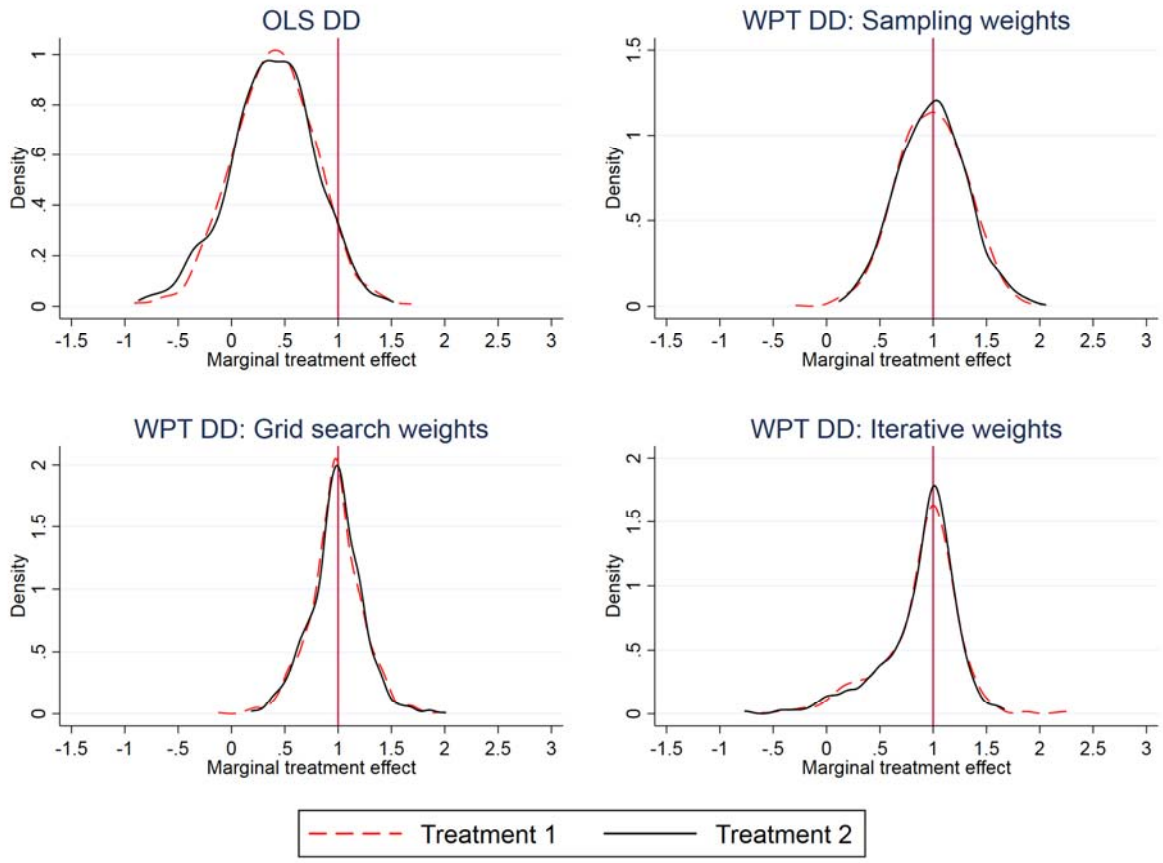
2.4 Results

Figure 1 illustrates the distribution of the estimated marginal treatment effects $\hat{\alpha}_n$ by treatment and estimation method. In the reported experiments, individual trend effects interact with changes in the outcome through a linear time trend *f(t)=t.* Table 1 reports the mean, the median, and the standard deviation (s.d.) for the same distributions. It also shows the results for a similar series of experiments in which I impose a quadratic or a cubic trend (see Figure A1 in the appendix for an illustration). Given the DGP, a distribution of $\hat{\alpha}$ from an unbiased estimator will have a unity mean.

As expected, the OLS DD $\hat{\alpha}_n$ estimates are way below unity, on average. In contrast, the distribution of $\hat{\alpha}_n$ from the WPT DD models are centred on values that are much closer to unity. The WPT DD $\hat{\alpha}_n$ estimates using sampling weights and weights recovered from the grid search algorithm are close to unity, on average. This is not surprising, given that the weights were constructed using procedures that leverage the knowledge of the sampling process. In comparison to the other WPT DD estimators, the WPT DD estimates with the weights generated by the iterative algorithm are more dispersed. The estimates are also smaller on average. But the estimates are generally much closer to unity than the OLS DD estimates.

In this context, it is worth recalling that I stop the interactive process after 100 iterations to speed up the Monte Carlo, even if the absolute standardised marginal effects of the treatments on the pre-treatment outcome trends ($\widehat{c_s^n}$) exceed a threshold of $B^M = \max\left(|\widehat{c_s^1}|, |\widehat{c_s^2}|\right) < v = 0.005$ (the objective that would normally stop the algorithm). It is possible that the Monte Carlo results would improve further if the algorithm was given more time to achieve its objective, but the experiments in which the objective is not achieved are interesting in their own right. In the experiments with a linear time trend in the DGP, the mean value of $\widehat{c_s^{n=1,2}}$ (pooled across treatments) is -0.012, i.e. the distribution is skewed to the left. Across the about 50% of the cases in which $\left|\widehat{c_s^{n=1,2}}\right| < 0.01$ (the objective is nearly achieved), the mean $\hat{\alpha}_{n=1,2}$ are almost exactly unity. The effects of failing to achieve the objective become even more apparent in the experiments with the more demanding cubic trend in the DGP. The mean estimated $\hat{\alpha}_{n=1,2}$ across all experiments (pooled treatments) is 0.58, so way below unity. But, again, the mean $\hat{\alpha}_{n=1,2}$ is almost exactly unity across the experiments in which $\left|\widehat{c_s^{n=1,2}}\right| < 0.01$.

Fig. 1. Distribution of estimated marginal treatment effects with linear trend interaction



Notes: Reported distributions are the estimated $\hat{\alpha}_n$ from 1000 estimations of equation-(1) type DD models, which do not control for $\omega_i f(t)$. The marginal treatment effect in the DGP is $\alpha = 1$ (in the data universe). Monte Carlo experiments are run on a subsample that is selective with respect to characteristics that determine individual trends. Sampling weights are the inverse of the sampling probabilities used in creating the selective subsample. "Grid search" are weights recovered from the algorithm described in section 2.3.1. "Iterative" are the weights recovered by the algorithm described in section 2.3.2. Two positive outliers in both distributions are omitted in the bottom-right panel to improve the presentation.

The upper panels of Figure 2 further support the argument made here that failure to ensure orthogonality between a treatment and the pre-treatment outcome trend will result in biased DD estimates. Within a Monte Carlo experiment (all values drawn are the same), the difference between the WPT DD with weights from the iterative algorithm and the WPT DD with sampling weights can be considered a proxy for the bias of the former because the letter is unbiased by construction. The upper panels illustrate how the performance of the WPT DD critically depends on the achievement of the objective. If $\widehat{c_s^{n=1,2}}$ is large, the bias is large. Equivalently important, provided $\widehat{c_s^{n=1,2}}$ is sufficiently small, little bias can be expected.

The lower panels of Figure 2 exploit the information generated by having multiple treatments in the Monte Carlo experiments. They show that if the algorithm fails to minimise $\widehat{c_s^n}$ for one treatment,

it usually succeeds on the other treatment. This suggests a trade-off in minimising multiple conditional correlations at the same time. One interpretation is that it will be more challenging to achieve unbiased treatment estimates with WPT DD the larger the number of treatments is, although other algorithms may be less sensitive to this problem.

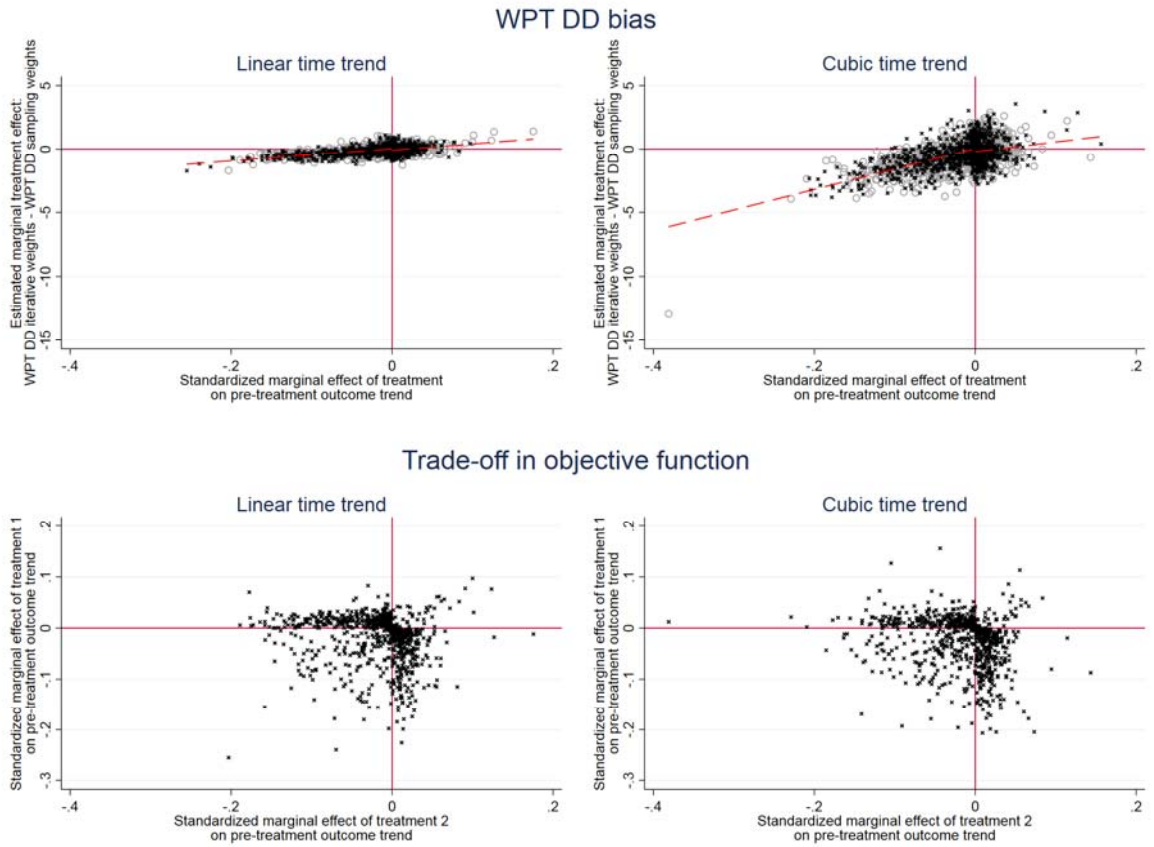
Overall, it seems fair to conclude that in this Monte Carlo setting, the WPT DD generally reduces OLS bias owing to violations of the parallel-trends assumption. To the extent that the correlation between a treatment and the pre-outcome trend is eliminated, the bias is also eliminated.

Tab. 1. Distribution of estimated marginal treatment effects: Summary of all experiments

	Polynomial	Treat-	Experi-	Estimated marginal treatment effects $\hat{\alpha}_n$		
Weights	order of *f(t)*[a]	ment *n*	ments	Mean	Median	S.D.
Uniform (OLS)	1	1	1000	0.402	0.401	0.387
Uniform (OLS)	1	2	1000	0.378	0.395	0.401
Sampling weights	1	1	1000	0.986	0.989	0.326
Sampling weights	1	2	1000	0.990	0.993	0.324
Grid search	1	1	1000	0.979	0.980	0.254
Grid search	1	2	1000	0.982	0.990	0.249
Iterative	1	1	1000	0.884	0.954	0.365
Iterative	1	2	1000	0.877	0.961	0.356
Uniform (OLS)	2	1	1000	0.230	0.260	0.490
Uniform (OLS)	2	2	1000	0.267	0.266	0.507
Sampling weights	2	1	1000	0.987	1.005	0.415
Sampling weights	2	2	1000	1.005	1.003	0.428
Grid search	2	1	1000	0.966	0.975	0.323
Grid search	2	2	1000	0.979	0.995	0.342
Iterative	2	1	1000	0.817	0.914	0.486
Iterative	2	2	1000	0.851	0.942	0.465
Uniform (OLS)	3	1	1000	-0.907	-0.914	1.281
Uniform (OLS)	3	2	1000	-0.982	-0.949	1.289
Sampling weights	3	1	1000	1.037	1.007	1.065
Sampling weights	3	2	1000	0.952	0.929	1.079
Grid search	3	1	1000	0.989	0.991	0.809
Grid search	3	2	1000	0.931	0.955	0.807
Iterative	3	1	1000	0.638	0.954	1.092
Iterative	3	2	1000	0.517	0.890	1.107

Notes: Reported distributions are the estimated $\hat{\alpha}_n$ from 1000 equation-(1) type DD models, which do not control for $\omega_i f(t)$. The marginal treatment effect in the DGP is $\alpha = 1$ (in the data universe). Monte Carlo experiments are run on a subsample that is selective with respect to characteristics that determine individual trends. Sampling weights are the inverse of the sampling probabilities used in creating the selective sub-sample. "Grid search" are weights recovered from the algorithm described in section 2.3.1. "Iterative" are the weights recovered by the algorithm described in section 2.3.2. [a] See Figure A1 in the appendix for an illustration of the functional form of the polynomial functions.

Fig. 2. Bias and achievement of objective in WPT DD with iterative weights



Notes: Circles [crosses] in the upper panel refer to treatment 1 [2]. Linear fits in the upper panels are pooled across treatments 1 and 2. All observations are from experiments using the WPT DD with weights from the iterative algorithm described in section 2.3.2. The time trend in the DGP is defined as indicated in each panel.

3 Application

In this section, I apply the WPT DD approach to a case study that draws from and expands on Ahlfeldt et al. (2016). I provide a brief description of the institutional setting and the data below and refer to Ahlfeldt et al. (2016) for details.

3.1 Background and data

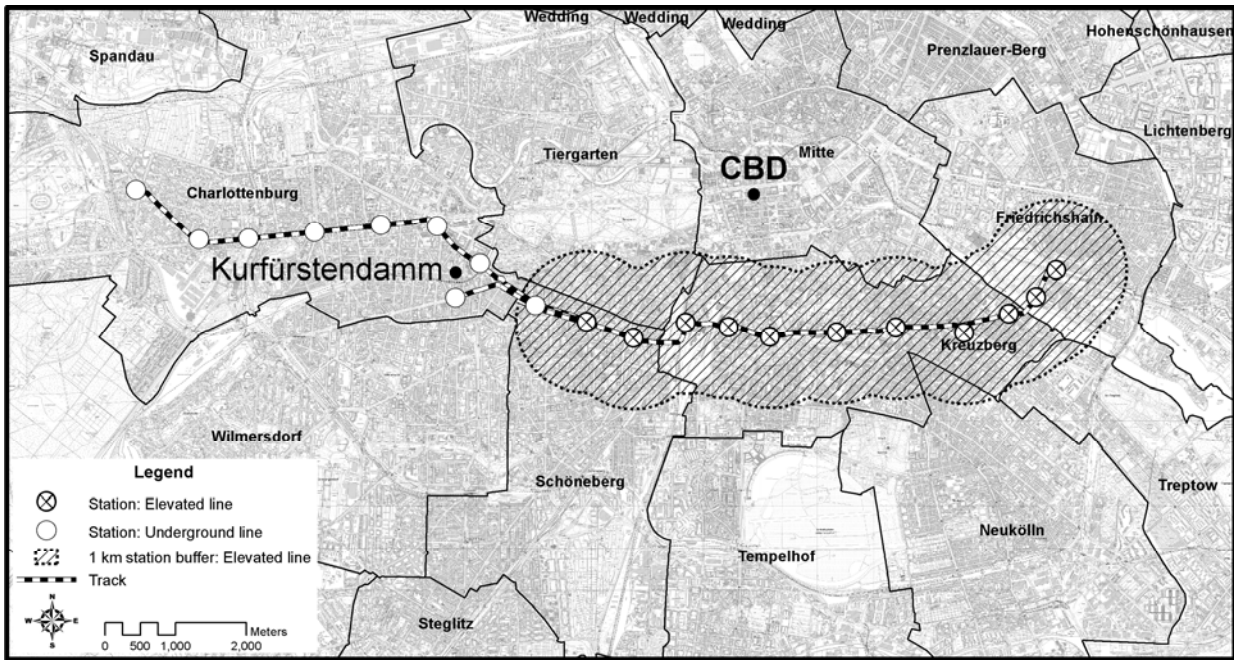
3.1.1 The case study

In 1895, Berlin's (Germany) city government (*Magistrat*) granted a concession for the establishment of an electrified elevated metro rail connecting the eastern parts of Berlin, at the station Warschauer Brücke, and the wealthy western city of Charlottenburg, at the station Zoologischer Garten. At that time, the *Magistrat* considered the project to be too risky and non-profitable and decided not to participate in its final execution (Baltzer, 1897). Therefore, in 1897 (only five years before

the inauguration of the line), Siemens & Halske founded the Elevated Railway Company (*Hochbahngesellschaft*) in cooperation with the Deutsche Bank to guarantee the funding.

The construction began immediately, starting from the eastern parts. However, Berlin residents quickly realised how unpleasant the viaduct's appearance would be. As a result, the city of Charlottenburg managed to ensure, in a last-minute move, that the tracks would run beneath the street surface once the line reached its boundaries. The line was inaugurated in 1902 and was called "Line A" (*Linie A* or *Stammstrecke*). The final routing, depicted in Figure 1, was later described by historians as an outcome of agreements and accidents (Bousset, 1935).

Fig. 3. Routing of Berlin's first metro line (Linie A)



Source: Ahlfeldt et al. (2016).

3.1.2 Data

Land Prices

The measure of land price is extracted from various editions (1881, 1890, 1896, 1900, 1904, 1910 and 1914) of assessed land value maps created by the chartered surveyor Gustav Müller in cooperation with official planning authorities. Müller's maps provide data at a remarkably disaggregated level of individual plots. The stated objective was to provide official and representative guides for both private and public investors participating in Berlin's real estate market. For the considered area of Berlin, our data set contains the full wealth of information provided by the maps. After creating a balanced panel for the final analyses, the data set contains close to 40k parcel-year observations.

While Müller himself did not describe in detail the exact procedure of land valuation, the imperial valuation law (*Reichsbewertungsgesetz*) of the German Reich contained a strict order to use capital values for the assessment of commercial plots based on fair market prices. In line with the valuation laws for commercial land, Müller claims that his assessment refers to the pure value of land, which is adjusted for all building and even garden characteristics. He also corrected values for specific location characteristics such as single and double corner lots, subsoil and courtyard properties. Nowadays, Müller's maps are an established data source. They have been used, among others, by Ahlfeldt et. al (2015), who also provide an extensive data appendix that describes in detail the nature of the data.

Noise

To assess the disamenities from increasing noise levels, we consult a highly disaggregated map, obtained from the Berlin Senate Department for Urban Development, which contains 2007 estimates of the continuous sound level by the source of noise on a 10x10-metre grid. Given that the built-up structure within the affected area remained virtually unchanged after the inauguration of the line,[3] it can reasonably be argued that contemporary noise levels reflect the dissemination of sound in relative terms about 100 years ago. As discussed by Ahlfeldt et al. (2016), there are several reasons why the noise exposure might have been higher in absolute terms (e.g. poor noise insulation, absence of automobile noise). This should not affect the significance levels of the noise estimates reported below, but it gives the estimated per-db noise effects of the character of upper-bound estimates. Arguably, this is an issue of subordinate relevance for the purposes of this paper, so I refer to Ahlfeldt et al. (2016) for a detailed discussion of the implications.

Routing of Line A

Historical network plans provide sufficient information on tracks and individual stations along the route to extract and digitise the whole line.[4] The elevated section of the line consists of 11 stations, while the entire line (including the underground section) consists of 20 with a total length of 15.2km

[3] Note that for very few plots, where the building structure changed, we impute historic noise levels using adjacent plots.

[4] Network plans are also available online; see, for instance, http://www.berlineruntergrundbahn.de and http://www.berliner-verkehr.de.

3.2 Empirical strategy

3.2.1 Baseline difference-in-differences

The baseline empirical strategy combines hedonic (Rosen, 1974) and DD methods (Ashenfelter and Card, 1985). With the hedonic approach the price of a parcel of land is expressed as a function of various attributes, including rail noise and rail access, and their implicit prices. The DD method identifies a treatment effect (e.g. of rail access or rail noise) by differentiating across space (different degrees of exposure) and time (before and after exposure). The baseline empirical specification takes the following form:

$$\begin{aligned}\ln(PRICE_{it}) = {} & \beta^{DIST}[DIST_i \times P(t>1902)_t] + \beta^{NOISE}[NOISE_i \times P(t>1902)_t] \\ & + \sum\nolimits_{A=(1896,1900)} \left[\beta_A^{DIST} DIST_i \times P(t=A)_t + \beta_A^{NOISE} NOISE_i \times P(t=A)_t\right] \\ & + \mu_i + \varphi_t + \varepsilon_{it},\end{aligned} \tag{12}$$

where $PRICE_{it}$ is the land price of a parcel i at time t, μ_i is a parcel fixed effect controlling for unobserved time-invariant heterogeneity onto which I cluster standard errors (Bertrand et al., 2004), φ_t is a year fixed effect controlling for common macroeconomic shocks, and ε_{it} is a random error term. $DIST_i$ is the straight-line distance from the nearest Line A station and the emitted noise level is $NOISE_i$. Both are time-invariant and measured after the metro opening. A=(1896, 1900) is a vector of scalars defining pre-opening years in which anticipation effects could occur. $P(t>1902)_t$ and $P(t=A)_t$ are indicator variables taking the values of one if the conditions are met and zero otherwise. β^{DIST} and β^{NOISE} are the DD parameters of interest, which correspond to $\alpha^{n=1}$ and $\alpha^{n=2}$ in the DGP in section 2.1. Because there was no metro rail noise prior to the elevated train line, the noise measure reflects the increase in noise due to the elevated rail line (such that $NOISE_i = \Delta NOISE_i$, where $\Delta NOISE_i$ is the before-after change in noise). Therefore, $\hat{\beta}^{NOISE}$ provides a first-difference estimate of the effect of rail noise on land prices that can be interpreted as a hedonic implicit price. $\hat{\beta}^{DIST}$ gives the change in the hedonic implicit price of station distance from the before (t<1902) to the after (t>1902) period, i.e. $\hat{\beta}^{DIST} = \delta_{t>1902}^{DIST} - \delta_{t<1902}^{DIST}$, where δ_t^S is the hedonic implicit price in the respective period. $\hat{\beta}^{DIST}$ can also be interpreted as the hedonic implicit price $\delta_{t>1902}^{DIST}$ of station distance since during the before period the stations could not be anticipated and, thus, $\delta_{t<1902}^{DIST} = 0$. The terms in brackets $\left[\beta_A^{DIST} DIST_i \times P(t=A)_t + \beta_A^{NOISE} NOISE_i \times P(t=A)_t\right]$ control for anticipated rail effects in 1896 and 1900 to avoid attenuation bias (Ahlfeldt et al., 2016).

3.2.2 Parallel-trend weights

The untestable assumption of any DD analysis is that, in the absence of a treatment, all subjects (irrespective of the intensity of treatment) would have followed the same trend. A selection problem exists if the treated and the non-treated subjects differ in observable or unobservable dimensions in a way that is correlated with the treatment intensity, and this heterogeneity interacts with time.

The idea of the WPT DD estimator introduced here, is to find weights that minimise the conditional correlations between treatments and pre-treatment outcome trends (the objective). The identifying assumption is that if there is no correlation during the pre-treatment period, there is also no correlation between treatments and trends in potential outcomes in the absence of a treatment (the counterfactual) during the post-treatment period. Finding weights that achieve this objective, is, thus, a critical task in the application of WPT DD.

I follow the assumption conceptualised in the DGP in section 2 that suitable weights can be expressed as a function of observable parcel characteristics. I use the first period in the data – 1881–1890 as the period to be targeted by an algorithm that identifies a set of weights $\hat{S}_i$. In the baseline, I use a grid search approach similar to the one described in section 2.3.1 because – with a sufficiently fine grid – it is likely to get close to a global minimum in the objective function (within a defined parameter space). Concretely, I define weights as:

$$\hat{S}_i = \frac{W_i}{\sum_i W_i}, W_i = \sum\nolimits_m q_m K\left(\lambda_m, h_i^m\right), \tag{13}$$

where, as in section 2.3.1, $Q(q_1, \dots, q_m)$ is a vector of parameters q_m to be identified. Unlike in the Monte Carlo experiments, h_i^m, one of M variables capturing observable time-invariant parcel characteristics, enters the weights in a Gaussian transformation defined as follows:

$$K\left(\lambda_m, h_i^m\right) = \frac{1}{\lambda_m\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{h_i^m - \bar{h}_i^m}{\lambda_m}\right)^2\right), \tag{14}$$

where the bandwidths λ_m are set according to the Silverman (1986) rule and the upper bar indicates the mean of a distribution. I use the Gaussian transformation because I presume that parcels that are more "normal" with respect to a plot characteristic h_i^m are more likely to be on a similar trend. Furthermore, I presume that parcels that are representative with respect to different characteristics h_i^m are likely on different trends. This approach has been chosen so as to mix these different trends in a way that ensures that the average trend in the weighted sample is orthogonal to the treatments. A positive collateral of the Gaussian transformation is that all $K_{i,m} = K(\lambda_m, h_i^m)$ are

non-negative and in the same dimension. I note that the Gaussian transformation works well in terms of the tests presented in section 3.3.1. However, I do not suggest it as a general rule. Different transformations may lead to better results in other applications.

In searching for a vector Q that minimises the objective function, I use a finer grid than in the Monte Carlo. The parameter space over which I search is defined by $q_1 = 0, 0.01, 0.02, \dots, 1$, $q_2 = 0, 0.01, 0.02, \dots, 1$, $q_3 = 0, 0.01, 0.02, \dots, 1$, which equates to 101^3=1,030,301 combinations. I use the additive objective function introduced in section 2.3.1 in the baseline and consider the "min-max" function (introduced in section 2.3.2) and a multiplicative function $\left(\prod_n \left(\widehat{c_n^q}\right)^2\right)$ In a further alteration, I also use the iterative algorithm (section 2.3.2).

3.2.3 Benchmark estimates

Ahlfeldt et al. (2016) provide estimates of the causal effect of noise on land prices, exploiting the noise discontinuity at the tunnel entrance where Line A vanishes below the surface to become an underground line in Charlottenburg. Their boundary discontinuity design in time differences controls for arbitrary unobserved time-invariant effects and unobserved changes in land prices over time that follow a smooth trend in space across the source of the noise discontinuity. Their results suggest that a 10-db increase in rail noise causes a decline in land prices of 5%. This is an estimate identified from plausibly exogenous, but local variation, so the results do not necessarily generalise to the entire elevated part of Line A studied here. However, given that the area connected by Line A is reasonably homogeneous in terms of density, building structure, and amenities, I expect the average effect along Line A to be in the same ballpark. Thus, I view the 5%-figure as a reasonable benchmark against which the OLS DD and WPT DD estimates can be benchmarked for a plausibility check.

A similar benchmark is not available for the station distance effect in the present case-study context. However, there is a wider literature that has estimated the capitalisation effects of metro rail systems exploiting spatiotemporal variation. This literature suggests that a 1km reduction in distance from the nearest station increases house prices by 2–9% (Debrezion et al., 2007; Dubé et al., 2013; Gibbons and Machin, 2005). Assuming competitive markets, a Cobb-Douglas housing production function and a share of land at the property value of 0.25, the house price capitalisation effect can be translated into a land price effect of 8–36% (Ahlfeldt et al., 2015).

3.3 Results

3.3.1 Parallel-trend weights

Before proceeding to the DD analysis of treatment effects it is critical to test whether the non-parallel trends assumption is likely violated in an empirical setting and, if so, to what extent this problem can be addressed by reweighting observations.

In Table 2, I provide two tests of the conditional correlations between treatment variables and pre-treatment outcome trends. Models (1–7) regress the change in ln land price over the 1881–1890 period (the period targeted by the algorithms) against both treatment variables. Models (9–14) replicate the exercise using the change in ln land price over the 1890–1900 period as a dependent variable. This (non-targeted) pre-treatment period has not been inputted into the computation of the weights $\hat{S}_i$, so it can be used in an overidentification test.

Models (1) and (8) present OLS estimation results. There is a significant correlation between station distance and land price growth over the targeted period. Compared to prices right next to a to-be-constructed station, prices at a 1km distance grow at a 0.221 log points higher rate (24%). There is also a significant correlation during the non-targeted period, however, with the opposite sign, suggesting the presence of unobserved effects that interact non-linearly with time. Conditional on the station-distance effect, the noise effect is insignificant. However, station distance and noise are correlated, which explains why the unconditional correlation between noise and the change in prices is significant (to save space, I omit the presentation of formal tests). The main takeaway from these results is that the parallel-trends assumption is violated during the pre-treatment period, thus, it seems likely that it does not hold during the post-treatment period.

The remaining models use weights to address this problem, which are constructed using different algorithms, objective functions and covariates. All approaches succeed in achieving their formal objective of reducing the correlation among treatments and trends during the targeted period (models 2–7). In several instances, the effects of both treatment variables are close to and not statistically distinguishable from zero. The models, using the grid-search algorithm and the Gaussian transformation of land price growth as a covariate, perform best in terms of the overidentification tests reported throughout models (8–14). Apparently, the treatment-trend correlation is low among parcels that experienced “normal” growth over the targeted period. The weights obtained using the iterative algorithm described in section 2.3.2 reduce the treatment-trend correlation to virtually zero (see the r2 close to zero) over the targeted period (model 7). However, they do not pass the

overidentification test as the effect of station distance on land price growth over the non-targeted period is not reduced, but amplified.

The weights used in models (2) and (9) are the most promising in terms of addressing non-parallel trends in the data, as they minimise the treatment variables' effects on outcome trends over the targeted and the non-targeted period. I will use these weights in what I refer to as the baseline specification in the remainder of the paper.

Tab. 2. Marginal treatment effects on pre-outcome trends (placebos)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Ln land price 1890 – ln land price 1881 (targeted period)						
Distance (km)	0.221***	-0.007	-0.024***	-0.006	-0.022**	-0.009	0.002
	(0.028)	(0.010)	(0.009)	(0.076)	(0.009)	(0.010)	(0.019)
Noise (db)	0.008	-0.004	0.001	-0.036**	0.000	-0.004	-0.000
	(0.009)	(0.004)	(0.003)	(0.015)	(0.003)	(0.003)	(0.009)
r2	.0146	.0005	.0051	.0071	.0031	.0004	4.58e-06
	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	Ln land price 1900 – ln land price 1890 (not targeted period)						
Distance (km)	-0.052***	-0.038	-0.054	-0.172***	-0.051	-0.040	-0.193***
	(0.015)	(0.033)	(0.033)	(0.058)	(0.033)	(0.033)	(0.044)
Noise (db)	0.007	-0.011	-0.014	-0.012	-0.014	-0.011	0.031
	(0.006)	(0.011)	(0.012)	(0.011)	(0.011)	(0.011)	(0.021)
r2	.0045	.0011	.0023	.0120	.0021	.0013	.0474
Objective	-	Additive	Additive	Additive	Multi.	Min-max	Min-max
Algorithm	-	Grid search	Grid search	Grid search	Grid search	Grid search	Iterative
Covariates	-	Land price growth, distance from CBD, distance from sub-centre	Land price growth, distance from station, rail noise	Distance from rail track, distance from CBD, distance from sub-centre	Land price growth, distance from CBD, distance from sub-centre	Land price growth, distance from CBD, distance from sub-centre	Land price growth, distance from CBD, distance from sub-centre
N	5,456	5,456	5,456	5,456	5,456	5,456	5,456

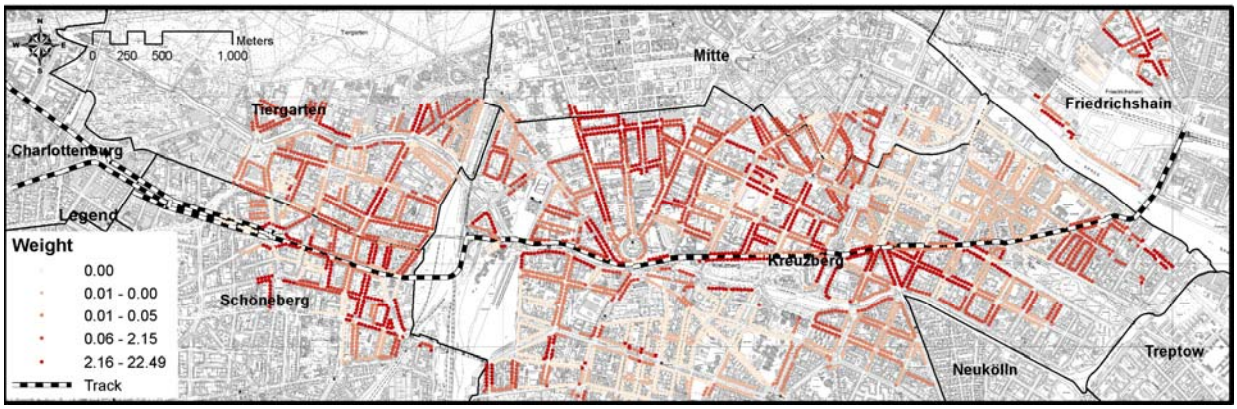
Notes: Unit of observation is parcel. Columns (1) and (8) show results of separate OLS regressions of land price growth over the first (1) and second (2) period in the data against the treatment measures. The subsequent columns show results of weighted regressions, where the weights are recovered using the algorithms, objective functions, and a Gaussian transformation of the covariates indicated in the bottom of the table. The algorithms are outlined in sections 2.3.1. and 2.3.2. Robust standard errors in parentheses. Additive /multi./min-max minimises the sum/product/the largest of squared standardised coefficients on distance and noise. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Throughout section 2, I have assumed that the treatment effect in the DGP is constant across all individuals. Under this constant-effects assumption, the treatment effect estimated by the WPT DD, by definition, is the average (marginal) treatment effect (all units are treated). In an empirical application, however, there may be heterogeneous treatment effects. If the weights are correlated with dimensions in which the treatment effect varies, the WPT DD naturally does not give the aver-

age treatment effect, but a weighted average that corresponds to a parcel with the mean characteristics of the weighted sample. For a judgement of how representative this weighted average is for the full population, it is useful to inspect the selectivity of the weighted sample.

The baseline weights are mapped in Figure 4. Overall, parcels with high weights are distributed relatively evenly across the study area. The most notable pattern is that of areas with relatively low parcel weights in the southern central section and the north-eastern part of the study area. Table 3 compares descriptive statistics of the weighted sample to the unweighted population. The distributions are fairly similar. In line with Figure 4, the mean parcel in the weighted sample is somewhat closer to the CBD (Stadtmitte, in the north) and the sub-centre (Kurfürstendamm in the west). But, overall, the weights inspection suggests that the results in the WPT DD will not be driven by a small number of non-representative parcels, so the estimates will hopefully be not too far from an average effect. Most likely, the WPT DD will have greater external validity than the benchmark estimate from the boundary discontinuity design discussed in section 3.2.3, which is identified from a small number of parcels around the tunnel entrance.

Fig. 4. Spatial distribution of weights



Notes: Weights are constructed using the algorithm described in section 2.3.1 and Gaussian transformations of the 1881 to 1890 land price growth, the distance from the CBD and the distance from the most important sub-centre (*Kurfürstendamm*). Classes defined based on quintiles. Own illustration using the Urban Environmental Information System of the Berlin Senate Department (Senatsverwaltung für Stadtentwicklung Berlin, 2006).

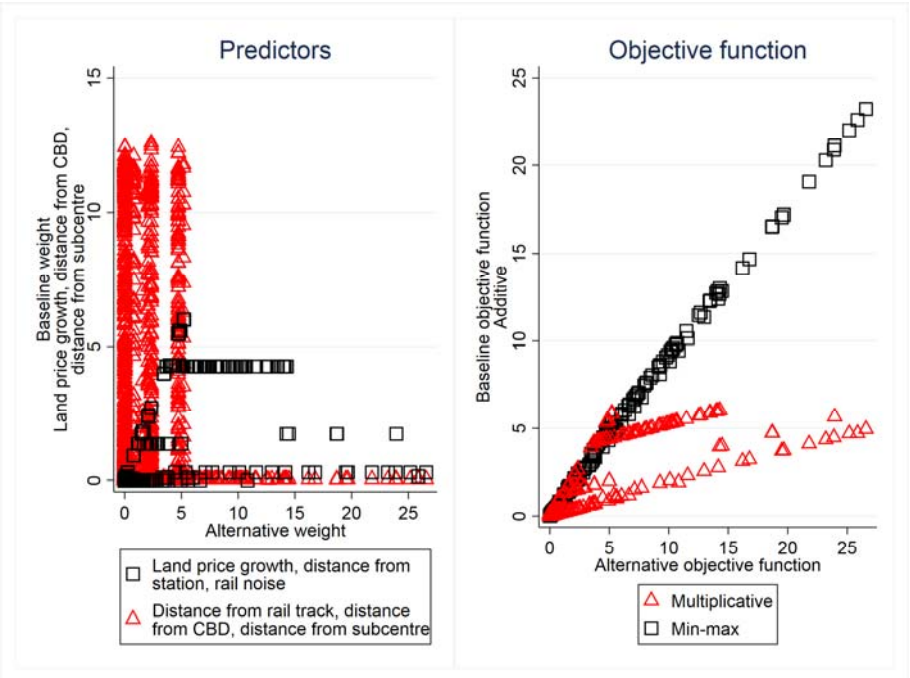
Tab. 3. Descriptive statistics in weighted vs. non-weighted sample

	Non-weighted			Weighted		
	Mean	Median	S.D.	Mean	Median	S.D.
Ln land price 1881	4.213	4.094	0.605	4.388	4.094	0.615
Ln land price 1914	5.854	5.768	0.521	6.058	5.991	0.591
Station distance (km)	0.502	0.491	0.237	0.467	0.486	0.226
Noise (10 db)	0.229	0.010	0.553	0.321	0.013	0.665
Distance from CBD	2.018	2.061	0.797	1.764	1.733	1.033
Distance from sub-centre	4.212	4.258	1.725	3.999	3.703	1.712
Distance from Line A track	0.543	0.517	0.265	0.559	0.503	0.310

Notes: Weights are constructed using the algorithm described in section 2.3.1 and Gaussian transformations of the 1881 to 1890 land price growth, the distance from the CBD and the distance from the most important sub-centre (*Kurfürstendamm*).

In Figure 5, I compare the weights obtained by using different covariates and objective functions in the grid search algorithm. While variations in the objective function tend to result in similar weights, different sets of covariates result in more substantial variation. The weights used in Table 2, models (4) and (11), as an example, are virtually uncorrelated with the baseline weights from models (2) and (9) (correlation coefficient: 0.076). Ideally, WPT DD results will be replicable using different sets of uncorrelated weights as this suggests that identification is not driven by a limited number of units receiving high weights.

Fig. 5. Correlation among WPT weights



Notes: Figure shows weights created using the grid search algorithm using different predictor variables and objective functions in the algorithm.

3.3.2 OLS difference-in-differences

In Table 4, I present (unweighted) OLS DD estimation results from variations of the model described by equation (1). I consider distance and noise estimates separately (1–2 vs. 3–4) as well as versions excluding (1,3,5) and including (2,4,6) anticipation effects. The preferred full specification is model (6). The results are clearly not within a plausible range, given the evidence discussed in section 3.2.3. The estimates of the noise effect are either positive, which is implausible, or insignificant. The station distance effects are generally relatively small and even insignificant in the preferred model (6). Controlling for anticipation effects reduces the treatment effects, whereas the opposite is theoretically expected. In sum, the OLS DD results are inconclusive, which seemingly confirms that DD estimates are not reliable if the parallel trends assumption is violated.

Tab. 4. OLS difference-in-differences estimates

	(1)	(2)	(3)	(4)	(5)	(6)
	Ln land price					
Distance (km) x (t > 1900)	-0.076***	-0.038**			-0.070***	-0.033
	(0.013)	(0.018)			(0.015)	(0.021)
Noise (10 db) x (t > 1900)			0.020***	0.012*	0.005	0.005
			(0.005)	(0.006)	(0.005)	(0.007)
Parcel effect	Yes	Yes	Yes	Yes	Yes	Yes
Year effect	Yes	Yes	Yes	Yes	Yes	Yes
Anticipation Effect	-	Yes	-	Yes	-	Yes
N	38,192	38,192	38,192	38,192	38,192	38,192
r2	.889	.889	.889	.889	.889	.889

Notes: Unit of observation is parcel-year (balanced panel). Announcement effects are distance and noise variables interacted with 1896 and 1900 effects. Balanced panel of repeated parcel observations for 1881, 1890, 1896, 1900, 1904, 1910 and 1914. Standard errors in parentheses are clustered in parcels. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.3.3 Weighted-parallel-trends difference-in-differences

In Table 5, I replicate all Table 4 models in the same order, this time weighting observation by the baseline weights tested in Table 2, models (2) and (9) and illustrated in Figure 4. The results are much more intuitive to interpret. Proximity to a station has a positive effect on land prices and the effect increases significantly if the negative effect of rail noise is controlled for (5–6 vs. 1–2). The noise effect is significantly negative if – and only if – station access is controlled for (5–6 vs. 3–4), as expected in the presence of strong countervailing spatial externalities. Controlling for anticipation effects now increases the treatment effect (6 vs 5) as expected. Most notably, the estimated per-10-db noise effect in the preferred specification (6) is now very close to the benchmark (-4.6% vs. 5%). The estimated effect of 0.191 log points (about 21%) per-station-distance-kilometre effect (6) is well within the 8–36% range prevailing in the literature (see section 3.2.3). Overall, the WPT DD results are more plausible than the OLS results.

Tab. 5. WPT DD estimates: Baseline models (grid search weights)

	(1)	(2)	(3)	(4)	(5)	(6)
	Ln land price (1881-1914)					
Distance (km) x (t > 1900)	-0.119***	-0.116***			-0.174***	-0.191***
	(0.025)	(0.032)			(0.030)	(0.039)
Noise (10 db) x (t > 1900)			-0.001	-0.010	-0.034***	-0.046***
			(0.007)	(0.009)	(0.008)	(0.011)
Parcel effects	Yes	Yes	Yes	Yes	Yes	Yes
Year effects	Yes	Yes	Yes	Yes	Yes	Yes
Anticipation effects	-	Yes	-	Yes	-	Yes
N	38,192	38,192	38,192	38,192	38,192	38,192
r2	0.930	0.930	0.930	0.930	0.930	0.931

Notes: Results replicated based on models reported by Ahlfeldt et al. (2016). Unit of observation is parcel-year (balanced panel). WPT DD models use weights constructed to minimise the conditional correlations between noise and the 1881–1890 land price trend as well as access (distance from station) and the 1881–1890 land price trend. Weights are constructed using the algorithm described in section 2.3.1 and Gaussian transformations of the 1881 to 1890 land price growth, the distance from the CBD and the distance from the most important sub-centre (*Kurfürstendamm*). Announcement effects are distance and noise variables interacted with 1896 and 1900 effects. Balanced panel of repeated parcel observations for 1881, 1890, 1896, 1900, 1904, 1910 and 1914. Standard errors in parentheses are clustered in parcels. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In Table 6, I present results from the same models in the same order as in Tables 4 and 5, this time using weights generated by the iterative algorithm introduced in section 2.3.2. The pattern of results is generally similar to that of the baseline WPT DD results in Table 5. In particular, the estimated noise effects in the preferred specifications (6) are close to each other. At -0.295 log points the effect of a station distance increase by one kilometre is larger than in the baseline, but much closer than in the OLS DD estimates. As discussed above, I prefer the results of the baseline specification (in Table 5) over those presented in Table 6 because the baseline weights do better in terms of the tests reported in Table 2. Still, I consider the relative similarity of the results in both tables as reassuring.

In Table 7, I replicate models (5) (excluding anticipation effects) and (6) (including anticipation effects) from Table 5, changing the set of covariates used in the weights construction. The alternative specifications yield results within reasonably close range, which is reassuring given the low correlation between the different sets of weights (see in Figure 5, left panel). Using different objective functions in the algorithm in Table 8 yields similar results, as expected given the high correlation among the weights (see Figure 5, right panel).

Tab. 6. WPT DD estimates: Weights from iterative algorithm

	(1)	(2)	(3)	(4)	(5)	(6)
	Ln land price (1881-1914)					
Distance x (km) x (t > 1900)	-0.171***	-0.246***			-0.221***	-0.295***
	(0.026)	(0.036)			(0.035)	(0.049)
Noise (10 db) x (t > 1900)			0.007	0.027*	-0.051***	-0.050**
			(0.010)	(0.016)	(0.015)	(0.022)
Parcel effect	Yes	Yes	Yes	Yes	Yes	Yes
Year effect	Yes	Yes	Yes	Yes	Yes	Yes
Anticipation	-	Yes	-	Yes	-	Yes
N	38,192	38,192	38,192	38,192	38,192	38,192
r2	.919	.920	.918	.918	.919	.920

Source: Unit of observation is parcel-year (balanced panel). WPT DD models use weights constructed to minimise the conditional correlations between noise and the 1881–1890 land price trend as well as access (distance from station) and the 1881–1890 land price trend. Weights are constructed using the algorithm described in section 2.3.2 and Gaussian transformations of the 1881 to 1890 land price growth, the distance from the CBD and the distance from the most important sub-centre. Announcement effects are distance and noise variables interacted with 1896 and 1900 effects. Balanced panel of repeated parcel observations for 1881, 1890, 1896, 1900, 1904, 1910 and 1914. Standard errors in parentheses are clustered in parcels. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Tab. 7. WPT DD: Varying predictors (grid search weights)

	(1)	(2)	(3)	(4)	(5)	(6)
	Ln land price (1881-1914)					
Distance x (km) x (t > 1900)	-0.174***	-0.191***	-0.183***	-0.214***	-0.256***	-0.315***
	(0.030)	(0.039)	(0.031)	(0.040)	(0.044)	(0.061)
Noise (10 db) x (t > 1900)	-0.034***	-0.046***	-0.039***	-0.051***	-0.018*	-0.037***
	(0.008)	(0.011)	(0.008)	(0.011)	(0.010)	(0.014)
Parcel effect	Yes	Yes	Yes	Yes	Yes	Yes
Year effect	Yes	Yes	Yes	Yes	Yes	Yes
Anticipation effects	-	Yes	-	Yes	-	Yes
Predictors	Land price growth, distance from CBD, distance from sub-centre	Land price growth, distance from CBD, distance from sub-centre	Land price growth, distance from station, rail noise	Land price growth, distance from station, rail noise	Distance from rail track, distance from CBD, distance from sub-centre	Distance from rail track, distance from CBD, distance from sub-centre
N	37,933	37,933	37,898	37,898	38,192	38,192
r2	.931	.931	.929	.93	.915	.916

Notes: Unit of observation is parcel-year (balanced panel). WPT DD models use weights constructed to minimise the conditional correlations between noise and the 1881–1890 land price trend as well as access (distance from station) and the 1881–1890 land price trend. Weights are constructed using the algorithm described in section 2.3.1 and Gaussian transformations of the listed covariates. Land price growth is the deviation from the mean 1881 to 1890 land price growth. Announcement effects are distance and noise variables interacted with 1896 and 1900 effects. Balanced panel of repeated parcel observations for 1881, 1890, 1896, 1900, 1904, 1910 and 1914. Standard errors in parentheses clustered in parcels. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Tab. 8. WPT DD: Varying objective functions (grid search weights)

	(1)	(2)	(3)	(4)	(5)	(6)
	Ln land price (1881-1914)					
Distance x (km) x (t > 1900)	-0.175***	-0.192***	-0.182***	-0.211***	-0.180***	-0.205***
	(0.030)	(0.039)	(0.031)	(0.040)	(0.031)	(0.040)
Noise (10 db) x (t > 1900)	-0.034***	-0.046***	-0.038***	-0.050***	-0.034***	-0.047***
	(0.008)	(0.011)	(0.008)	(0.011)	(0.008)	(0.011)
Parcel effect	Yes	Yes	Yes	Yes	Yes	Yes
Year effect	Yes	Yes	Yes	Yes	Yes	Yes
Anticipation effects	-	Yes	-	Yes	-	Yes
Objective function	Additive	Additive	Multipli-cative	Multipli-cative	Min-max	Min-max
N	37933	37933	38052	38052	37933	37933
r2	.931	.931	.93	.93	.93	.93

Notes: Unit of observation is parcel-year (balanced panel). Weighted models use weights constructed to minimise the conditional correlations between noise and the 1881–1890 land price trend as well as access (distance from station) and the 1881–1890 land price trend. Weights are constructed using a Gaussian transformation of the 1881 to 1890 land price growth, the distance from the CBD and the distance from the most important sub-centre. Announcement effects are distance and noise variables interacted with 1896 and 1900 effects. Balanced panel of repeated parcel observations for 1881, 1890, 1896, 1900, 1904, 1910 and 1914. Standard errors in parentheses clustered in parcels. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4 Conclusion

In this paper, I discuss an approach to improving counterfactuals in difference-in-differences (DD) analysis when the parallel trends assumption is likely violated. In the weighted-parallel-trends (WPT) DD estimator discussed here, weights are used that minimise the conditional correlation between one or multiple treatment variables and pre-treatment trends in an outcome. I argue that if the researcher is able to identify a set of (time-invariant) weights that reduce this correlation in several pre-treatment periods, it is also likely that this correlation will be reduced in the remaining (post-treatment) periods. Thus, it is more likely that the critical parallel-trends assumption holds.

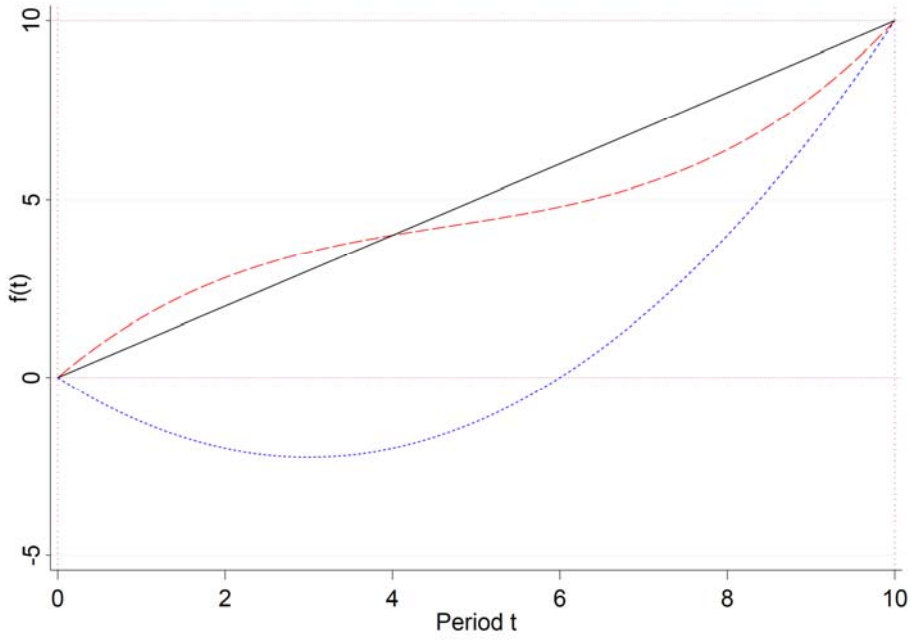
I subject this intuition to a test in a Monte Carlo study and provide an application to a case study that draws from and expands on Ahlfeldt et al. (2016). The Monte Carlo results suggest that the WPT DD has the potential to reduce OLS bias to the extent that the objective of minimising treatment-trend correlations (over one or several pre-treatment periods) is achieved. In the case study application, the WPT DD provides results that are more plausible, given theoretical expectations and previous evidence, than the OLS DD. Reassuringly, different implementations of the WPT DD yield similar results. One of the strengths of WPT DD is that it is applicable to a wide range of empirical settings, e.g. cases with multiple continuous treatment variables and intervention studies that aim at estimating time-varying treatment effects. Moreover, the WPT DD is relatively straightforward to implement and is transparent in the sense that its ability to ensure orthogonality between treatments and individual trends over targeted and non-targeted pre-treatment periods is

easy to test. It is also straightforward to test for a selectivity of the weighted sample, which may create an external validity problem if there is heterogeneity in treatment effects. Therefore, the WPT DD represents a potential avenue to be explored by applied researchers in instances where the ready-made tools for dealing with a non-parallel-trends problem do not fit the empirical setting.

At this stage, I do not recommend a specific procedure for the identification of weights that serve the aforementioned objective. My view is that it is up to the researcher's creativity to develop an approach that suits a particular application. As discussed above, it is easy enough to test the plausibility of proposed weights. I leave the development of a 'one-size-fits-all' algorithm to future research.

Appendix

Fig. A1: Polynomial time trend specifications



Notes: Linear (solid) trend: f(t)=t
Quadratic (short dash) trend: f(t) =-1.5t + 0.25t^2
Cubic (long dash) trend: f(t) = 2t - 0.35t^2 + 0.025t^3.

Literature

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

Abadie, A., & Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review, 93*(1), 113-132.

Ahlfeldt, G. M., Koutroumpis, P., & Valletti, T. (2017). Speed 2.0: Evaluating Access to Universal Digital Highways. *Journal of the European Economic Association, 15*(3), 586-625.

Ahlfeldt, G. M., Moeller, K., Waights, S., & Wendland, N. (2017). Game of Zones: The Political Economy of Conservation Areas. *The Economic Journal, 127*(605), F421-F445.

Ahlfeldt, G. M., Moeller, K., & Wendland, N. (2015). Chicken or egg? The PVAR econometrics of transportation. *Journal of Economic Geography, 15*(6), 1169-1193.

Ahlfeldt, G. M., Nitsch, V., & Wendland, N. (2016). Ease vs. Noise: On the Conflicting Effects of Transportation Infrastructure. *CESifo Working Paper No. 6058*.

Angrist, J. D., & Krueger, A. B. (1999). Chapter 23 - Empirical Strategies in Labor Economics. In O. C. Ashenfelter & D. B. T. H. o. L. E. Card (Eds.), (Vol. 3, pp. 1277-1366): Elsevier.

Ashenfelter, O., & Card, D. (1985). Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs. *The Review of Economics and Statistics, 67*(4), 648-660.

Athey, S., & Imbens, G. W. (2006). Identification and Inference in Nonlinear Difference-in-Differences Models. *Econometrica, 74*(2), 431-497.

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.

Bai, J. (2003). Inferential Theory for Factor Models of Large Dimensions. *Econometrica, 71*(1), 135-171.

Baltzer, F. (1897). Die elektrische Stadtbahn in Berlin von Siemens & Halske. *Zeitschrift für Kleinbahnen*.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust difference-in-difference estimates? *The Quarterly Journal of Economics, 119*(1), 249-275.

Besley, T., & Case, A. (2000). Unnatural Experiments? Estimating the Incidence of Endogenous Policies. *The Economic Journal, 110*(467), 672-694.

Blundell, R., & Macurdy, T. (1999). Chapter 27 - Labor Supply: A Review of Alternative Approaches. In O. C. Ashenfelter & D. B. T. H. o. L. E. Card (Eds.), (Vol. 3, pp. 1559-1695): Elsevier.

Blundell, R., Dias, M. C., Meghir, C., & van Reenen, J. (2004). Evaluating the employment impact of a mandatory job search program. *Journal of the European Economic Association, 2*(4), 569-606.

Bousset, E. H. J. (1935). *Die Berliner U-Bahn*. Berlin: Ernst & Sohn.

Conley, T. G., & Taber, C. R. (2011). Inference with "Difference in Differences" with a Small Number of Policy Changes. *The Review of Economics and Statistics, 93*(1), 113-125.

Debrezion, G., Pels, E., & Rietveld, P. (2007). The Impact of Railway Stations on Residential and Commercial Property Value: A Meta-analysis. *The Journal of Real Estate Finance and Economics, 35*(2), 161-180.

Dubé, J., Thériault, M., & Des Rosiers, F. (2013). Commuter rail accessibility and house values: The case of the Montreal South Shore, Canada, 1992–2009. *Transportation Research Part A: Policy and Practice, 54*(Supplement C), 49-66.

Gibbons, S., & Machin, S. (2005). Valuing rail access using transport innovations. *Journal of Urban Economics, 57*(1), 148-169.

Gobillon, L., & Magnac, T. (2016). Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls. *The Review of Economics and Statistics, 98*(3), 535-551.

Gomez, M. (2015). REGIFE: Stata module to estimate linear models with interactive fixed effects. https://ideas.repec.org/c/boc/bocode/s458042.html.

Guardabascio, B., & Ventura, M. (2013). Estimating the dose-response function through the GLM approach. Stata program. https://mpra.ub.uni-muenchen.de/45013/.

Hainmueller, J., Abadie, A., & Diamond, A. (2011). SYNTH: Stata module to implement Synthetic Control Methods for Comparative Case Studies. https://ideas.repec.org/c/boc/bocode/s457334.html.

Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis, 20*(1), 25-46.

Hainmueller, J., & Xu, Y. (2013). ebalance: A Stata Package for Entropy Balancing. *Journal of Statistical Software, 54*(7).

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

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies, 64*(4), 605-654.

Hernán, M. A., Brumback, B., & Robins, J. M. (2001). Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments. *Journal of the American Statistical Association, 96*(454), 440-448.

Kline, P. (2011). Oaxaca-Blinder as a Reweighting Estimator. *American Economic Review, 101*(3), 532-537.

Lee, M.-J. (2015). Generalized Difference in Differences With Panel Data and Least Squares Estimator. *Sociological Methods & Research, 45*(1), 134-157.

Lee, M.-J. (2016). *Matching, regression discontinuity, difference in differences, and beyond.* Oxford: Oxford University Press.

Leuven, E., & Sianesi., B. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. http://ideas.repec.org/c/boc/bocode/s432001.html. .

McMillen, D. P. (1996). One Hundred Fifty Years of Land Values in Chicago: A Nonparametric Approach. *Journal of Urban Economics, 40*(1), 100-124.

Meyer, B. D. (1995). Natural and Quasi-Experiments in Economics. *Journal of Business & Economic Statistics, 13*(2), 151-161.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy, 82*(1), 34-55.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Senatsverwaltung für Stadtentwicklung Berlin. (2006). *Urban and Environmental Information System*. Berlin.

Shaun, R. S., & Ian, R. W. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research, 22*(3), 278-295.

Silverman, B. W. (1986). Density Estimation For Statistics and Data Analysis. *Monographs on Statistics and Applied Probability.*

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