

The Use of Two Control Groups in Quasi-Experimental Program Evaluation

Craig McIntosh*

September 9, 2004

Abstract

The structure of many applied quasi-experiments generates two control groups. This paper shows that by matching to the second control across the same space which assigns treatment status, we can test and relax the assumptions required for the estimation of treatment effects. By varying the number of matches used, the standard difference-in-differences estimator and a minimum bias nearest-neighbor estimator can be nested within the same technique. The empirical properties of these estimators are investigated using simulated data, and the paper concludes with an extension of the method to panel data and to non-parametric estimation.

JEL Classification: C14, C21, C52, C93

Keywords: Impact Analysis, Ignorability, Triple-Differencing, Matching

*Graduate School of International Relations and Pacific Studies, University of California, San Diego. 9500 Gilman Drive, La Jolla, California, 92093-0519. Phone: (858)822-1125. Fax: (858)6534-3939. Email: ctmcintosh@ucsd.edu. Many thanks to Alain de Janvry, Gordon Hanson, Guido Imbens, Ethan Ligon, Elisabeth Sadoulet, and Michael Ward for comments on earlier versions of this paper. All remaining errors are my own.

1 Introduction

The estimation of the impact of binary policy treatments is a central focus of applied econometricians. When faced with quasi-experimental data, the researcher's problem is to generate a counterfactual; units which are 'like' the treated but did not receive the program. A difference in differences (DID) regression achieves this parametrically through the use of control variables, and matching estimators do so through direct comparison to similar untreated units. Many real-world programs present two layers of selection; for example units must first qualify for the program and then elect to take it, or they must be within an administrative region which receives the treatment, and then they must qualify as units. In this case, both DID and matching estimators make assumptions as to the ignorability of unobservable variables in this joint decision in order to estimate impacts, an assumption which usually cannot be tested within-sample. Failure of such estimators to replicate experimental impact measures (LaLonde 1986) has cast doubt on the validity of the ignorability assumption, fuelling much of the recent interest in randomization.

This paper shows how a second control group can be used to test and relax the ignorability assumption. Standard techniques, not using this second control group, must identify treatment through unexplained differences between the treatment and control and therefore are forced to assume ignorability. The presence of this second control, subject to assumptions laid out in the body of the paper, allows us to test for ignorability, and in the event that it fails, to estimate the counterfactual distribution of outcomes across the rule which selects agents into the treatment and the control. We develop the theory of this spatial matching to the second control, implement it on simulated data, and conclude the paper with extensions of the technique, including application to panel data and a fully non-parametric

estimator.

We denote the space across which groups of agents are selected to be offered the treatment as the treatment criterion; depending on the application the space may be physical, temporal, or a group qualification criterion. For the second control to exist, there must be an additional layer of selection; this may arise as a result of unit-level eligibility criteria or as a result of a voluntary treatment. This second layer of selection which causes units within the treatment space actually to receive it we call the selection criterion, and those selected by this criterion we call compliers. What is suggested here is that the group of non-compliers contains information which allows us to test for ignorability across the treatment criterion. Specifically, are there unexplained effects within this untreated group which violate the assumption necessary to use differences between treatment and control regions to estimate impact among the compliers?

The possibility of using this information to perform such a test was first suggested by Rosenbaum (1982), and several recent papers conduct some form of triple-differencing which utilizes the information in both control groups. This comparison is usually accomplished either by calculating lump-sum averages among non-compliers (Morduch 1998) or through calculating fixed effects which include non-compliers (Gruber 1994), (Pitt & Khandker 1998). We show that methods which use aggregated groups of non-compliers are unlikely to remove bias from impact estimates. Some authors have implemented non-parametric forms of triple-differencing (Ravallion, Galasso, Lazo & Philipp 2002), (Duflo 2000); what is offered here is a general method which can be implemented either semi-parametrically or non-parametrically to a wide range of applied policy issues. By estimating a surface which represents unexplained heterogeneity across the treatment criterion, we

can match each individual complier to this surface and generate quasi-experimental outcomes which satisfy ignorability by construction.

Intuitively, the meaning of this spatial matching estimator is that we will only ascribe impact to a treatment/control regime if it altered outcomes among compliers above and beyond the spatial outcome surface observed among local non-compliers. Such a definition of impact has several attractive features. First, it allows us to distinguish the impact of innovations in a chaotic environment, where unexplained shocks in the treatment space may lead to violations of ignorability. Secondly, the pattern of residuals across the treatment criterion is often itself of direct interest. Thirdly, this technique allows us to measure a kind of impact invisible to standard techniques: the potential of a treatment to insulate agents against such shocks. Finally, failures of ignorability often arise as a result of endogenous placement of quasi-experiments; we show that as long as the endogeneity of the treatment criterion is defined by the treatment space itself, this technique will allow for unbiased estimation of impact.

2 Estimation Under Ignorability

We study changes in changes in unit-level outcomes, with this change denoted by Y_i . The first observation is pre-treatment for all units and the second observation includes a treatment effect for treated units. By using a change as the dependant variable in a cross-section we set up a simple form of double-differencing, and avoid the autocorrelation problems discussed in Bertrand, Duflo & Mullainathan (2004). These changes are determined by some function $Y_i = f(X_i, s_i, Z_i) + \epsilon_i$, where X_i denotes a vector of observable variables, Z_i a vector of unobservable variables, and

s_i agent i 's location in the space of the treatment criterion. The treatment criterion is a rule that maps $s_i \mapsto [0, 1]$; we denote this mapping by τ , so $T_i = \tau(s_i)$. In cases where the treatment is also voluntary or has a second-tier eligibility requirement, the second level of selection is indicated by the binary variable ω_i . Those with $\omega_i = 1$ are called compliers, and those agents with $T_i = \omega_i = 1$ are subjected to a treatment effect $t(X_i, Z_i, s_i)$. In potential outcomes notation, Y_{1i} represents the outcome for a treated agent, and Y_{0i} represents the counterfactual untreated outcome for the same agent. Without loss of generality, we can think of the treatment effect as additive, so that $Y_{1i} = Y_{0i} + t(X_i, s_i, Z_i)$.

Under perfect and involuntary randomization, regardless of the space s used to select agents into the treatment, incidence of the treatment is orthogonal to all observable and unobservable effects, and so $E(Y_0 | T = 1) = E(Y | T = 0)$. In this case, with no parametric methods we can easily recover the average treatment effect (ATE) by calculating $E(Y | T = 1) - E(Y | T = 0) = E(t) + E(Y_0 | T = 1) - E(Y | T = 0) = E(t)$. Much of the recent evaluation literature (Kremer 2003), (Duflo & Kremer 2003), has promoted the use of randomized evaluations as the surest method of estimating treatment effects. Where a voluntary program is randomized, so $\omega_i = 0$ for some agents for whom $T_i = 1$, we can recover the Intention to Treat Effect (ITE) with the above estimator but cannot estimate the Treatment Effect on the Treated (TET) without assumptions over the selection process.

Using non-experimental data, if $\omega_i = 1 \forall i$ we can calculate the average treatment effect (ATE) using a DID estimator by $Y_i = \beta X_i + \delta T_i + \mu_i$ (Card & Krueger 1994), or more complex forms of double-differencing as in Duflo (2001). Similarly, it can be identified using a matching estimator to calculate counterfactual outcomes among untreated agents with 'similar' observations of X . In certain contexts

instrumental variables may be used to infer average treatment effects (Angrist & Krueger 1991), (Imbens, Rubin & Sacerdote 2000) or local-average treatment effects among a specific group of compliers (Angrist, Imbens & Rubin 1996), however such instruments are difficult to come by in most applications.

Where not all agents in the treatment space are treated, so $\omega_i = 0$ for some agents for whom $T_i = 1$, the intention to treat effect (ITE) can be estimated using the DID estimator. In order to estimate the TET, we can try to construct a counterfactual either by comparison to non-compliers, or by identifying a parallel population which is comparable but was not offered the treatment. Much of the matching literature focuses on the first approach, where ignorability requires that ω be a function strictly of observable variables (Rosenbaum & Rubin 1983), (Rosenbaum & Rubin 1984), (Ahn & Powell 1993), (Angrist 1995). Since it is only where we match to a group not offered the treatment that we generate two controls, we focus on this application of matching. Here, we must make one assumption in order to be able to identify the agents in the control who would have been compliers:

$$\textit{Selection Assumption} : E(\omega | X, T = 1) = E(\omega | X, T = 0)$$

(Ignorability of treatment on selection decisions)

and another assumption which says that counterfactual outcome among compliers would have been comparable in the treatment and control:

$$\textit{Outcome Assumption} : E(Y_0 | X, \omega, T = 1) = E(Y | X, \omega, T = 0).$$

(Ignorability of treatment and selection on outcomes)

Using only the Outcome Assumption, the intention to treat effect (ITE) can

be estimated using the DID estimator above. Estimation of the average treatment effect on the treated (TET) requires us to use both assumptions (Heckman 1979). This can be achieved through matching on propensity to select into the treatment (Heckman, Ichimura & Todd 1997), (Abadie 2003), or directly observed through surveys in the control and estimated through the use of a DID regression of the form $Y_i = \beta X_i + \delta T_i + \mu_i \forall \omega_i = 1$.

When we explain Y_i with a linear regression, we have

$$Y_i = \beta X_i + \phi(X_i, s_i, Z_i) + \epsilon_i,$$

where $\phi(X_i, s_i, Z_i)$ is the entire systematic component of outcomes that is orthogonal to the linear specification on observables. This allows us to rewrite the outcomes assumption as $E(\phi | \omega, T = 1) = E(\phi | \omega, T = 0)$, or that the unexplained systematic component of outcomes is equal in the treatment and the control regions. This assumption must be satisfied in order for either the TET or the ITE to be estimated consistently. Thus any shocks, unobserved variables that effect rates of change, or non-linear relationships that differ between treatment and control will bias the standard DID approach. The presence of these shocks cannot be tested for within the sample of compliers jointly with an impact analysis since they rest on the same identification.

3 Matching Across the Treatment Criterion

The purpose of this paper is to suggest a means by which the relevant counterfactual surface, representing unexplained spatial heterogeneity in the absence of the treatment, can be estimated using non-compliers as a second control group. We

relax the outcomes assumption by modifying the selection assumption to say that the differences between compliers and non-compliers should not change over time in ways that are related to the incidence of the treatment.

The strong form of unconfoundedness needed to produce unbiased DID estimates when treatment assignment is non-random is $E(Y | X) \perp \tau(s)$. This will only be the case if $\phi \perp \tau(s)$. Our approach, instead of assuming this term to be orthogonal to the treatment is directly to estimate that projection using a second control group and to subtract it from the dependant variable. The resulting dependant variable is unconfounded by construction. This methodology replaces an assumption about the comparability of two regions with an assumption about the comparability of two populations.

We denote changes in outcomes among the group of non-compliers by Y_i^N . For this group, the true process determining these changes is $Y_i^N = \beta^N X_i^N + \phi_i^N + \epsilon_i$.

Our first assumption allows us to identify the unobserved shock at each point in space for compliers, and so to generate a dependant variable that is unconfounded:

$$\text{Spatial Assumption: } E(\phi | s_i) = E(\phi^N | s_i).$$

Because the X-vector contains a constant, this assumption states that unexplained spatial deviations from group mean outcomes are the same in expectations for a complier and non-complier located in the same place on the treatment criterion. In some applications (such as triple-differencing) a weaker form of this assumption, $E(\phi - \phi^N | T = 1) = E(\phi - \phi^N | T = 0)$, is sufficient.

Second, in order to use the parallel population as a spatial counterfactual, we

require that there be no spillover effects of the treatment, which is itself spatial.

$$\text{No Spillovers Assumption: } E(t_i \mid \omega_i = 0) = 0$$

Finally, we must make the standard matching assumption that the probability of compliance is bounded away from one and zero:

$$\text{Overlap: } 0 < Pr(\omega = 1 \mid s) < 1$$

For what is to follow, we maintain these assumptions throughout.

A straightforward test of the outcomes assumption which can be conducted is the False difference-in-differences (FDID). The FDID is a spurious impact regression which tests for a treatment effect in a group which did not receive the treatment, and for whom we assume there has been no spillover effect. It can be written as

$$\text{FDID: } Y_i = \beta X_i + \delta_{FDID} T_i + \mu_i \quad \forall \omega_i = 0.$$

This coefficient δ_{FDID} in this regression estimates unexplained heterogeneity across the space of the treatment criterion.

The Nearest Neighbor estimator uses the information present in these residuals to estimate location-specific counterfactual outcomes for every complier. First, we estimate a regression in the population of non-compliers using the X -vector but no treatment dummy, so as to leave spatial information in residuals: $Y_i^N = \beta^N X_i^N + \mu_i$. Here we confront potential spatial clustering in the observable variables: if the control variables are clustered across space in some way that proxies for unobserved shocks, then the estimated errors from this regression will be ‘too small’. We solve

this problem through a technique known as backfitting, or Gauss-Seidel Regression (Telser 1964). This iterative technique smooths the error terms across the relevant definition of space, subtracts the predicted spatial errors off of the dependant variables, and re-runs the regression until it converges. The β s which arise from this regression do not include any spatial component, and so if we predict outcomes using this set of β s then we recover an error term which includes the desired full degree of spatial variation.

The residuals from this regression will equal

$$\hat{\mu}_i^N = Y_i^N - \hat{\beta}^N X_i^N = \phi_i^N + \epsilon_i.$$

We denote the non-complier who is closest in the space of the treatment criterion to chooser i as i' . The Nearest-Neighbor (NN) regression is performed by matching *with replacement* to $\hat{\mu}_{i'}$, the residual of the nearest non-complier. We subtract this residual from the outcomes of compliers, and re-run the DID estimator:

$$\text{NN: } Y_i - \hat{\mu}_{i'} = \beta X_i + \delta_{NN} T_i + \mu_i \quad \forall \omega_i = 1.$$

We now present several propositions related to the use of these estimators.

Proposition 1: If the spatial distribution of compliers and non-compliers is identical, the False DID provides a test for the bias present in the DID estimator.

Proof: The ‘impact’ term in the false DID estimator will equal $P_{\tau(s)}(\phi^N)$. The nearest-neighbor residual equals $\phi_{i'}^N + \epsilon_{i'}$. Since this true error term $\epsilon_{i'}$ is i.i.d., $E(\hat{\mu}_{i'} | s_i) = E(\phi^N | s_i)$. By the spatial assumption, $E(\phi^N | s_i) = E(\phi | s_i) \quad \forall i$, and so $E(\hat{\mu}_{i'} | s_i) = E(\phi | s_i)$. The overlap assumption says that non-compliers

will exist in both the treatment and control, and because the spatial distributions are identical and there are no spillovers, $P_{\tau(s)}(\phi^N)$ will equal $P_{\tau(s)}(\phi)$. Letting $\bar{\delta}$ be the true TET, the DID estimate will be $\hat{\delta}_{DID} = \bar{\delta} + P_{\tau(s)}(\phi) = \bar{\delta} + \hat{\delta}_{FDID}$, and so $\hat{\delta}_{FDID} \neq 0 \Leftrightarrow \hat{\delta}_{DID} \neq \bar{\delta}$. \diamond

Proposition 2: If the spatial distribution of compliers and non-compliers is identical, then the NN estimator is unbiased even in the presence of unexplained spatial effects.

Proof: We can write Y_i as

$$Y_{1i} = \beta X_i + t_i + \phi_i + \epsilon_i, \text{ and}$$

$$Y_{0i} = \beta X_i + \phi_i + \epsilon_i.$$

From the proof of Proposition 1, $E(\hat{\mu}_{i'}) = E(\phi | s_i)$, and so

$$E(Y_i - \hat{\mu}_{i'}) = E(Y_i - E(\phi | s_i)) = \beta X_i + t_i + \phi_i - E(\phi | s_i) + \epsilon_i.$$

Let M_s be the orthogonalizing matrix which projects off of the space of the treatment criterion. The unexplained effect ϕ_i can be decomposed as $\phi_i = E(\phi | s_i) + M_s(\phi_i)$, and thus $\phi_i - E(\phi | s_i) = M_s(\phi_i)$. In expectations, the NN outcomes can then be rewritten as:

$$Y_{1i} - E(\phi | s_i) = \beta X_i + t_i + M_s(\phi_i) + \epsilon_i.$$

$$Y_{0i} - E(\phi | s_i) = \beta X_i + M_s(\phi_i) + \epsilon_i.$$

Crucially, since $\tau(s)$ defines a subset of the elements of s , $M_s(\phi) \perp T_i$. All of the unobservables which remain in the problem are either correlated only with the

observables, or are uncorrelated with both the observables and treatment. Thus letting $\bar{\beta}$ be the coefficient estimates from the DID estimate of the TET, when we estimate the NN regression

$Y_i - \hat{\mu}_{i'} = \beta X_i + \delta_{NN} T_i + \mu_i$, the expected estimates will equal:

$$\hat{\beta} = \bar{\beta} + P_X(M_s(\phi))$$

$$\hat{\mu}_i = \epsilon_i + (M_s(\phi) \perp X)$$

$$\hat{\delta}_{NN} = \bar{\delta},$$

and so although the coefficients on observables may be biased by unobservables, we have recovered an unbiased impact term. \diamond

Proposition 3: In the absence of unexplained spatial effects, both the DID and the NN are unbiased, but the DID has lower variance.

Proof: The DID estimator measures $\hat{\delta}_{DID} = \bar{\delta} + P_{\tau(s)}\phi$. Without spatial effects, $E(\phi | s_i) = E(\phi^N | s_i) \equiv 0 \forall i$. Thus $P_{\tau(s)}\phi = 0$ and so $\hat{\delta}_{DID} = \bar{\delta}$. The expected value of the dependant variable of the NN regression is $E(Y_i - E(\phi | s_i)) = Y_i$ without spatial effects, and so $\hat{\delta}_{NN} = \hat{\delta}_{DID} = \bar{\delta}$. Where there are n observations and K controls, and letting χ represent the block matrix $[XT]$, the estimated variance of the DID estimator will equal the square root of the $K+1$ th diagonal element of $\frac{\epsilon'\epsilon}{n-K}(\chi'\chi)^{-1}$. Without spatial effects, the conditional residual from the group of non-compliers will consist only of an error term, ϵ^N , which is i.i.d. and thus $\epsilon^N \perp \{X_i, T_i\}$. The NN regression can thus be rewritten as $Y = \beta X_i + \delta_{NN} T_i + \epsilon_i + \epsilon^N$, giving rise to an estimator with variance equal to $\frac{(\epsilon + \epsilon^N)(\epsilon + \epsilon^N)'}{n-K}(\chi'\chi)^{-1}$. Because ϵ, ϵ^N are

independent random variables, this variance will be strictly larger than that of the DID estimator. \diamond

Proposition 4: If the spatial distribution of compliers and non-compliers is identical, then the NN estimator is equal to the coefficient from the DID estimator minus the coefficient from the False DID estimator.

Proof: Flows directly from the proofs of Propositions 1 through 3:

$$\hat{\delta}_{DID} - \hat{\delta}_{FDID} = \bar{\delta} + P_{\tau(s)}(\phi) - P_{\tau(s)}(\phi) = \bar{\delta},$$

and since from Proposition 2, under these conditions the NN estimator is unbiased, it follows that

$$\hat{\delta}_{DID} - \hat{\delta}_{FDID} = \bar{\delta} = \hat{\delta}_{NN}.\diamond$$

Following from these proposition, if the FDID is insignificant, we have validated the conditions for the DID estimator, and the NN will be inefficient. An alternate way of testing for the presence of spatial effects is to bootstrap from the empirical marginal distribution of the errors in the parallel population, smoothing each set of errors across the treatment criterion. The envelope that contains 95 percent of these smoothed contours is a spatial confidence region, and if this interval contains zero at every location, we have rejected the presence of spatial heterogeneity. We should not proceed to utilize methods designed to remove spatial bias across the treatment criterion unless it has been demonstrated to exist.

A valuable feature of the NN estimator is its ability to recover unbiased treatment terms in the presence of endogenous placement. While Rosenzweig & Wolpin (1986) suggest a parametric method for estimating the process by which units are

endogenously selected into the treatment, we estimate the direct impact of endogenous placement rules on outcomes. The requirement is that the endogenous placement take a specific form; namely that *regions* on the treatment criterion were selected for some non-random reason to be exposed to the treatment. In this case, because it is reasonable to assume that non-compliers in the same space will have similar unexplained deviations from their own group means, we can back out an estimate of the counterfactual effects of this non-random program placement.

Proposition 5: An unbiased impact term can be estimated even in the presence of endogenous placement, as long as that placement takes the form $T_i = \tau(s_i)$.

Proof: Endogenous placement indicates that $E(Y_i | X_i, T_i = 0)$ is not a proper counterfactual for $E(Y_{0i} | X_i, T_i = 1)$, because growth in outcomes in the treatment differs from the control in some manner not eliminated by controls. This in turn implies that $E(\phi | T_i = 0) \neq E(\phi | T_i = 1)$. By the overlap assumption, however, there exist non-compliers in the control who allow us to estimate $E(\phi^N | s)$, and the treatment rule is spatially assigned through $\tau(s_i)$. In this case $E((Y_i - \hat{\mu}_{i'}) | X_i, \tau(s_i) = 0)$ is a proper counterfactual for $E((Y_{0i} - \hat{\mu}_{i'}) | X_i, \tau(s_i) = 1)$ because $E(\hat{\mu}_{i'}) = E(\phi | s_i)$ by assumption and $P_{\tau(s)}(Y_{0i} - E(\phi | s_i)) = 0$ by construction, and so the results of Propositions 2 cover this form of endogenous placement. \diamond

4 Bias from Imperfect Matching

Under the conditions outlined above, the nearest-neighbor approach will generate the same treatment effects as a triple-difference (3D) regression of the form

$$Y_i = \omega_i\beta_1X_i + (1 - \omega_i)\beta_2X_i + \alpha_1T_i + \alpha_2\omega_i + \delta_{3D}(T_i * \omega_i) + \mu_i \forall i.$$

Once we allow the spatial distribution of compliers and non-compliers to differ, the mapping across the treatment criterion is no longer one-to-one, causing the NN and 3D estimators to diverge. In practical applications we may see large differences between the 3D and NN estimates because it is the intersection of the spatial distribution of agents and the spatial distribution of unexplained effects which determines the projection into the treatment dummy in each population. The NN estimator will not be biased by, for example, a greater density of non-compliers than compliers in a region which experiences significant negative spatial effects, since it takes into account the precise location of compliers. The 3D regression makes no allowance for such differences, and hence mis-estimates the counterfactual.

Matching to nearby agents also forces us to think about the exact nature of the spatial correlation present in ϕ . One body of literature (Case 1991), (Conley 1999), (Conley & Topa 2002) estimates spatial correlations as an econometric phenomenon, remaining agnostic as to the reason for the observed patterns. Using the language of Manski (1993), it is difficult to maintain the assumption of no spillovers if spatial correlation arises as a result of endogenous effects, rather than exogenous or correlated effects. Hence the methodology presented here is most appropriate where adjacent agents experience common spatial effects, rather than where outcomes are transmitted directly from one agent to another.

A major advantage of the matching across the treatment criterion is that this rule usually maps across a low-dimensional space, providing us with a technically straightforward distance metric. Most matching estimators must resolve the problem of weighting the different dimensions across which the match is performed in order to define proximity. In practice the Mahalanobis metric (Mahalanobis 1930), which weights the square of the distance across each matching dimension by the inverse of the sample covariance matrix, is most commonly used. Proposition 2 demonstrates that using the NN estimator it is only the dimension of the treatment criterion across which matching must be performed in order to remove bias: a single dimension in the case of time or a qualification score, and two equally weighted dimensions in the case of physical space.

Proposition 6: Given a continuous, stochastic spatial distribution of compliers and non-compliers, the NN estimator may be biased, but as the number of non-compliers goes to infinity, the NN becomes unbiased.

Proof: Let N^N be the number of non-compliers. The location of each complier is an isolated point within the enveloping space of the treatment criterion, and hence is of measure zero. Thus when N^N is finite, the Euclidean distance $\|s_i - s_{i'}\| \neq 0$, and so even under our assumptions, $E(\hat{\mu}_{i'} | s_{i'}) = E(\phi | s_{i'}) \neq E(\phi | s_i)$. This implies that we are imperfectly removing spatial heterogeneity, and so the NN estimator will still contain spatial effects in the dimension of the treatment criterion and thus $P_{\tau(s)}(Y_{0i} - E(\phi | s_{i'})) \neq 0$. The direction of the bias is indeterminate *a priori*.

As $N^N \rightarrow \infty$, $\|s_i - s_{i'}\| \rightarrow 0$, so the quality of the match improves. In the limit $s_{i'} = s_i$, so $E(\hat{\mu}_{i'}) = E(\phi | s_i)$, whereupon the result of Proposition 2 holds. \diamond

In order to investigate the nature of the bias from imperfect spatial matching, we introduce some notation. The space of the treatment criterion is j -dimensional; let $\gamma_i = s_i - s_{i'}$ be a $j \times 1$ column vector representing the distance between i and i' in each element of s . The bias present in each individual match will equal $E(\phi | s_{i'}) - E(\phi | s_i)$, which we denote by $\Phi(s_i, \gamma_i)$, with $\Phi(s_i, 0) = 0$. This function returns the difference in the height of the smoothed residual surface among non-compliers as you move the distance and direction of γ_i away from s_i .

We can think of γ_i , the distance to the nearest neighbor match, as a vector of random variables with elements $\gamma_j \sim N(0, \sigma_j^2)$. Crucially, however, this does not imply that $E(\Phi(s_i, \gamma_i)) = 0$ without further assumptions on the shape of Φ . The reason flows from Jensen's Inequality; the residual matching surface $E(\phi | s)$ being convex implies that $\phi(s_i) = \phi(E(s_{i'})) < E(\phi(s_{i'}))$. Thus, if the residual surface $E(\phi | s)$ is globally convex (concave), then the expected value of the matching error will be positive (negative), leading to a downward (upward) bias because this residual estimate is subtracted off of the dependant variable. Following this reasoning, it will only be the case that $E(\gamma_i) = 0$ results in unbiased matching if the residual matching surface is planar in the dimensions of treatment criterion.

Abadie & Imbens (2004) implement a straightforward method for correcting for matching bias; they regress the outcome variable on a linear measure of the space over which the matching is conducted and so estimate the average effect of changes within that space. In our context their correction can be estimated by regressing

$$\hat{\mu}_i^N = \lambda s_i^N + \eta_i \quad \forall \omega_i = 0$$

where λ has j elements, which will return the linearized slopes of the residual surface across the elements of s . By correcting the match by the observed distance to the

nearest neighbor and by the estimated effect that this distance should have on the outcome, the matching estimator is made $N^{1/2}$ consistent.

This discussion suggests problems with implementing bias-correction which is linear in the dimension of s . While the bias correction improves the quality of each individual match and thus renders it more asymptotically efficient, if the residual surface is itself planar (meaning that the bias correction regression above is properly specified), in expectation there is no bias in the regression as long as $E(\gamma) = 0$. Bias resulting from non-linearities of ϕ across s are not corrected for, suggesting that the use of higher-order terms in the bias-correction regression may be warranted. As long as the mistakes made in matching are the same in the treatment and control regions, they will have no effect on impact estimates. What will generate bias in the estimate itself is if the convexity of the surface $E(\phi | s)$ differs between the treatment and control.

5 Matching to Multiple Units

The results of the previous sections provide two reasons that we might want to match to more than a single nearest neighbor. The first is from Proposition 6, which indicates that matches to a second control with a different spatial distribution and a finite number of agents will be imperfect. In this context it may be more reasonable to construct a localized average of the residual surface, since the nearest neighbor does not provide a point estimate of spatial effects which is unbiased anyway. Second, we are left with an empirical mean-variance tradeoff as we condition on space because $\hat{\mu}_{i'}$ is composed of the sum of $E(\phi^N | s_{i'})$, which we are interested in, and $\epsilon_{i'}$, which we are not. This would lead us to believe that we should

define some local neighborhood within which we average $\hat{\mu}_{i'}$ so as to remove the i.i.d. component ϵ_i . Such local averaging in a finite population of matches means, however, that in expectations the quality of the match will decrease as we increase the number of agents who define the ‘local’ spatial effect.

We investigate the nature of this tradeoff by varying the number of matches M among non-compliers which form the estimate $E(\phi | s_i)$. Giving an equal weighting to distance across the elements of γ , $d_{i,i'} = || s_i - s_{i'} ||$ is the Euclidean distance between i and i' . $Rank_{i,i'}(d_{i,i'})$ gives the rank of the distance of each non-complier i' from i (increasing in the distance), and $D_{i,i'}$ is an indicator function equal to 1 if $Rank_{i,i'}(d_{i,i'}) \leq M$ and 0 else. We then estimate spatial residuals as

$$E(\phi_i | s_i) = \frac{1}{M} \sum_{i'} D_{i,i'} \hat{\mu}_{i'}.$$

Proposition 7: Nearest-neighbor matching and the standard Difference-in-Differences estimators are both special cases of the Spatial Matching estimator.

Proof: With $M = 1$, $D_{i,i'} = 1$ only for the closest agent i' and so the estimator is trivially the nearest neighbor.

Once M has increased to equal the number of agents in the second control group, then the indicator function $D_{i,i'} = 1 \forall i'$. Because $\hat{\mu}_{i'}$ is a set of OLS residuals and we use the whole population across which this residual was estimated, $E(\phi_i | s) = \frac{1}{M} \sum_{i'} \hat{\mu}_{i'} \equiv 0 \forall i$. At this point, the dependant variable of the Spatial Matching estimator $Y_i - E(\phi | s) \equiv Y_i$, and so identically equals the DID estimator. \diamond

Given different spatial distributions and multiple matches, because we conduct matching with replacement, agents may be used as a match multiple times, requiring an upwards adjustment in the estimator of the variance. We can modify the

estimator provided in Abadie & Imbens (2004); if matching is conducted only one way (meaning that we match each complier to the residual of nearby non-compliers) then the variance of the spatial matching coefficients can be calculated the square roots of the corresponding diagonal elements of:

$$(X'X)^{-1} \frac{1}{N_c - K} \sum_i \left(\omega_i - (1 - \omega_i) \frac{K_M(i)}{M} \right)^2 \left(\frac{M}{M+1} (\mu_i - \hat{\mu}_{i'})^2 \right)$$

where N_c is the number of compliers, M is the number of nearest neighbors used as matches, and $K_M(i)$ is the number of times that each control agent is used as a match.

If we match in both directions, meaning that we also match each non-complier to nearby compliers, then we calculate variance using:

$$(X'X)^{-1} \frac{1}{N - K} \sum_i \left(\frac{1 + K_M(i)}{M} \right)^2 \left(\frac{M}{M+1} (\mu_i - \hat{\mu}_{i'})^2 \right).$$

6 Implementation on Simulated Data

In this section, we generate datasets in order to be able to investigate the empirical properties of the spatial matching estimator as we vary the number of matches. We generate a data set which allows us to examine the use of two control groups where physical spaces have been designated as treatment & control. Our data has two ‘cities’, meaning two random concentrations of agents. The treatment criterion is two-dimensional, and the cities lie at (2,0) and at (4,0). Agents are randomly assigned to be compliers or non-compliers, which means that the spatial assumption is satisfied by definition. Then, our space is divided into a treatment and control; compliers with $0 < X < 3$ are subjected to a treatment effect which increases

outcomes by one unit. In addition, spatial effects are generated as a function of the sine of agents' location in the X and Y dimension; Figure 1 illustrates the spatial shock by showing the residuals among the non-compliers (the random component ϵ_i of the residuals has been removed so as to represent the surface more clearly). Because there is no treatment effect within this group, the only source of spatial heterogeneity is the shock. The set of blue lines represent the border between the treatment and control regions, with the treatment region to the right. As a check, we perform the False DID regression among this group and find a significant negative 'impact' of $-.9108$, with a t-statistic of -3.9 .

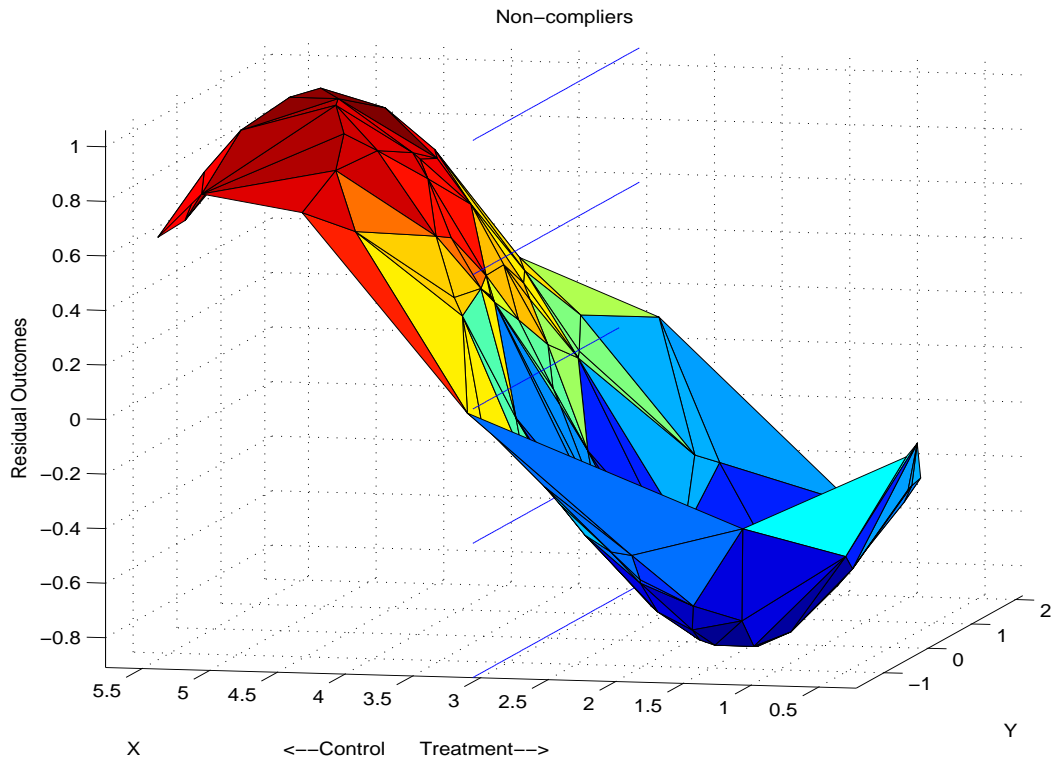


Figure 1: Smoothed residual surface for non-compliers

Figure 2 shows the backfitted residuals of outcomes (not including a treatment

dummy) among the compliers. The simulation has been set up so that the shock and the treatment effect counteract each other, with the treatment experienced in the area with the highest negative shock. The DID estimator will estimate impact by taking the mean difference in residuals between the treatment and control area.

There is no clear pattern in the residuals between the treatment and control areas; the DID will see no impact because $P_{\tau(s)}(\phi)$ almost exactly counteracts $E(t_i)$. The DID estimate is .0903 with a t-statistic of .4.

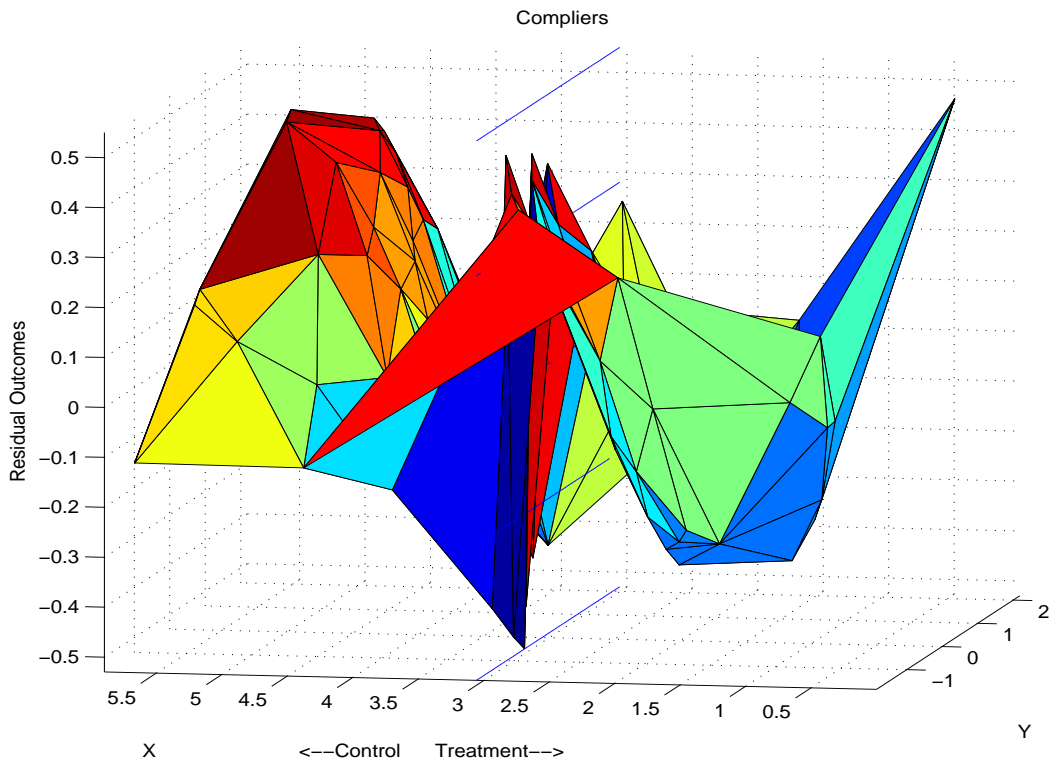


Figure 2: Smoothed residual surface for compliers

Because the NN estimator will find the closest non-complier and subtract that residual from the dependant variable, it will generate a spatial contour that is equivalent to the distance between the two contour sets, calculated at the location

of each complier. We can represent this graphically by Figure 3, which shows the contour that the NN estimator will use to estimate impact: namely, the mean difference in this residual surface between the treatment and control areas. The NN estimator will compare the average value of this difference between the treatment and control and control.

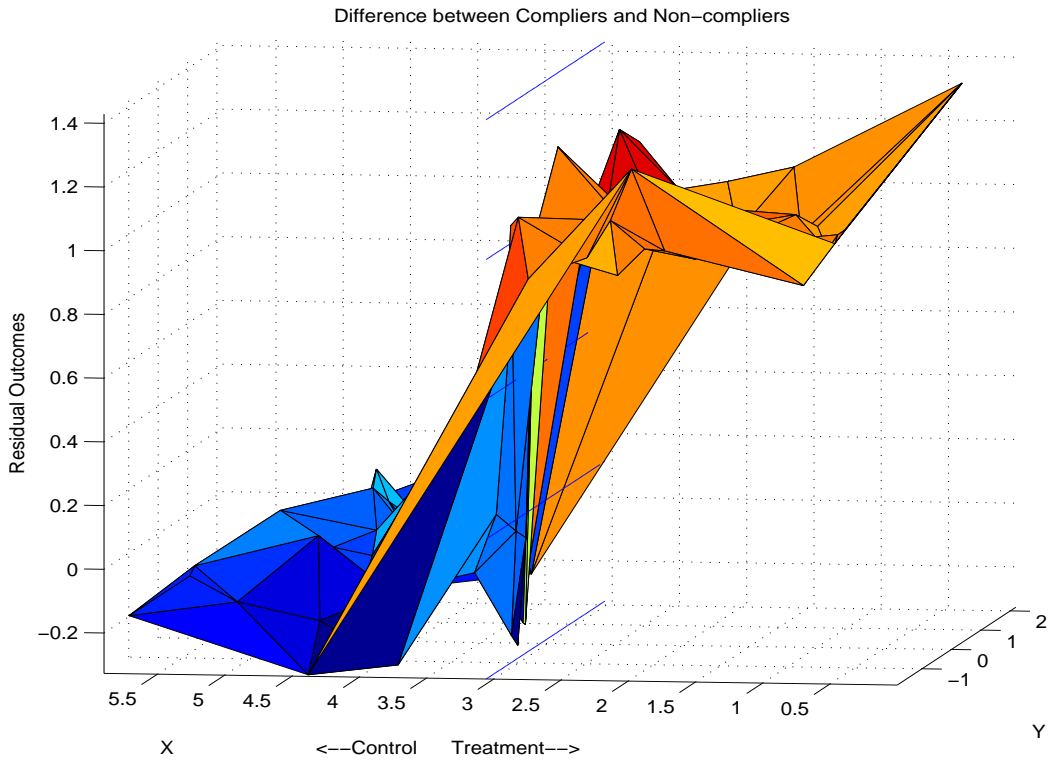


Figure 3: Distance between the residual surfaces for compliers and non-compliers

We generate 1,000 datasets where the location of agents, the identity of compliers, and the stochastic error term are resampled each time. Within each dataset we vary the number of nearest neighbors used in the matching, thus iterating between the NN and DID estimators. The impact estimates and confidence intervals are represented in Figure 4, and we see that the NN estimator achieves almost zero

bias, with the true impact of 1 well within its confidence interval. Once the number of matches equals the size of the population of non-compliers, the spatial matching estimate exactly equals the DID estimate.

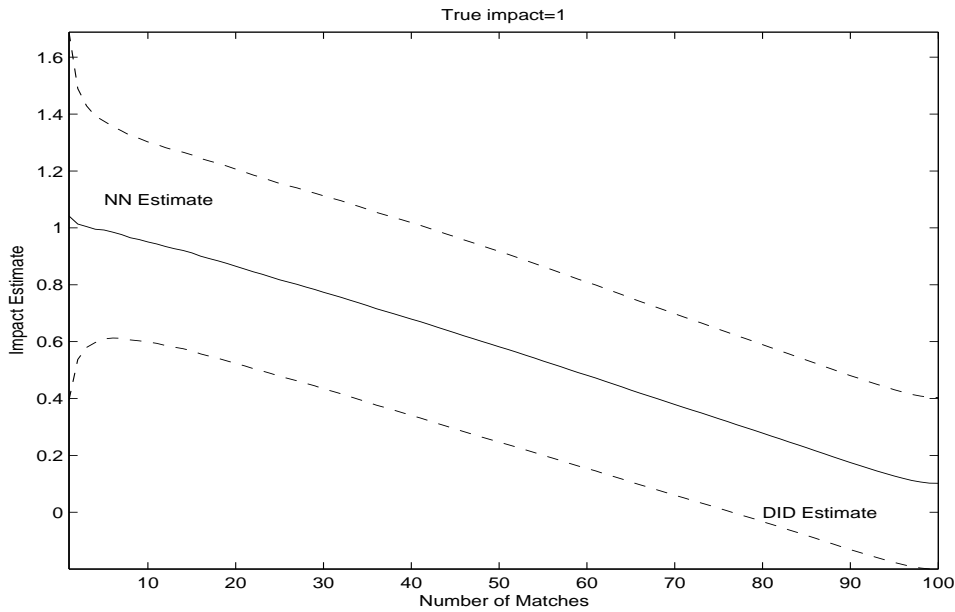


Figure 4: Change in impact estimate as # of matches increases

Figure 5 reports the variance of the estimator as we increase the number of matches, using the variance estimator for one-way matching defined in Section 5. The results in this simulation show that a great deal of variance reduction can be achieved by using multiple matches, but that this effect tails off between 10 and 20 matches.

Since the bias introduced is almost linear in the number of matches, the conclusion we draw from this simulation is that matching to several nearest neighbors is warranted in terms of the mean-variance tradeoffs, but that the use of more than 20 neighbors is not justified given our simulation.

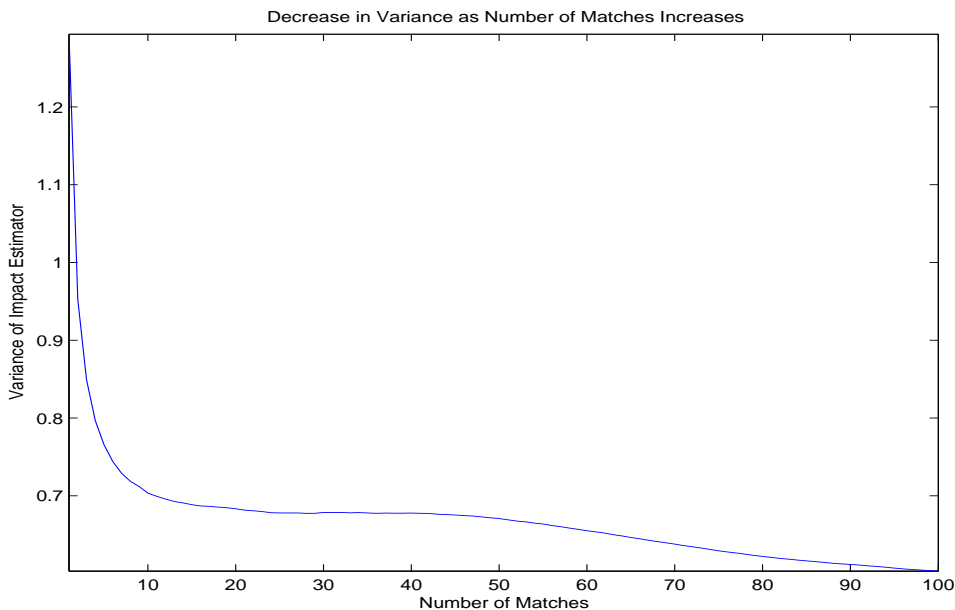


Figure 5: Reduction in estimator variance as # of matches increases

Having investigated the mean-variance tradeoff, we are left with the question of the empirical properties of the bias-reduction technique. The results presented above contain no bias-correction. Since the simulated shock surface is concave where positive and convex where negative, $E(\phi | s_i)$ is biased towards zero in both the treatment and the control. The fact that the NN estimate is slightly higher than 1 indicates that the shock surface in the treatment is more convex than the surface in the control is concave, leading us to underestimate $E(\phi | s_i)$ by a smaller amount in the control than the treatment, hence biasing the NN estimator upwards. In simulations where the shock structure was planar, NN estimates based on resampling large numbers of datasets were anyway unbiased. Where shock structures exhibited curvature and non-linear bias-correction techniques were implemented, the resultant bias in the NN estimator arises from differences in the concavity of the *differences*

between the actual shock structure and the regression specification used in the bias correction. Our conclusion is that the use of non-linear terms in the bias-correction is warranted, but it is unlikely that any reasonable functional form will remove matching bias given the complexity of the shock surfaces encountered in real data.

7 Some Applications of the Method

For an application of this technique to a spatial experiment performed by a micro-finance agency where different districts of Uganda were specified as treatment and controls, see McIntosh (2004).

Matching across a qualification criterion:

Imagine that a voluntary social welfare program was offered only to agents below a certain value on a scalar poverty index which can be calculated for all units. Compliance status is observed only in the treatment, and thus must be established in the control either through propensity methods or through surveys which elicit hypothetical choices. We cannot compare the treated to non-compliers in the treatment due to selection bias, nor can we compare the treated to compliers in the control because other, unmeasured components of wealth which are imperfectly removed by observed variables and correlated to outcomes are very likely to bias results. While a discontinuity design could be used to test for impacts in the immediate vicinity of the cutoff in the selection criterion, these are local treatment effects and are unlikely to provide a proper estimate of the TET across the distribution of the treatment rule. Spatial matching can be implemented by constructing the residual surface among non-compliers across the univariate dimension of the poverty index, subtracting off of compliers' outcomes the residual of the closest matches across the

index, and running a DID. This regression is identified by the assumption that the differences between compliers and non-compliers will be constant across the range of the poverty index.

Time as the Treatment Criterion:

In contexts where we cannot difference outcomes from within a given unit over time, we may wish to select time as the treatment criterion and match across it instead. As an example, a firm began offering stock options to highly qualified new employees over the past two years, and wishes to tell how they have effected subsequent performance. Again, we have two layers of selection into the program; in the first an employee must have qualified, and in the second they must have entered the company during the period during which the treatment was offered.

The two control groups would be the unqualified cohort that entered the firm at the same time, and the qualified cohort that entered the company in the years before the treatment was offered. The first suffers from selectivity bias, and the second suffers from any changes in the business climate over the years which might have caused employees's outcomes to shift. So, by estimating the residual surface across time (the more straightforward selection rule) among non-compliers, we can match the qualified entrants to this surface depending on time of entry to the firm, and compare the compliers with and without the treatment. This residual surface reveals how employee performance shifted in unexplained ways across time over the course of the study, and the estimate is robust to selection problems and to dynamic shocks.

Non-Parametric Implementation:

Instead of using regression to make treated and untreated compliers similar

across X , we can use matching. The matching estimator for the TET is

$$\frac{1}{N_c} \left[\sum_{\omega_i, T_i=1} (Y_i - \hat{Y}_i) \right],$$

where \hat{Y}_i is the average outcome over the M closest matches across X among compliers in the control, and N_c the number of compliers in the treatment. Correspondingly, the FDID matching estimator is

$$\frac{1}{N_{nc}} \left[\sum_{\omega_i=0, T_i=1} (Y_i - \hat{Y}_i) \right],$$

with \hat{Y}_i the match across X to non-compliers in the control, N_{nc} the number of non-compliers in the treatment, and asymptotic variance as described in Abadie & Imbens (2004).

The assumptions required to tie the information in the FDID to the compliers is altered substantially because the non-parametric control for X has deprived us of a residual surface. To write the spatial assumption directly in outcomes, we have $E(Y | s_{it}) = E(Y^N | s_{it})$, a rather draconian assumption over the spatial distribution of X_i across compliance status.

If we proceed with this assumption, for each complier in treatment and control we can calculate $\tilde{Y}_{i'}^N$, the counterfactual across s for complier i , by averaging outcomes among the M nearest non-compliers. The non-parametric spatial matching estimator for the TET will then be:

$$\frac{1}{N_c} \left[\sum_{\tau_i=1} (Y_i - \tilde{Y}_{i'}^N) - \sum_{\tau_i=0} (\hat{Y}_i - \tilde{Y}_{i'}^N) \right],$$

with $\tilde{Y}_{i'}^N$ representing the closest match across the treatment criterion for each of the matches closest to the treated across X among the compliers in the control. If treatment status in the control is not known, then we can proceed to use exactly the same estimator, however we now must add the standard selection assumption, because we are relying on the matching across X to eliminate selection effects.

If we wish to relax the strictures that the spatial assumption as written above places on differences in the distribution of X across s and ω , we can augment the matching rule across the treatment criterion with information in X . By using a Manalanobis distance metric that combines X and s to form the match to non-compliers, we alter the calculation of $\tilde{Y}_{i'}^N$ and $\tilde{Y}_{i'}^N$ above. The interpretation of this use of triple differencing is to ascribe as impact only differences between the treatment and control in excess of average outcomes among *similar* nearby units.

Since the method for establishing counterfactuals across the selection and treatment criterion is now the same, the only distinction between them lies in the order in which $\tilde{Y}_{i'}^N$ is calculated. In the regression case, it is clear that we should select as the treatment criterion the rule which is easily observable and across whose dimensions we expect complex, non-linear effects. When using only matching, there may be no clear reason to prefer one direction over another.

As an example, consider evaluating the impact on test scores of a school lunch program, to receive which students must be both in schools that qualify for the program and meet an individual qualification criterion. Assume both metrics are a known weighting of observable variables. The triple-difference can be constituted as either of the following questions:

1. Was the difference between qualified and unqualified children bigger in schools that did not subsidize lunches than those that did?

2. Was the difference between lunch and non-lunch schools smaller among qualified children than it was among unqualified children?

In either case, we will use the qualified students in non-lunch schools who are similar to the treated, and we will use the unqualified students in the lunch schools. Where we have little reason to prefer one direction to another, performing the matching both ways may provide a useful robustness check.

Implementation on Panel Data with Fixed Effects:

In a fixed-effects context, the only value for the second control is in estimating unexplained, idiosyncratic, time-varying counterfactuals. If the unexplained spatial heterogeneity is time-invariant, then the use of fixed effects in a standard panel estimation of the TET would be sufficient; nothing would be achieved by matching to the fixed-effect of the nearest non-complier and subtracting it off of the dependant variable. Similarly, if using both unit- and time-level fixed effects, any temporal shocks common to the whole population will fall out (and would not anyway bias a binary impact estimator). Our assumptions translate with no difficulty to panel data; no spillovers as $E(T_{it} | \omega_i = 0) = 0 \forall i, t$, the spatial assumption covers only time-varying, idiosyncratic component: $E(\phi | s_{it}) = E(\phi^N | s_{it})$ where $E(\phi^N | s_{it})$ is the spatial expectation of the residuals from

$$y_{it}^N = \beta X_{it}^N + \eta_i + \gamma_t + \mu_{it} \forall \omega_i = 0.$$

The spatial matching is conducted as above; the residuals are subtracted off of compliers' outcomes and the standard fixed effects DID is run:

$$y_{it} - \hat{\mu}_{i't} = \beta X_{it} + \delta_{FE} T_{it} + \eta_i + \gamma_t + \mu_{it} \forall \omega_i = 1.$$

This allows for an analogous relaxation of ignorability; the estimator permits a very general set of time-varying, unexplained phenomena as long as the differences in the way that compliers and non-compliers experience these phenomena is similar in treatment and control regions.

8 Conclusion

The counterfactual used to identify difference-in-difference treatment effects is only valid if the treatment is the sole source of unexplained heterogeneity in the space that defines the treatment criterion. In quasi-experiments, there is no reason to believe that this will in general be the case. This paper suggests techniques for using a second control population to estimate and remove the other sources of spatial heterogeneity. We replace the assumption that no unobserved differences exist between treatment and control with the assumption that differences between compliers and non-compliers will be the same across the treatment and control. The use of this technique allows for much simpler, less intrusive methodologies for conducting policy tests.

Many real-world examples of unexplained heterogeneity conform to the assumptions laid out in this paper. Inaccurate data collection by enumerators is likely to cluster spatially, in which case non-compliers provide us with a way of estimating and removing errors. Differences in rules or managerial quality across administrative units are likely to effect compliers and non-compliers in similar ways. Organizations often eschew randomization precisely because they want to expose a non-random subset of agents to a treatment; as long as this subset contain non-compliers, we have a way of backing out impact estimates when policies are endogenously placed.

The impact technique suggested here is data-intensive. We must have information over where agents are located on the treatment criterion, however for spatial quasi-experiments the advent of inexpensive GIS handsets has made collection of physical location data fairly routine. We must also have data on the parallel population, yet in many cases institutions collect data on compliers and non-compliers alike. Most problematic is the need to establish compliance status in the control region; this is not routinely a part of many studies. Where this information is missing, it can be predicted through the use of propensity scores at the cost of reinstating the standard assumption of selection on observables. The fact that compliance status in the control allows for an alternate form of identification suggests that this information may usefully be made a part of a quasi-experimental data collection strategy.

The payoff in the use of the spatial matching technique is that it allows us to utilize the extra counterfactual provided by two-tiered quasi-experiments. Given the broad range of applied problems which feature two layers of selection, these techniques can be used in many contexts. This form of triple-differencing is most applicable to a simple, observable treatment criterion across which we expect complex, non-linear phenomena. This suggests that quasi-experiments with straightforward treatment rules will be more amenable to analysis than those with complex rules such as a failed randomization. Given sufficient data, the methods outlined here allow us to test and then relax the assumptions underlying most forms of quasi-experimental identification and to recover robust impact estimates under a wide range of circumstances.

References

- Abadie, A. (2003). Semiparametric difference-in-differences estimators. Review of Economic Studies, forthcoming.
- Abadie, A. & Imbens, G. (2004). Large sample properties of matching estimators for average treatment effects. Unpublished working paper.
- Ahn, A. & Powell, J. (1993). Semi-parametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* (58), 3–29.
- Angrist, J. (1995). Conditioning on the probability of selection to control selection bias. *NBER Technical Working Paper 181*.
- Angrist, J. & Krueger, A. (1991). Does compulsory school attendance affect schooling and earnings?. *Quarterly Journal of Economics* **CVI**(4), 979–1014.
- Angrist, J., Imbens, G. & Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**(434), 444–472.
- Bertrand, M., Duflo, E. & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?. *Quarterly Journal of Economics*, forthcoming.
- Card, D. & Krueger, A. (1994). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review* **8**(4), 772–793.
- Case, A. (1991). Spatial patterns in household demand. *Econometrica* **59**(4), 953–965.
- Conley, T. (1999). Gmm estimation with cross-sectional dependence. *Journal of Econometrics* **92**(1), 1–45.
- Conley, T. & Topa, G. (2002). Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics* **17**(4), 304–327.
- Duflo, E. (2000). Child health and household resources in south africa: evidence from the old-age pension program. *American Economic Review* **90**(2), 393–398.

- Duflo, E. (2001). Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment. *American Economic Review* **91**(4), 795–813.
- Duflo, E. & Kremer, M. (2003). Use of randomization in the evaluation of development effectiveness. Paper for the World Bank Operations Evaluation Department (OED).
- Gruber, J. (1994). The incidence of mandated maternity benefits. *The American Economic Review* **84**(3), 622–641.
- Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* **47**(1), 153–162.
- Heckman, J., Ichimura, J. & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies* **64**, 605–654.
- Imbens, G., Rubin, D. & Sacerdote, B. (2000). Estimating the effect of unearned income on labor earnings, savings, and consumption: evidence from a survey of lottery players. *American Economic Review* **90**(2), 778–794.
- Kremer, M. (2003). Randomized evaluations of educational programs in developing countries: some lessons. *AEA Papers and Proceedings* **93**(2), 102–106.
- LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review* **76**(4), 604–620.
- Mahalanobis, P. (1930). A statistical study of certain biometric measurements from sweden. *Biometrika* **22**(1), 94–108.
- Manski, C. (1993). Identification of endogenous social effects: the reflection problem. *Review of Economic Studies* **60**(3), 531–542.
- McIntosh, C. (2004). Estimating treatment effects from spatial policy experiments: An application to ugandan microfinance. Unpublished working paper.
- Morduch, J. (1998). Does microfinance really help the poor? new evidence from flagship programs in bangladesh. Unpublished working paper.

- Pitt, M. & Khandker, S. (1998). The impact of group-based credit programs on poor households in bangladesh. does the gender of participants matter?. *Journal of Political Economy* **106**(5), 958–995.
- Ravallion, M., Galasso, E., Lazo, T. & Philipp, E. (2002). Do workfare participants recover quickly from retrenchment?. World Bank Working Paper.
- Rosenbaum, P. (1982). The role of a second control group in an observational study. *Statistical Science* **2**(3), 292–316.
- Rosenbaum, P. & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
- Rosenbaum, P. & Rubin, D. (1984). Reducing bias in observational studies using subclassifications on the propensity score. *Journal of the American Statistical Association* **79**(387), 516–524.
- Rosenzweig, M. & Wolpin, K. (1986). Evaluating the effects of optimally distributed public programs: Child health and family planning intervention. *The American Economic Review* **76**(3), 470–482.
- Telser, L. (1964). Iterative estimation of a set of linear regression equations. *Journal of the American Statistical Association* **59**(307), 845–862.