ESTIMATING TREATMENT EFFECTS FROM SPATIAL POLICY EXPERIMENTS: AN APPLICATION TO UGANDAN MICROFINANCE

Craig McIntosh*

Abstract—This paper demonstrates a method for estimating treatment effects in spatial tests, utilizing a second control group to measure unexplained spatial phenomena. The technique is implemented on two innovations in Ugandan microfinance, and we measure the ways in which concurrent shocks such as an Ebola outbreak and a contentious presidential election altered outcomes differentially across regions. By correcting for this spatial heterogeneity, we measure the impact of the policies; a program that increased borrowers' control over the terms of their loans improved outcomes, while the results of a program that bundled health insurance into the lending contract were more mixed.

I. Introduction

SPATIAL policy tests are subject to biases that randomized tests are not (Kremer, 2003), and yet they are likely to remain a permanent feature of the quasi-experimental evidence available to researchers. States, school districts, administrative branches, and the rule of law all describe physical spaces, and so designating physical regions as treatment and control will continue to be the least intrusive way for many kinds of institutions to experiment (Card & Krueger, 1994). The trouble with this approach is that if outcomes are being determined by some process which is itself spatial and we fail to control for this process, the resulting spatial heterogeneity is indistinguishable from a treatment effect in a spatial policy experiment and so it biases difference-in-differences (DID) estimators. Many policies, however, either are voluntary or have eligibility requirements. In this case we have a group of agents who did not choose or qualify to receive the treatment, and so there will be untreated units even within the treatment region.

The value of this parallel untreated group is that it allows us to examine the differences between the treatment and control region in the absence of a treatment effect (Rosenbaum, 1982; Gruber, 1994). Under assumptions laid out below, ineligible groups experience similar relative spatial heterogeneity to what would have been observed in the eligible group had none of these units received the treatment. Specifically, after running the difference-in-difference regression (without a treatment term) in this nonparticipant group, we suggest mapping the residuals across space as a means to measure whether there are significant unexplained phenomena that differ across the treatment and control regions. These residuals form a spatial surface that allows us to answer the following question: What is the unexplained

outcome that we would expect to see for a qualified unit in a given location if there had not been a spatial policy test? This contour measures exactly the quantity that biases the difference-in-differences, and so it is subtracted off of outcomes in the group that *did* qualify for (or choose) the treatment. The result is a fine-grained form of triple-differencing that measures the treatment effect on the treated, accounting for localized shocks or misspecification in the estimating equation. This not only makes spatial impact analysis robust to unexplained shocks and endogenous placement, but allows us to measure a kind of impact invisible to standard techniques: the ability of a treatment to insulate agents against shocks.

We apply the technique to two new policies introduced by FINCA Uganda, the country's largest microfinance institution. The first innovation allows clients to change the repayment frequency of their loans, and the second bundles a health insurance package into the lending contract. A straightforward spatial testing strategy was implemented, wherein whole administrative branches of the institution were designated as treatment and control regions. A standard analysis of a voluntary, spatial test would either compare treated and untreated regions to estimate the intention-to-treat effect, or identify those likely to choose the program in the control in order to estimate the treatment effect on the treated. Here, we dealt with the selection problem by conducting mock elections in the control region in order to establish which groups would have chosen the programs had they been offered them. With this extra degree of identification, we can use differences between choosers and non-choosers as well as the differences between those offered and not offered the program to identify the impact of the two treatments.

The results of the analysis have several implications for best practices in microfinance. Weekly repayment is widely perceived to be central to the low default rates observed in microfinance (Morduch, 1999), yet this system entails large transaction costs for both lender and borrower. We find that moving from repaying loans every week to repaying every two weeks causes none of the predicted negative effects, instead causing dropout to fall by 10 percentage points (a 40% reduction) and triggering a slight improvement in repayment performance. The implication of this study is that when clients are allowed to decide whether their fellow members are capable of repaying reliably biweekly, information revealed by the joint-liability mechanism allows transaction costs to be reduced without pushing up default. The entire impact of the health insurance program arose because of insulation against shocks, rather than from directly observable changes in outcomes. New client enrollment

increased sharply, but client composition shifted for the worse in insured groups, suggesting that the asymmetric information problems in credit and insurance markets need to be solved separately rather than using a single joint-liability contract (Stiglitz, 1990), (Ghatak, 1999) to solve them both.

II.   Bias in the Difference-in-Differences

The joint-liability group is the unit of observation. The change in outcomes is denoted by Y_i, which is determined by some function $Y_i = f(X_i, s_i, Z_i) + \varepsilon_i$. We use simple changes in outcomes rather than a panel because most of our control data is cross-sectional, and also because using before-after differences eliminates the concerns raised in Bertrand, Duflo, and Mullainathan (2004) over the inconsistency of standard errors in a panel DID when outcomes are serially correlated. X_i is a vector of observable control variables, s_i is group i's location in physical space, and Z_i is a vector of unobservables. Treatment status is a binary variable indicating whether a group was in the region in which the treatment was offered, and is denoted by T_i. Our application involves a spatial treatment-control strategy, which implies a mapping from location s_i to the treatment status; we denote this mapping by τ, so $\tau(s_i) = T_i \in \{0, 1\}$. In addition, the treatment is voluntary, and so groups may either choose or not choose to accept the treatment, a decision denoted by ω_i (this can represent either choice or eligibility to participate; the terms are used interchangeably here). Receipt of the treatment, then, indicates that $T_i = \omega_i = 1$.

In potential outcomes notation, Y_{1i} represents changes in outcomes for a treated group, and Y_{0i} represents the counterfactual untreated outcome for the same group. Without loss of generality, we can think of the treatment effect as additive, so that $Y_{1i} = Y_{0i} + t(X_i, s_i, Z_i)$. Since both choices and location determine treatment status in this kind of test, we have a two-tiered selection problem. The rule by which agents select into $\omega_i = 0$ and $\omega_i = 1$ is called the selection criterion, and $T_i = 0$ and $T_i = 1$ the treatment criterion; both rules are, in general, endogenous. Estimating impact in such two-tiered tests requires us to account for variation in the determinants of Y_i across both criteria.

We investigate the bias due to spatial heterogeneity by explicitly denoting the portion of outcomes that is systematic but not explained by a linear function b of observables as $\phi(X_i, s_i, Z_i)$. Therefore the equations determining outcomes are

$$Y_{0i} = bX_i + \phi(X_i, s_i, Z_i) + \varepsilon_i,$$

and

$$Y_{1i} = bX_i + \phi(X_i, s_i, Z_i) + t(X_i, s_i, Z_i) + \varepsilon_i.$$

The difference-in-differences estimator for the average treatment effect on the treated (ATT) when we observe treatment choices in the control is

$$Y_i = \beta_1 X_i + \delta_1 T_i + \mu_i \ \forall \ \omega_i = 1.$$

Although the ability to compare choosers to choosers has removed selection bias from this estimate, the DID will be biased in a spatial test unless unexplained spatial effects are identical in the treatment and control regions. In the DID estimator, $\phi(X_i, s_i, Z_i)$ is an omitted variable which is orthogonal to X by definition. Consequently, it will project into T but not into X, and we estimate impact to be $\hat{\delta}_1 = E(t) + P_T(\phi)$. Since $T_i = \tau(s_i)$ is a binary variable defined over space, the projection of ϕ into T will equal $E(\phi|\tau(s) = 1) - E(\phi|\tau(s) = 0)$.

If we consider the non-choosers of a program to represent a counterfactual for the spatial variation in outcomes that would have been present in the absence of a treatment, then we are provided with a spatial surface that forms a natural baseline from which to estimate the true effects of the treatment. A simple way to proceed would be to subtract off locally averaged outcomes among non-choosers from the outcomes of each chooser. Upon further reflection, however, we see that it is only unexplained spatial effects which will bias estimates, and so we should not include in this local outcome estimate anything that is directly explained by observables. This suggests that it is the residuals among non-choosers in a local area that possess the most information about unexplained spatial effects. The reason for the concern with shocks *which vary across space* is that it is only unexplained effects which have some spatial component that will project into the treatment and cause bias. The direct implication is that if we are able to estimate residual spatial effects, we can recover an impact term free of both selection bias and spatial bias.

We now introduce two assumptions that allow us to estimate and utilize these spatial effects. The first is essentially an extension of the typical selection assumption:

Spatial Assumption: $E(\phi|s_i, \omega_i = 1)$
$= E(\phi|s_i, \omega_i = 0).$

Because we include a constant term in the X vector, $E(\phi|\omega_i = 1) = E(\phi|\omega_i = 0) \equiv 0$. So the spatial assumption says that a chooser and a non-chooser located at the same place should experience unexplained spatial effects, relative to their own group mean, that are the same.

The second required assumption is that there be no spillover effects of the treatment from choosers to non-choosers within this same region.

No Spillovers Assumption: $E(t_i|\omega_i = 0) = 0.$

These assumptions suggest two straightforward ways of using our assumptions. The first is the OLS analogy to the general estimators described in Rosenbaum (1982) for testing whether an unobserved covariate differs across the treatment criterion. We can estimate the false difference-in-differences (FDID) regression among non-choosers;

$$Y_i = \beta_2 X_i + \gamma_1 T_i + \mu_i \,\forall\, \omega_i = 0.$$

$\hat{\gamma}_1$ will equal $E(P_T(\phi)|\omega=0)$, which is a spurious treatment effect arising from unexplained spatial phenomena. If the spatial dispersion of choosers and non-choosers is identical, the spatial assumption implies that

$$E(P_T(\phi)|\omega = 0) = E(P_T(\phi)|\omega = 1),$$

and so $\hat{\gamma}_1$ is a precise measure of the spatial bias among choosers. In this case, we can recover an unbiased estimate of the ATT by subtracting the coefficient from the FDID off of the DID, which is analogous to conducting a triple-difference regression:

$$Y_i = \omega_i \beta_1 X_i + (1 - \omega_i)\beta_2 X_i + \alpha_1 T_i + \alpha_2 \omega_i$$
$$+ \delta_2(T_i \times \omega_i) + \mu_i.$$

In other words, $\hat{\delta}_2 = \hat{\delta}_1 - \hat{\gamma}_1$.

In the absence of spatial effects, the estimate from the triple-difference will be the same as from a DID, but less efficient. Since the FDID will also be insignificant in such cases, this suggests the use of the FDID as a preliminary test for the presence of spatial effects. Only in the presence of a significant false impact in the FDID should we proceed to utilize methods designed to remove spatial bias.

In recent years, an increasing number of researchers have been attempting to overcome the limitations of difference-in-difference techniques by some form of triple-differencing. Examples include the comparison of pretreatment growth to posttreatment growth (McKenzie & Mookherjee, 2003; Banerjee, Duflo, & Munshi, 2003), or comparison of current and past members of a training program to a matched sample of nonparticipants (Ravallion et al., 2002). Closer to the application presented here are efforts to compare eligible and ineligible agents in regions that offer and do not offer a program. Hammermesh and Trejo (2000) use a dummy-variable approach to compare the differences between women and men in California versus nonwestern states before and after overtime benefits had been extended to men in California. Kugler (2005) follows Gruber (1994) in estimating a triple-difference by interacting dummies for being eligible, for being in a treatment region, and for being in the posttreatment time period. Morduch (1998) presents a set of tables that demonstrate the mean difference between households that are eligible and ineligible for microfinance, across villages that do and do not have microfinance institutions. To date this literature, however, has used the second control group simply to calculate lump sum spatial effects through regional means or dummy variables.

So, under what circumstances will a dummy variable approach to triple-differencing form the correct counterfactual? The purpose of the second control group is to give an estimate of unexplained changes in outcomes across the space that defines the treatment criterion. Hence if the

dummies represent very small units such as villages (Pitt & Khandker, 1998), if spatial effects happen to line up neatly with the dummy used for the treatment unit, or if the spatial distribution of eligible and ineligible is identical, then a dummy-variable approach to triple-differencing recovers the correct counterfactual. In many cases, however, the spatial unit of treatment is large, spatial effects are localized, and eligible and ineligible units do not have exactly the same spatial distribution. In this case, the average spatial effect among the eligible within a given region differs from the ineligible, and so even under our assumptions, the triple-difference estimator is inconsistent. This problem cannot in general be overcome through fine-grained spatial dummies because they will be collinear with the treatment term.

III.   The Spatial Matching Estimator

The contribution of this paper is to show that, if we have precise location information, we can recover a consistent estimate of unexplained spatial effects even with highly localized shocks. Using the non-chooser (or ineligible) units, we can first construct a spatial surface that represents this heterogeneity. We can then utilize techniques designed for propensity-score matching across the selection criterion (Rosenbaum & Rubin, 1983; Angrist, 1995; Dehejia & Wahba, 2002) in order to match agents for whom $\omega_i = 1$ to the counterfactual surface estimated among $\omega_i = 0$, using physical location. By taking the value of this surface at the exact location of each chooser (or eligible) unit, we correctly form the counterfactual regardless of differences in the exact spatial distribution across the groups, as long as they share a common support.

The closest analogy to this is the "regression-adjusted conditional difference-in-differences matching estimator" discussed in Heckman, Ichimura, and Todd (1997). They invoke an exclusion restriction to match across the selection criterion, and subtract off counterfactual fitted values from a linear regression. The resulting estimator performs well when compared with experimental identification of the true impact of a job training program. The estimator suggested here is able to add an additional level of differencing because we observe selection status in treatment and control alike, and so we can eliminate selection effects and match instead across the treatment criterion. Seen in this light, we can recast our assumptions as the exclusion restriction required to use physical location as an instrument for treatment status in comparing the difference-in-differences across choosers and non-choosers.

To operationalize the method, we first run the FDID regression among $\omega_i = 0$, except that the treatment term is omitted so that all spatial effects remain in the residuals. We then calculate a local average using linear distance around each chooser i of the residuals among non-choosers, denoted by $\hat{\mu}_{i'}$. For each i and number of nearest neighbors M, we have a set $J(M, i)$ of nearest neighbors.

The local average is denoted by

$$\tilde{\mu}_i = \frac{1}{M} \sum_{i' \in J(M,i)} \hat{\mu}_{i'}.$$

This is the value at chooser's location s_i of the residual surface among non-choosers i', smoothed by M. This value can then be subtracted off the dependent variable and the difference-in-differences is run on the choosers.

So, the spatial matching estimator estimates the ATT as follows:

$$Y_i - \tilde{\mu}_i = \beta_3 X_i + \delta_3 T_i + \mu_i \; \forall \; \omega_i = 1.$$

Because it is a mean, for small M, $E(\tilde{\mu}_i \mid s_i) = E(\hat{\mu}_{i'} \mid s_i)$. In the absence of the treatment, $\hat{\mu}_i = Y_i - \hat{\beta} X_i = \phi_i + \varepsilon_i$. Epsilon is an i.i.d. error term and we assume that the untreated receive no treatment effect, and therefore the spatial expectations of residuals contain only spatial effects: $E(\mu_{i'} \mid s_i) = E(\phi_{i'} \mid s_i) \; \forall \; s$. Applying the spatial assumption to this expression shows that $E(\tilde{\mu}_i \; \forall \; s_i) = E(\phi_i \; \forall \; s_i)$.

This implies that $Y_i - \tilde{\mu}_i$ is an outcome that allows for unbiased estimation of the ATT even when $P_\tau(\phi) \neq 0$. The counterfactual dependent variable among choosers is determined by

$$Y_{0i} - E(\phi \mid s_i) = bX_i + \phi_i - E(\phi \mid s_i) + \varepsilon_i.$$

The term $\phi_i - E(\phi \mid s_i)$ represents the unobserved effect minus the unexplained effect conditional upon being in that location. Crucially, the resulting term is itself orthogonal to space, and so it can be written as $\phi_i - E(\phi \mid s_i) \equiv M_s(\phi_i)$: this is the residual vector that remains when ϕ has been projected into s, or the component of the outcome vector that projects off of both X and s.

The unexplained residual is orthogonal to X by definition, and is orthogonal to the treatment by construction, since

$$M_s(\phi) \perp [\tau(s), \beta X].$$

Because $P_T(M_s(\phi_i)) \equiv 0$, when we use our modified dependent variable to run a DID regression

$$Y_i - \tilde{\mu}_i = \beta X_i + \delta_3 T_i + \mu_i \; \forall \; \omega_i = 1,$$

$E(\hat{\delta}_3) = E(t) + P_T(M_s(\phi_i)) = E(t)$, giving an unbiased estimate of the ATT.

One attractive feature of this approach is that it allows for consistent estimation of treatment effects even under endogenous placement. If a program is (intentionally or unintentionally) placed in a region that has inexplicably different rates of growth from the control, this is usually fatal to our ability to estimate impact. The spatial assumption, however, will apply as long as it is the case that the non-choosers of the program share the outcome differences seen among nearby choosers. In other words, if treatment regions contain nonrandom draws of administrative quality, economic prospects, or ethnicity, this only biases spatial impact assessments because it introduces spatial heterogeneity. To the extent, then, that this spatial effect is *not* related to selection into the program in a way that varies across space, the effects of the endogenous placement will be removed by the use of spatial matching to the second control.

Although there is no way to directly estimate the intention-to-treat effect when spatial effects are present, we can back out an estimate from the ATT. The reason is that we have already assumed no spillover effects, which means that $E(t_i \mid \omega_i = 0) = 0$, and so the intention-to-treat effect, can be estimated by

$$ITE = \hat{\delta}_3 \times Pr(\omega = 1).$$

where $Pr(\omega = 1)$ is the fraction of choosers. To the effect that the treatment effect itself is spatial, we estimate the ATT over the control space.

A technical problem arises in the use of OLS residuals because of the potential clustering of the Xs in space, which will cause observables to proxy for underlying spatial effects. If we do nothing to address this issue, then the spatial assumption in reality requires a similar spatial distribution of the Xs between choosers and non-choosers, which is an unpalatable assumption. To address the issue, we use an orthogonalizing procedure known as Gauss-Seidel regression (Telser, 1964) to "backfit" the residual vector among non-choosers. This procedure uses a spatial smoother to extract spatial information present in the residuals, subtracts it off of the dependent variable, and iterates until there is no spatial component remaining. The resultant βs are used to predict a vector of residuals that contain the full degree of spatial variation. This procedure is deemed to make the spatial assumption more realistic, and so it is implemented throughout the paper.

There are alternative ways to use the two controls that do not require backfitting. One would be to use raw residual surfaces and to match choosers to non-choosers across both location and other observable variables. This solution, however, does not correct for the misestimation of the residual surface itself and so still requires an assumption over the similarity of the spatial distribution of the Xs between choosers and non-choosers. An alternative to this would be directly to match outcomes across both location and observable variables. Either of these methods requires that we use a more complex weighting matrix (typically the Mahalanobis metric) to measure multidimensional distances. The nonparametric approach also deprives us of the residual surface altogether, and insofar as this surface is of direct interest in the application at hand, we proceed to estimate it using backfitting.

We draw on the extensive literature developed to solve matching problems across the selection criterion. We utilize a bias-correction technique suggested by Abadie and Imbens (2006) to estimate the following OLS regression:

$$\hat{\mu}_{i'} = \lambda s_{i'} + \eta_i \; \forall \; \omega_i = 0.$$

The estimated shock at location s_i is then adjusted according to the distance and direction between s_i and $s_{i'}$ and the average slope of the residuals across longitude and latitude measured by λ. For the case where $M = 1$ with i'' as i's nearest neighbor, we would have

$$\tilde{\mu}_i = \hat{\mu}_{i''} + \hat{\lambda}(s_i - s_{i''}).$$

This removes bias that would result from imperfect spatial matching when the residuals have a clear tilt across s.

The second technique taken from that paper is to establish a double counterfactual; we estimate not only what all of the choosers would have looked like in the absence of the treatment, but also what all of the non-choosers would have looked like had they received the treatment. Because we backfit residuals for both choosers and non-choosers, we can estimate a residual vector among choosers that contains the treatment effect. The regression used to estimate impact can be written in block matrix form as follows:

$$\begin{vmatrix} Y^C - \tilde{\mu}_i^{NC} \\ Y^{NC} - \tilde{\mu}_i^{C} \end{vmatrix} = \begin{vmatrix} X^C & 0 & T \\ 0 & X^{NC} & -T \end{vmatrix} \begin{vmatrix} \beta^C \\ \beta^{NC} \\ \delta \end{vmatrix} + \mu.$$

Finally, we can vary the number of agents used to form the estimate of the local spatial residual. Matching to the single nearest residual generates a low-bias, high-variance estimate. Matching to increasing numbers of agents is analogous to smoothing the residual surface. Because the sum of the residuals among non-choosers is identically zero, as we match to a larger number of agents the spatial matching estimator approaches the DID estimator because the residual surface goes toward zero everywhere. In our empirics we use a single match, which is the minimum-bias estimate, as well as sixteen matches, to investigate how altering the size of the local area affects the impact estimates.

IV. Two Policy Innovations

FINCA Uganda is among the oldest formal microfinance institutions in the country and is one of the largest and best established in Africa. Their standard lending product utilizes a group-lending methodology, wherein the thirty members of "village banks" are jointly liable for each others' loans. Loans are made almost exclusively to women. There is no formal screening of new clients, so membership in groups is constrained only by the selection imposed by current clients on members of their community for whom they will accept liability. Loans begin at $50, and subsequent increases are based on fixed multiples of clients' savings determined by the client's grade, which in turn is based on repayment and attendance at weekly meetings. The village banking methodology, using large groups and frequent meetings, in general targets the poorest market segment served by major microfinance organizations. Wealthier clients with rapidly growing businesses may eventually "graduate" to higher-tier lenders who use smaller groups and offer more custom-

ized financial services (McIntosh, de Janvry, & Sadoulet, 2005). The standard loan has a sixteen-week cycle, and clients pay 4% per month flat interest (87% effective). Each client is covered by a life insurance policy, whose premiums are included in the interest payments. FINCA Uganda now has more than 25,000 clients in 1,000 village banks, spread over most of the conflict-free parts of Uganda.

At the beginning of 2000 FINCA Uganda began offering two new policies, as shown in figure 1. The programs were offered simultaneously, but to different parts of the country. The health treatment was offered in downtown Kampala and in the area around Masaka, while the biweekly treatment was offered in the east and north of the country. This leaves the central region, with the exception of the capital itself, as the control. The biweekly treatment allowed groups to elect (by a unanimous vote) to change from the standard practice of repaying loans weekly, to repaying them every other week instead. While it is clear that making fewer payments is preferable to clients, the widespread perception that frequent repayment is central to preventing default has made institutions slow to offer this service, and indeed makes members of joint-liability groups reluctant to accept it when it is offered. Frequent repayment is likely to be particularly important to delinquency rates for the poor (de Aghion & Morduch, 2000). Thus, the primary concern for those groups that switch to biweekly repayment is whether the reliability of repayment drops. Since clients in groups that have switched to biweekly payment have lower transaction costs, dropout should fall. The same effect would cause new client enrollment to rise, however biweekly groups may also be more *selective* in admitting new clients: thus the effect on the fraction of new clients is ambiguous. Weekly repayment also places a very tight cash-flow constraint on client businesses; from anecdotal evidence, the amount that can be repaid in the worst typical week often determines what clients are willing to borrow. Thus, we hypothesize that biweekly repayment will cause loan volume to increase.

The second new policy offered a voluntary health insurance package to village banking clients and their families. The package costs roughly $13 per four-month cycle and covers the client, spouse, and four dependents against routine medical expenses. As an attempt to control adverse selection, in order for groups to be eligible for the health insurance package, more than 60% of the individuals in any village banking group had to enroll (in effect, causing the adverse selection of unhealthy *groups* rather than unhealthy *individuals*). Uganda is an environment characterized by high mortality and morbidity; consequently, medical costs can constitute a major burden for poor families. It is thus almost certainly the case that FINCA will be more attractive to new clients as it adds the health insurance option.

In practice, the insurance program suffered from major cost overruns; premia calculated on the basis of preinsurance health expenditures turned out to be far too low to

FIGURE 1.—TREATMENT-CONTROL REGIMES CIRCLE = CONTROL, SQUARE = HEALTH TREATMENT OFFERED, TRIANGLE = FLEXIBILITY TREATMENT OFFERED.
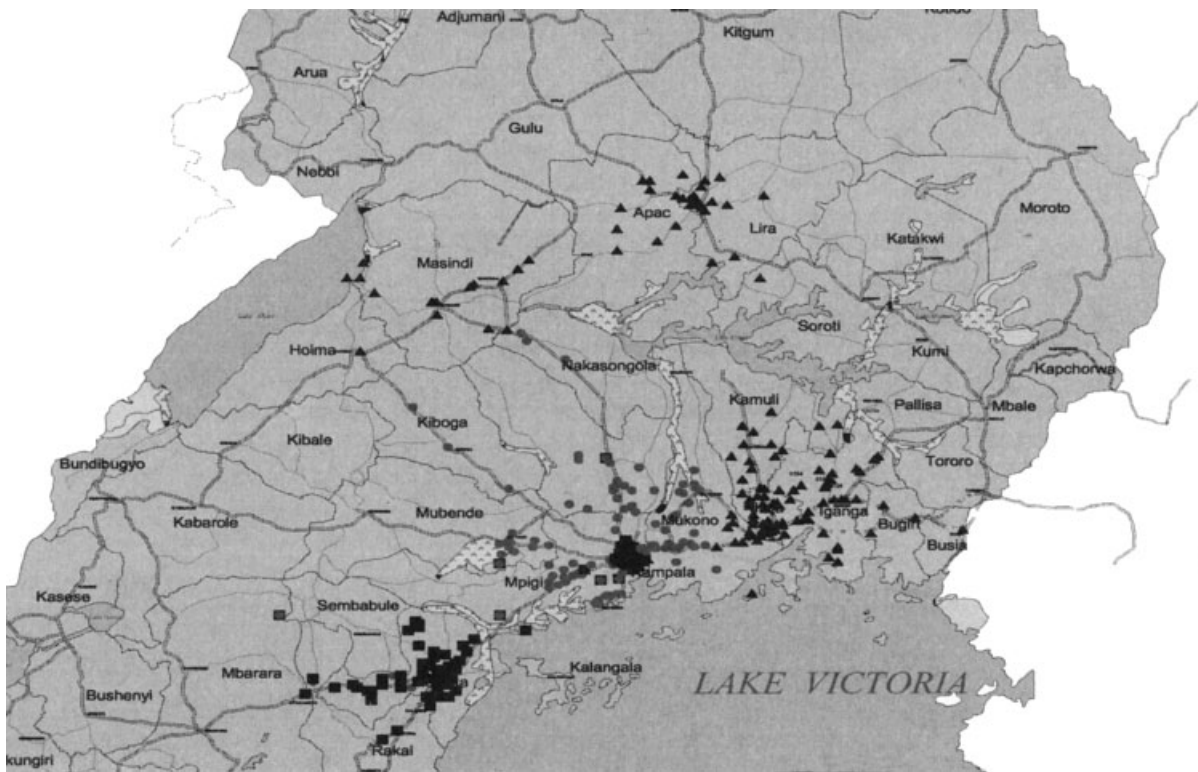
cover costs among the insured. Given the relative destitution of this client base, there is good reason to think that this is a result of more than the standard moral hazard story; if agents were truly constrained in their preinsurance health expenditures, then the problem they face is fundamentally not one of insurance (which relates to smoothing) but one of limited income. Hence, even in the presence of a well-designed program, we might expect to encounter cost over-runs. These were likely exacerbated by the fact that the 60% voting rule does very little to control adverse selection and does nothing to utilize information possessed by clients about their fellow group members.

Given the somewhat chaotic time period over which these data were collected, we have prior reason to believe that spatial shocks may be present in outcomes. An Ebola outbreak occurred in October of 2000; the disease was contained to the northern part of the country but the town of Masindi, a part of the biweekly treatment, was affected. Figure 5 indicates that there was a significant increase in credit demand in the region around Masindi, potentially as a result of the epidemic. In addition, Uganda experienced political unrest surrounding the presidential election in March of 2001 as Kizza Besigye mounted an unexpectedly strong challenge to the incumbent president Yoweri Museveni. This episode led to insecurity and some minor rioting in the capital (the health treatment region), potentially explaining the dramatic drop in credit demand there. It is also conceivable that pork-barrel spending in the run-up to the election was apportioned in a nonrandom way across elec-

toral districts, creating differences-in-differences in the spatial pattern of business outcomes. While we cannot easily determine the causality behind such spatial differences, we confirm that they are present, and their mere presence is enough to bias the difference-in-differences.

The data are taken from the accounts of FINCA Uganda and from surveys conducted at the individual and the group level during the test. The number of usable groups, restricted to those that were in existence prior to the beginning of the test and that are in the relevant areas, is 450. Data were missing from roughly 10% of the surveys, and groups that participated in either treatment were eliminated from the analysis of the other, reducing the overall number of observations to roughly 400 for each program. Table 1 lists the outcome and explanatory variables used in the analysis. The chooser/non-chooser status of groups in control areas was established through a mock election as a part of the group survey; voting rules were applied in the same way that they had been in the treatment; for example, 60% was required for health insurance, and unanimity for biweekly repayment.

In table 2, we see the number of groups that were offered and that chose the treatments. In each case there is a slightly higher fraction of choosers in the control than in the treatment, possibly indicating that hypothetical decisions are more easily taken than real decisions; to the extent that this bias exists it creates problems for any assumption across the selection criterion. If some of the controls who said they would accept the treatment would not actually do so when

TABLE 1.—VARIABLES USED IN THE ANALYSIS

Outcome Variables:
1. Dropout (percentage that took a loan last cycle and do not return)
2. New clients (percentage of clients starting this cycle that are new)
3. Grades (repayment performance and meeting attendance)
4. Average loans in a village bank
5. Average savings in a village bank

Control Variables:
1. Loan cycle number (e.g., loans taken by this village bank)
2. Ethnic homogeneity of the group
3. Borrower's perception of their local business climate
4. A dummy equal to 1 if the village bank is in a rural area
5. The average number of children in clients' households
6. The average number of nonworking adults in clients' households
7. The share of clients in a village bank that own their own homes
8. Does the village bank conduct a Rotating Savings and Credit Association (ROSCA)
9. Did the group preexist in some form prior to formation of the village bank

offered it, then we are comparing the treated to an insufficiently selected group of controls, and so some selection bias may still be present in our estimates. We perform *t*-tests of the differences in take-up rates across treatment and

control regions, and find that the difference is significant for the insurance treatment and insignificant for the biweekly treatment.

Table 3 allows us to make conjectures as to the direction of this selection bias. Here we compare means of the exogenous and (pretreatment) endogenous variables for both programs. We see evidence of program placement bias in both cases through differences across the treatment criterion. Selection effects are present in the exogenous variables for both programs; choosers of the biweekly program were more urban and optimistic, and choosers of the insurance program came from older groups and had larger families. There is evidence of selection bias in outcomes for the biweekly treatment, where choosers are better repayers and less likely to drop out (consistent with de Aghion & Morduch, 2000). If we had significant differences in take-up rates in combination with selection effects in pretreatment outcomes, we would have reason to think that both double- and triple-difference estimators would contain selection bias. In this case, because the take-up rate is not different where a selection effect in outcomes exists (biweekly), and no selection effect in outcomes exists where take-up rates are different (insurance), there is no obvious selection problem. Any bias that exists would be in the direction of the effects seen at the bottom of table 3.

V. Investigating Spatial Effects

We begin our empirical analysis by investigating spatial heterogeneity, using the variation in loan volume among non-choosers of the biweekly treatment as a test case. To visualize these spatial effects, we create a moving average composed of the twelve nearest neighbors for each outcome. In figure 2, we plot the spatial surface of deviations from the average loan size. The view is from the southwest, over Lake Victoria, and the defining feature of this image is the

TABLE 2.—NUMBERS OF GROUPS IN TREATMENT AND SELECTION CATEGORIES

Biweek:	Treatment Criterion		
	$T = 0$	$T = 1$	Total
Selection ω = 0	102	168	270
Criterion ω = 1	51	59	110
Total	153	227	380

Insurance:	Treatment Criterion		
	$T = 0$	$T = 1$	Total
Selection ω = 0	140	91	231
Criterion ω = 1	99	18	117
Total	239	109	348

TABLE 3.—*t*-TESTS OF DIFFERENCES IN MEANS ACROSS THE TREATMENT AND SELECTION CRITERIA

	Treatment Criterion (Mean difference between treated & untreated)		Selection Criterion (Mean difference between chooser & non-chooser)	
	Biweek	Insurance	Biweek	Insurance
Exogenous:				
Located in village	0.0334	−0.1285**	−0.1418**	−0.0073
Cycle no. of group	0.6935**	−0.0859	0.2689	0.8595**
Group conducts ROSCA	−0.0106	−0.0133	−0.0106	0.1042**
Ethnic heterogeneity	0.0881	0.0867	−0.0279	−0.1762**
Perceived econ. climate	−0.0502	0.2185**	0.2885**	−0.1048
No. of children in HH.	0.1088	0.1643	−0.1019	0.1996**
No. of adults in HH.	0.1739**	−0.0829	−0.0345	0.2842**
Own home?	0.0096	0.0237	0.0081	−0.0231
Preexisting group	0.0111	−0.0199	−0.0606	−0.0157
Endogenous: (Pretreatment)				
% dropout	0.1581	−0.3488	−0.8554**	−0.4072
% new clients	−0.5844	−0.1820	−0.1757	−0.7421
Av. ind. loans, USD	−11.712**	12.0221**	−1.874	−6.2119
Av. ind. savings, USD	−12.0557**	16.4137**	−0.4804	−0.4287
Grades, A = 4, D = 1	−0.0418	0.0884	0.3845**	0.0018

** = Difference in means significant at 95% level.

FIGURE 2.—SPATIAL DISTRIBUTION OF LOAN SIZES
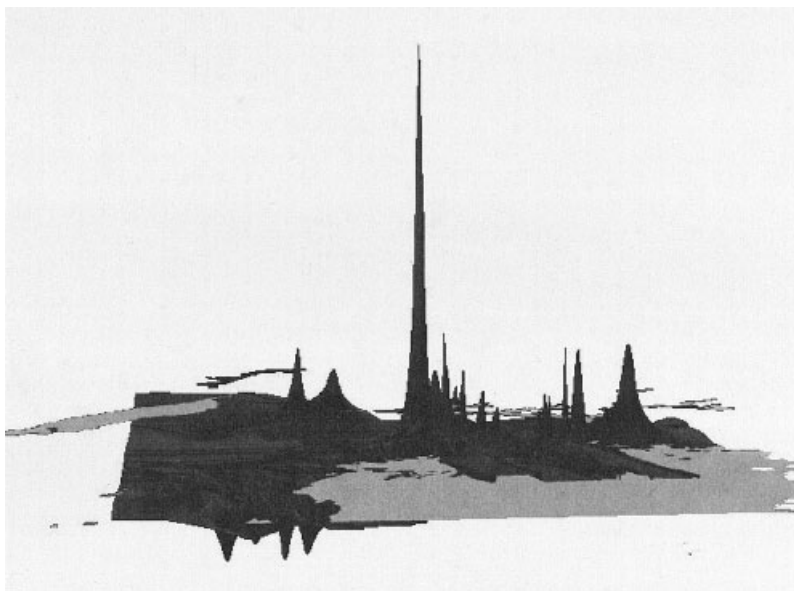
large spike over the capital Kampala, located on the northwest side of the lake. Obviously, conditions in this city induce clients to take vastly larger loans on average than anywhere else in the country.

Figure 3 is the same except that we now use *changes* in loan volume. At first glance, this picture appears to be the inverse of the previous, and again we see the enormous difference between the capital and the rest of the country. It would appear that changes in loans are inversely related to size, however the correlation between loan size and loan growth in this subsample is positive (0.014). Thus it is more likely that the strong negative spike under Kampala is related to shocks surrounding the 2001 presidential elections, which were disproportionately felt in the capital.

It is unsurprising to find large spatial differences in raw rates of change. The use of a DID estimator in spatial tests, however, relies crucially on the assumption that the control variables will remove all spatial heterogeneity other than that imposed by the test itself. Consequently, if we examine the spatial surface of the *residuals* among non-choosers (who have no treatment effect), under the assumptions of the DID we should see a flat surface, meaning that no significant spatial information remains in the residuals. The FDID, indeed, is measuring the difference in the average height of this surface between the treatment and control regions. In figure 4 we carry out the exercise of plotting residuals of loan changes, and we see that our control variables have achieved almost nothing in terms of remov-

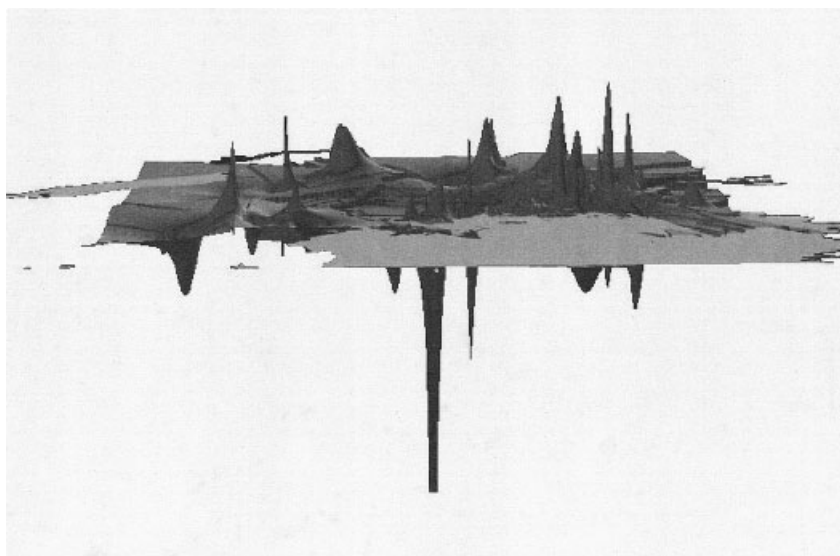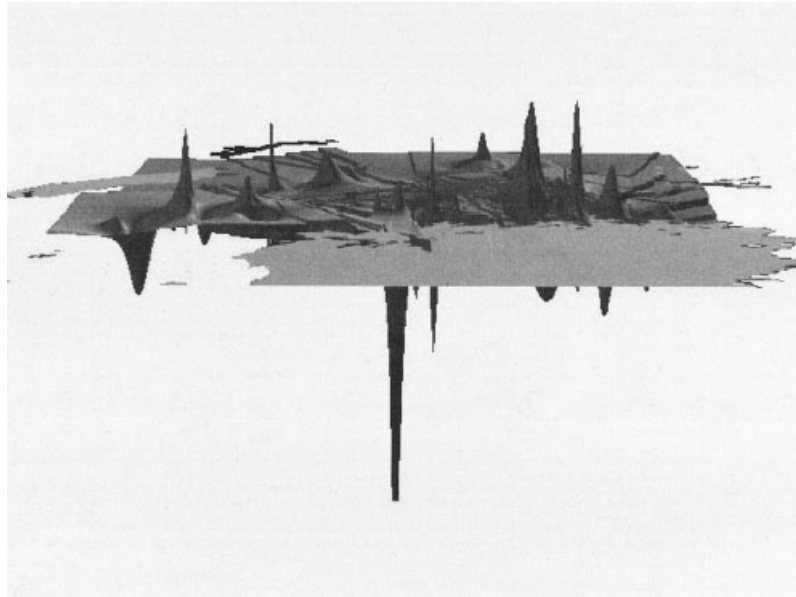FIGURE 3.—SPATIAL DISTRIBUTION OF LOAN CHANGES

FIGURE 4.—SPATIAL DISTRIBUTION OF LOAN CHANGE RESIDUALS

ing the kinds of spatial heterogeneity that will bias the DID; in fact, figures 3 and 4 appear virtually identical.

Under the two assumptions presented in section III, the surface in figure 4 is the counterfactual surface for choosers, showing what the spatial heterogeneity in residuals would have been in the absence of the test. The presence of such extreme spatial heterogeneity in this picture, coupled with the fact that the capital is a part of the control for the biweekly treatment, implies that the DID estimate of impact will be sharply biased upward. The lack of difference between the contours in figures 3 and 4 indicates that the DID estimate would be almost as biased as an impact estimate arrived at by simply subtracting the raw changes in the control from the raw changes in the treatment.

We can verify the problems inherent to the use of the DID regression in a context with this kind of spatial shocks in several other ways. First, we compose a simple table of the means and t-tests of differences for loan size, loan changes, and residuals of loans taken by choosers and non-choosers of the program across the treatment and the control (see table 4).

While these residuals are not backfitted, this table tells us several things. First, the use of differencing can exacerbate bias in a spatial DID if the source of bias is spatial shocks. We see this from the fact that the levels of loans are not significantly different between the treatment and control in the non-choosers, but loan changes and residuals are significantly different in both groups. Second, we see the inefficacy of this vector of controls at removing spatial heterogeneity, despite the fact that they are fairly standard and were chosen from large groups of potential controls for their explanatory power. Third, under the assumptions in section III, we see that the DID cannot be unbiased when the counterfactual surface (residuals among non-choosers) shows such strong differences between treatment and control. In this example we see that far from being equal to 0, the difference between treatment and control among non-choosers is stronger than among choosers, leading us to believe that the DID coefficient may actually have the wrong sign.

To investigate the surprising failure of our vector of controls at removing spatial heterogeneity, we regressed both the residuals and the changes in loan size on a set of district-level dummies. In the former regression the R^2 was 0.073 and in the latter 0.065, meaning that according to this rough measure our control variables removed only 10% of the spatial heterogeneity. This result may not generalize to other studies, but it is troubling that this fairly typical battery of controls does such a poor job of removing the effects that would bias a spatial treatment/control setup.

To get a clearer picture of the magnitude of these effects, we bootstrap the residual surface in figure 4. We would like

TABLE 4.—COMPARING DIFFERENCES ACROSS TREATMENT AND CONTROL

	Non-choosers			Choosers		
	T	C	t-test	T	C	t-test
Loans	113.9	119.8	1.10	117.4.9	110.4	−0.96
Loan changes	39.0	14.4	−3.57**	36.0	15.2	−2.28**
Residuals	8.8	−14.5	−3.97**	7.06	−8.18	−1.99**

** = difference in means significant at 95% level.

FIGURE 5.—SHOCKS AT THE LOCATIONS OF THE CHOOSERS X = INSIGNIFICANT SHOCK, AND AN ARROW INDICATES THE DIRECTION OF SIGNIFICANT SHOCKS.
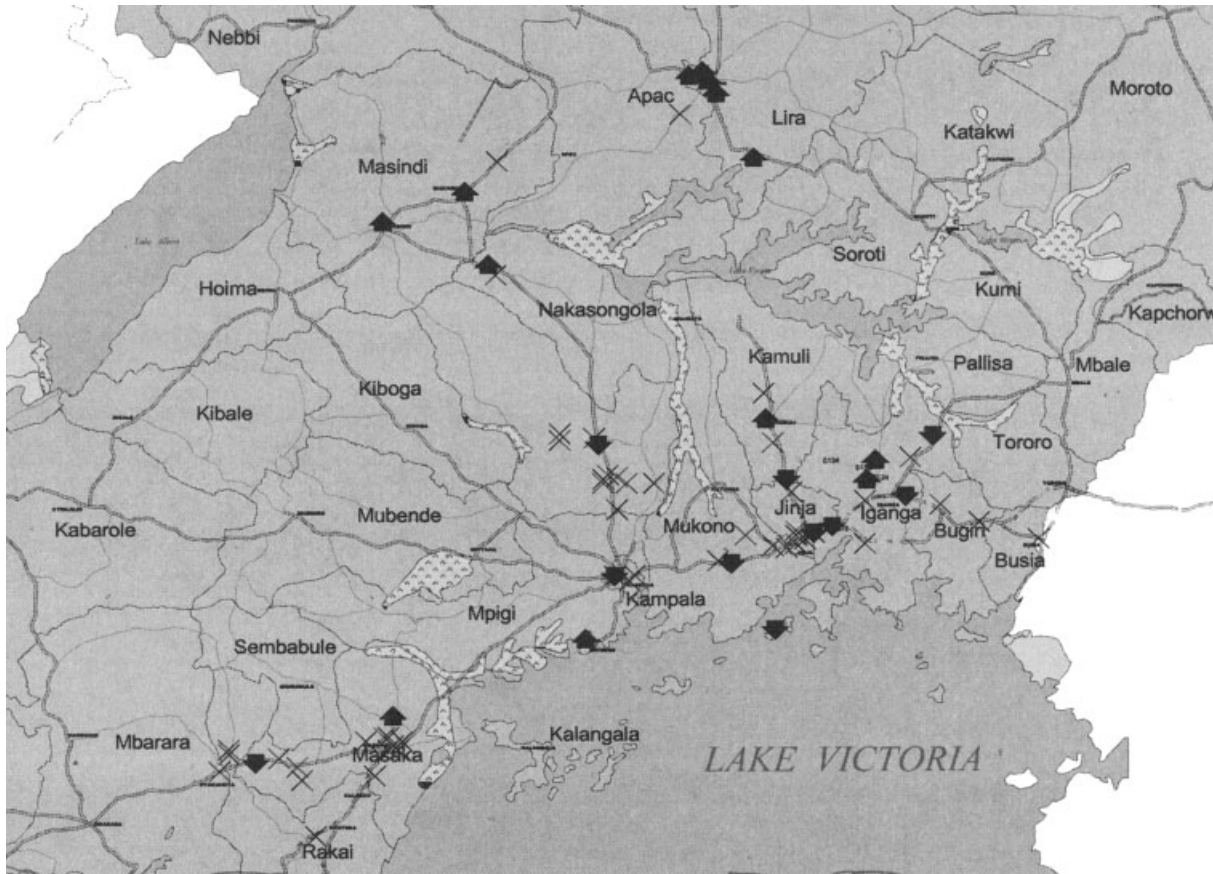
to know not only whether the spatial effects observed are significant, but also whether there are significant effects at the locations of the choosers. This is achieved by matching each chooser to the residual that is estimated at the location of the closest non-chooser. We then sample with replacement from the empirical marginal distribution of the residuals, randomizing over location, and so we bootstrap data sets that are spatially i.i.d. by construction. We then run a spatial smoother (a Gaussian smoother with a bandwidth of 0.2 standard deviations) over these bootstrapped data sets and select the envelope that contains 95% of the smoothed surfaces. Finally, we smooth our observed nearest-neighbor residuals with the same smoother, and then compare these pointwise smoothed outcomes with the pointwise smoothed i.i.d. outcomes. Any observations that lie below the confidence region have experienced a significant negative residual effect, and any observations above have experienced a positive effect.

Figure 5 shows the locations of all of the choosers. Those observations that experienced insignificant shocks are denoted with an X, and those with significant shocks are denoted by an arrow pointing in the direction of the effect. It is clear by inspection that the area around and to the west of Kampala, which is the control region, displays more significantly negative residuals, while residuals in the north-

ern and eastern sections of the country, the treatment, are disproportionately positive. While the treatment and the control contain six groups each that experienced significant negative effects, the treatment region contains twelve groups that experienced positive effects, while the control region contains only two. This is consistent with our prediction of a significant positive treatment effect in the false DID. We thus proceed confident that we do indeed experience significant spatial heterogeneity in our data, and thus the standard DID estimates will be biased (in this case upwards).

VI.   Regression Results

The first column in table 5 shows the DID impact estimates for five outcomes under each of the treatments. The programs appear to have few significant effects, with the exception of an increase in loan volume under the biweekly treatment, significant at the 95% level. These will be unbiased estimates of impact only in the case that no other unexplained spatial effects coincide with the treatment and control regions. To test this, we run the false DID for the same five outcomes in both programs, and these results are reported in column 2. We reject the absence of such effects among the group of non-choosers in three out of ten cases,

TABLE 5.—REGRESSION RESULTS

	Difference-in-Differences	False Difference-in-Differences	Spatial Matching $M = 1$	Spatial Matching $M = 16$
Biweekly:				
# obs:	110	270	380	380
% new clients	−5.3331	0.2736	−2.4280	−0.4867
	(−1.0223)	(0.0890)	(−0.7521)	(−0.1818)
% dropout	−5.0194	3.0012	−12.9544**	−9.5690**
	(−1.2097)	(1.2474)	(−4.7618)	(−4.5962)
Savings (USD)	−4.9995	−9.3477	13.3042	8.0221
	(−1.1161)	(−1.0105)	(1.4288)	(1.1692)
Loans (USD)	17.5045**	24.6750**	2.8716	−3.8805
	(2.0508)	(4.0369)	(0.4455)	(−0.7796)
Grades (A = 4, D = 1)	−0.0674	−0.0403	0.2468**	0.0743
	(−0.4842)	(−0.4418)	(2.4766)	(0.9671)
Insurance:				
# obs:	117	231	348	348
% new clients	0.2263	−3.6515	5.2698**	3.8479**
	(0.0549)	(−1.6399)	(2.1727)	(1.9612)
% dropout	2.5454	4.6773*	6.5050**	−1.5620
	(0.5080)	(1.6934)	(2.0535)	(−0.6322)
Savings (USD)	0.4842	12.9900	−10.4638	−12.9625
	(0.0787)	(1.2227)	(−1.2494)	(−1.6056)
Loans (USD)	−4.6143	−14.7495**	4.2655	9.0664*
	(−0.4485)	(2.1586)	(0.5716)	(1.6566)
Grades (A = 4, D = 1)	0.2357	0.0195	−0.0598	−0.0243
	(1.3368)	(0.2104)	(−0.5621)	(−0.2832)

t-statistics in parentheses, * = 90% significance, ** = 95% significance.

implying that significant unexplained spatial phenomena are present. Therefore, the DID estimator will be biased in this spatial test, and only because the non-choosers provide us with a counterfactual for these spatial effects are we able to back out an unbiased estimate of the impact of the programs.

In the cases where the false treatment effect observed is similar to the treatment effect measured in the DID, we suspect that the DID is picking up only spatial bias, and there is unlikely to be any real treatment effect. The increase in loan volume in the biweekly program measured by the false DID, for example, is similar to that observed in column 1, implying that all of the "treatment effect" picked up by the DID is bias. Where we have significant false effects in the FDID that differ from those found in the DID, we suspect that the bias in the DID may be masking real underlying treatment effects.

If the spatial mapping of choosers and non-choosers were identical, then we could get an unbiased treatment estimate by subtracting the coefficient from the second column off of the first. The spatial matching estimator differs from this triple-difference estimate because of differences in the spatial distributions of the two populations. Column 3 gives the estimate of the spatial matching estimator using only the nearest neighbor as a match, and column 4 uses the sixteen nearest neighbors. Column 4 thus uses a larger area to define the "local" shock.

These estimates demonstrate no impact of the biweekly program on loan volume, despite the significant positive DID estimate. We do see significant effects for the biweekly program in strongly decreased dropout and grades that are somewhat higher. Because these effects are similar to the selection effects seen in table 3, and the take-up rate in the treatment was weakly lower, some concern remains that these results could be contaminated by selection bias. While we have not seen the jump in loan volumes that was predicted by theory, we do observe that these groups that have tailored products to the clients have managed to become significantly more attractive to current members, and so have improved retention. The costs of lending are, obviously, dramatically lower when the credit officers halve the number of meetings they are required to oversee, and so if this transition leaves loan volumes unchanged and actually improves client quality, it suggests that extending the option of biweekly repayment to the rest of the country will both increase the sustainability of the institution and help to tailor products toward the needs of groups.

Intriguingly, every outcome for which the FDID is significant and the DID is not is found under the insurance program. The obvious intuitive interpretation is that the insurance program has insulated participants against a shock observed in an uninsured population, which is precisely what we would expect from such a treatment. New client enrollment is weakly lower, dropout weakly higher, and loans significantly lower for clients within areas offered the insurance treatment who did not choose it. For those who chose and received the treatment, however, we see no evidence of these spatial differences, leading us to infer that the treatment has played a causal role in insulating them.

The insurance program has been strongly effective in attracting new clients, or at least in avoiding a decrease in enrollment that was otherwise occurring. A study of individual clients who have joined groups after they received

TABLE 6.—REGRESSION RESULTS USING ROBUST VARIANCE ESTIMATOR

	Spatial Matching $M = 1$	Spatial Matching $M = 16$
Biweekly:		
# obs:	380	380
% new clients	−2.4280	−0.4867
	(−0.3955)	(−0.0867)
% dropout	−12.9544*	−9.5690**
	(−1.7730)	(−2.0624)
Savings (USD)	13.3042	8.0221
	(0.8912)	(0.6094)
Loans (USD)	2.8716	−3.8805
	(0.1842)	(−0.3860)
Grades (A = 4, D = 1)	0.2468	0.0743
	(0.9105)	(0.4311)
Insurance:		
# obs:	348	348
% new clients	5.2698	6.8398*
	(0.9924)	(1.6609)
% dropout	6.5050	−1.5620
	(0.7695)	(−0.2858)
Savings (USD)	−10.4638	−12.9625
	(−1.0575)	(−0.8947)
Loans (USD)	4.2655	9.0664
	(0.3129)	(0.8164)
Grades (A = 4, D = 1)	−0.0598	−0.0243
	(−0.1838)	(−0.1304)

t-statistics in parentheses, * = 90% significance, ** = 95% significance.

the insurance product shows that these clients are low-quality borrowers (low grades, small loans, small growth in loans) and hence are probably participating in FINCA only to get access to the insurance product. The fact that average default does not increase indicates that the screening process continues to be effective in preventing the admission of deadbeat clients. It is also important, however, that insuring clients against health expenditure risk creates no worsening in repayment, implying that illness is not driving default even in this high-mortality environment. The weak decrease in savings and increase in loans is likely driven primarily by the need to pay the insurance premia at the time when savings and loan volumes are recorded.

We perform two robustness checks. First, we modify a new method of calculating the variance of matching estimators presented in Abadie and Imbens (2006). Because we perform the spatial matching without replacement, agents may be used as a match multiple times, requiring an upward adjustment in the estimator of the variance. We recalculate the standard errors of the spatial matching estimator using the square roots of the corresponding diagonal elements of

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

where N is the number of agents, 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. The results are presented in table 6; the sample standard errors from this method are roughly twice as large as those calculated by the normal method. Only the decrease in dropout in the bi-

weekly program remains significant at the 95% level, and the increase in new clients for the insurance program falls to 90% significance. Different FINCA groups often meet at the same location, meaning that multiple groups in the parallel population may qualify as a nearest neighbor. We dealt with this problem in the data by averaging residuals for all units of a given status at a given location, but this multiple matching led to a substantial increase in the estimated variance. A study population with a more dispersed, uniform spatial distribution would likely not see this large difference between standard errors calculated by the two different methods. The additional uncertainty introduced by multiple matches to the same units substantially decreases our confidence in the conclusions, but the main results are still present.

Our second robustness check takes advantage of the fact that for all of our outcomes except savings, we have two pretreatment observations. This allows us to test the spatial assumption because we can verify whether the spatial shocks incurred by choosers and non-choosers were similar. The results of this exercise are reported in table 7. Because there are no treatment effects experienced anywhere in this population, both of our DID regressions are now false. The first column reports the false DID among choosers, and the second the false DID among non-choosers. We see once again that significant spatial phenomena exist that were not properly picked up by our control variables. While the number of observations among choosers is limited, we see that in every case where there are significant spatial effects among non-choosers, the magnitude of the effects among choosers is similar. When we run the spatial matching estimator (using the normal s.e. calculation and $M = 1$), we see that there are no significant differences between the spatial shocks experienced by these two populations in the eight months prior to treatment. We thus fail to reject the spatial assumption.

VII. Conclusion

In the presence of a second control group, we have a way of testing the assumptions underlying the use of a DID estimator of spatial treatment effects. Not only does the application in this paper display heterogeneity of a type that will result in bias in the DID, but our battery of control variables has been largely ineffective in removing this heterogeneity. While this result needs to be tested in other data sets, the implication is that studies which appear to be estimating spatial DID impacts, having controlled for all observable differences between regions, may be very sensitive to any differences in raw growth rates across the space of the treatment criterion. Particularly when legislative or administrative units are used to define treatment and control regions, we have myriad reasons to expect that unexplained differences will exist. Data collection, managerial talent, differing incentive structures, shocks, and endogenous pol-

TABLE 7.—COUNTERFACTUAL TEST OF SPATIAL ASSUMPTION
(FALSE IMPACT ESTIMATES)

	Choosers False DID	Non-choosers False DID	False Spatial Matching
Biweekly:			
# obs:	72	162	234
% new clients	−4.6611	−4.0120*	−3.0326
	(−1.4165)	(−1.7716)	(−1.1252)
% dropout	−0.2362	4.9902	−2.2653
	(−0.0415)	(1.4189)	(−0.5778)
Loans (USD)	8.0202	12.2790**	−2.3783
	(1.1398)	(3.5661)	(−0.5098)
Grades (A = 4, D = 1)	−0.0387	−0.0039	−0.0066
	(−0.1554)	(−0.0314)	(−0.0423)
Insurance:			
# obs:	61	173	234
% new clients	1.7701	−1.0621	−0.0663
	(0.3329)	(−0.5213)	(−0.0313)
% dropout	4.3708	−0.3149	2.5592
	(0.6838)	(−0.0907)	(0.6696)
Loans (USD)	−9.8486	−11.1089**	−1.6184
	(−1.0493)	(−3.2083)	(−0.3545)
Grades (A = 4, D = 1)	0.0908	0.2259*	−0.1672
	(0.3328)	(1.7591)	(−1.3548)

t-statistics in parentheses, * = 90% significance, ** = 95% significance.

icy placement can all cause units to differ in unexplained ways across the treatment rule, and hence the DID will be inconsistent.

Through the use of the spatial matching estimator we can directly measure other forms of unexplained spatial heterogeneity and recover unbiased impact estimates. Using the non-choosers of the program to eliminate this spatial heterogeneity, we find the biweekly repayment program improving retention of existing clients while also improving repayment. Since it lowers transaction costs on both sides of the contract, our results indicate significant welfare gains from extending the choice of this program to the rest of FINCA's groups in Uganda.

Insuring clients against health expenditure risk has no effect on their repayment performance. Since few environments on earth have a higher incidence of endemic diseases than Uganda, this is relatively strong evidence for the fact that health shocks are not a major determinant of microfinance delinquency. The insurance program retained existing clients more successfully and attracted new ones at a much greater rate than control groups. In addition, the program caused a weak drop in savings and increase in loans. It is very likely that the insurance caused increases in the household welfare of participants, however the program as currently constituted represents a mixed blessing from the perspective of the lender. Increasing debt loads and decreasing savings also indicate that the financial burdens of paying for the insurance policy may be undermining the economic self-sufficiency that should be the ultimate goal of microfinance.

The impact of these programs is closely related to the manner in which groups choose whether to accept them, and hence to the difference between the average treatment effect and the treatment effect on the treated. We have strong reason to suspect that the treatment effect of the biweekly repayment program on the average microfinance borrower would be negative, particularly in terms of repayment performance. What this experiment shows is that as long as a mechanism exists to expose group members to the negative repercussions of their choices (in this case, joint liability), then the groups themselves will make the correct decisions regarding insurance and risk. The more mixed success of the insurance program illustrates these same concepts. Insured groups were able to accept a large number of new clients who appeared less attractive without sustaining a drop in repayment, indicating an ability to overcome the asymmetric information present in the lending contract. The insurance program itself, however, suffered from adverse selection in client health, huge cost overruns, and use of the system that was dramatically higher than predicted. Because FINCA's joint-liability contract did not expose the groups to any health-related asymmetric information costs, local information was not used to overcome these problems. If we wished to learn the lessons of microfinance, a financially sustainable insurance program might include features like a joint copayment to be made collectively by all members of a group; this will mobilize local information and induce agents to screen on health quality.

It has become increasingly clear that program evaluation needs to be built into the design of programs in order for us to have any hope of measuring their effects accurately. In many contexts, the best way of doing this is to randomize some aspect of the design or implementation of the program. What is suggested here is a different path; we show that the use of a simple rule for the treatment criterion presents us with a measurable dimension along which we need to worry about bias. Because this dimension can be used for matching purposes, we can generate high-quality controls in situations

where a second control group is present and assignment into the treatment is based on easily observable criteria. The implication is that, in environments where randomization is not feasible, a spatial testing design with transparent eligibility requirements is an alternate way of producing high-quality impact estimates.

REFERENCES

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

Angrist, I., "Conditioning on the Probability of Selection to Control Selection Bias," *NBER technical working paper 181* (1995).

Banerjee, A., E. Duflo, and K. Munshi, "The (Mis)allocation of Capital," *Journal of the European Economic Association* 1 (2003), 484–494.

Bertrand, M., E. Duflo, and S. Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119:1 (2004), 249–275.

Card, D., and A. Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *The American Economic Review* 8:4 (1994), 772–793.

de Aghion, B. A., and J. Morduch, "Microfinance Beyond Group Lending," *Economics of Transition* 8:2 (2000), 401–420.

Dehejia, R., and S. Wahba, "Propensity Score Matching Methods for Non-experimental Causal Studies," this REVIEW 84:1 (2002), 151.

Ghatak, M., "Group Lending, Local Information, and Peer Selection," *Journal of Development Economics* 60 (1999), 27–50.

Gruber, J., "The Incidence of Mandated Maternity Benefits," *The American Economic Review* 84:3 (1994), 622–641.

Hammermesh, D., and S. Trejo, "The Demand for Hours of Labor: Direct Evidence from California," this REVIEW 82:1 (2000), 38–47.

Heckman, J., J. Ichimura, and P. Todd, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies* 64 (1997), 605–654.

Kremer, M., "Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons," *AEA Papers and Proceedings* 93:2 (2003), 102–106.

Kugler, A., "Wage-Shifting Effects of Severance Payments Savings Accounts in Colombia," *Journal of Public Economics* 89 (2005), 487–500.

McIntosh, C., A. de Janvry, and E. Sadoulet, "How Rising Competition among Microfinance Institutions Affects Incumbent Lenders," *The Economic Journal* 115 (2005), 987–1004.

McKenzie, D., and D. Mookherjee, "Distributive Impact of Privatization in Latin America: Evidence from Four Countries," *Economia* 3 (2003), 161–218.

Morduch, J., "Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh," Unpublished working paper (1998).

——— "The Microfinance Promise," *The Journal of Economic Literature* 37:4 (1999), 1569–1614.

Pitt, M., and S. Khandker, "The Impact of Group-Based Credit Programs on Poor Households in Bangladesh. Does the Gender of Participants Matter?" *Journal of Political Economy* 106:5 (1998), 958–995.

Ravallion, M., E. Galasso, T. Lazo, and E. Philipp, "Do workfare participants recover quickly from retrenchment?" World Bank working paper (2002).

Rosenbaum, P., "The Role of a Second Control Group in an Observational Study," *Statistical Science* 2:3 (1982), 292–316.

Rosenbaum, P., and D. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70 (1983), 41–55.

Stiglitz, J., "Peer Monitoring and Credit Markets," *World Bank Economic Review* 4 (1990), 351–366.

Telser, L., "Iterative Estimation of a Set of Linear Regression Equations," *Journal of the American Statistical Association* 59:307 (1964), 845–862.