Quantitative Impact Evaluation of the United Nations Mission in Liberia: Technical appendix

Cyrus Samii[*]

February 1, 2010

[*] Department of Political Science, Columbia University. Email: cds81@columbia.edu.

I. Overview

We present the methods used in Mvukiyehe and Samii (2010). Peacekeeping deployments can be considered as "treatments" applied at the level of locality. Outcomes of interest are individual and household level behaviors—displacee return, investment decisions, attitudes, etc. The appropriate sampling design is thus one that selects locality clusters that are defined based on their treatment status. Random samples of individuals within clusters are used to measure between cluster effects on individual outcomes. This is the standard multilevel analysis set-up (Snijders and Bosker 1999).

Our sampling design exploits the advantages of matched clusters in estimating cluster-level treatment effects (Imai et al 2008). This is an observational study insofar as we had no control in assigning treatment. Nonetheless, we use current best practices derived from the program evaluation literature to obtain a sample that allows us to best approximate the conditions of a randomized trial (Rubin 2001, 2008). We first conducted a power analysis to determine optimal cluster and within-cluster sample sizes under the practical constraints of the survey. We then used a matching algorithm to obtain a sample that matched comparison clusters to treated clusters. In all of the analysis below, we use Liberia's 306 clans as the relevant clusters. Clans aggregate villages and may contain as few as dozens and as many as tens of thousands of households. We consider them to be the relevant group, as they delineate clear economic and administrative systems to which deployments were tied. We now present details on the power analysis, matching, and household and individual sampling methods.

II. Power analysis

We determined the optimal number of matched cluster pairs as well as within-cluster sample sizes using simulation-based power analysis.[1] As discussed in the next section, we actually used matched trios in the sampling design. Nonetheless, the estimation of treatment effects will be based on making pairwise comparisons between treated clusters on the one hand and either "intermediate" or control groups on the other. As such, a power analysis based on pairwise inter-cluster comparisons is appropriate. The results below are stated in terms of optimal numbers of cluster pairs; for the sampling design, this should then be translated into optimal numbers of cluster trios.

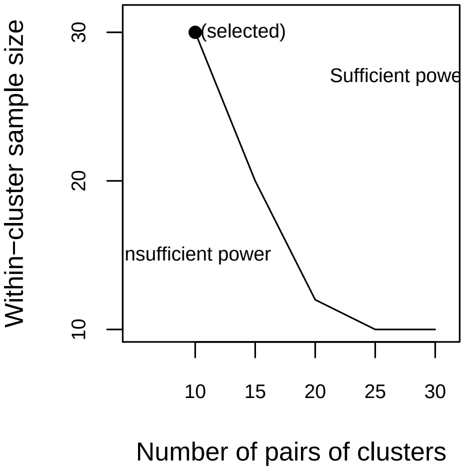
We are interested primarily in binary outcomes. As such, we model individual level outcomes (indexed by i), nested within clusters (indexed by c), as follows:

$$\begin{aligned}\Pr[y_i = 1] &= \text{logit}^{-1}(b_{0c[i]} + b_{1c[i]}t + b_2 x_i),\\ \text{with } b_{0c} &\sim N(b_{00}, \sigma^2_{b0})\\ b_{1c} &\sim N(b_{01}, \sigma^2_{b1}),\end{aligned}$$

where $t \in \{0, 1\}$ is the treatment indicator and x_i is a confounding covariate with within-cluster means that vary with t. The x_i represents a confounder that is only measurable from the survey, and thus cannot be used in the matching prior to cluster selection. The model implies that treatment effects vary over clusters, with cluster-specific treatment effects distributed with mean b_{01}. We are thus interested in recovering b_{01} as our measure of the treatment effect. A random intercept, b_{0c}, also varies over clusters; the variance of the inter-cluster intercept determines the level of between-cluster variance relative to within-cluster variance.

[1] See Gelman and Hill (2007) for a discussion of simulation-based power analysis.

Figure 1: **Numbers of Clusters and Within-Cluster Sample Sizes Providing Sufficient Power**



The curve shows combinations of numbers of pairs of clusters and within-cluster sample sizes that grant exactly .8 power. These are the sample allocations that minimally meet the criteria of providing a power level of .8. The dot at (10,30) shows the sample allocation that we selected.

The simulation settings were as follows. Under treatment, the probability that $y_i = 1$ is .55 on average; it is .45 on average under control. We say "on average" because as the model above makes clear, treatment effects are are modeled as being heterogeneous by cluster. Based on the variance of a binomial random variable, setting things so that average treated and control outcomes are symmetric around .5 provides the "hardest" test for a binary outcome. We choose an effect size of .1 as being the minimum that would be substantively meaningful. The covariate, x_i, has mean 0 for individuals in control clusters, and .125 in treated clusters. Thus, not controlling for x_i introduces bias of .05 on average in our measure of the marginal effect of treatment on $\Pr[y_i = 1]$. We set $\sigma^2_{b0} = .125$, which induces about 95% of clusters to have baseline inter-cluster heterogeneity within ±.05 in $\Pr[y_i = 1]$. We set $\sigma^2_{b1} = .25$, in which case for about 95% of clusters, treatment effects heterogeneity will be within ±.1. These variance settings imply substantial between-cluster heterogeneity in both baseline levels and treatment effects.

We chose a power level of .8 and a p-value of 0.05, which are conventional. Simulations were run, using the settings above, for different combinations of numbers of treated/control cluster pairs and sizes of within-cluster samples. Results of the power analysis are displayed in Figure 1. The curve shows estimates from the simulations for sample allocations that grant exactly .8 power. We assume that (i) the marginal rate of substitution generally favors increasing within-cluster sample sizes but that (ii) we face a hard constraint of 30 households within each cluster. Under those constraints, our optimal allocation is 30 households in 10 treated/control sets.

III. Selecting matched clusters

We wanted our sampling design to deal with two issues. First, we wanted to achieve maximal balance on potentially confounding cluster-level attributes to minimize potential confounding bias in our estimation of the cluster-level effects of deployments. Second, we wanted to minimize the potential contamination in our estimation of treatment effects due to the effects of deployments in one area "spilling over" into areas were deployments were not assigned per se. [2] Note that the two goals of maximal balance and minimum contamination are competing: the potential for balance over confounding covariates decreases in distance from treated clusters, but the potential for contamination also decreases in distance.

To deal with these two issues, we first created a three-way categorization of clusters: treated, intermediate, and control. Treated clusters were those that contained peacekeeping deployments at some point during the deployment period of 2004 to present. Control clusters were those that never contained peacekeeping deployments and were sufficiently distant from treated clusters so as to be minimally likely to experience spill-over. Intermediate clusters were those that never contained deployments, but were in close proximity to treated clusters and thus likely to experience spill-over. To determine whether a non-deployment cluster was far enough to be considered "control," we first examined the distance between the actual deployment base locations within clans and the other villages within that same clan. We found the maximum base-village distance and used that as a critical distance value.[3] Any non-deployment clan that contained villages that were within that critical distance value to any deployment location was considered "intermediate." We think this works as a reasonable way to classify clusters in terms of potential spill-over given the information available.

We used a matching algorithm to identify sets of intermediate and control clusters that would allow us to achieve maximal balance relative to treated clusters in the distributions of covariates. We employed a genetic matching algorithm, which matches directly on covariates (Diamond and Sekhon 2005). Matching without replacement was used, since our budget permitted us to sample a set of control and intermediate clusters for each treated cluster and to obtain ample within-cluster sample size. In any case, the set of matched clusters from a with-replacement algorithm is nested within the without replacement set. We can thus test for hidden bias by examining the magnitude of differences in treatment effect estimates when using the with-replacement versus the without-replacement control samples, although the power of the test may be less than the power of the full-sample analysis (Rosenbaum 1987).

The covariates used in the matching algorithm were as follows:

- Conflict exposure, operationalized in terms of cumulative "exposure" to major conflict events.

[2] We formalize spill-over by defining the potential outcomes for unit i in terms of $Y_i(T_i, T_{s[i]})$, and where $(T_i, T_{s[i]}) \in \{0,1\} \times \{0,1\}$. T_i is the treatment status of unit i and $T_{s[i]}$ is an indicator taking 1 if any unit $j \neq i$ within the area s around i has $T_j = 1$ and 0 otherwise. In principle, each unit has four potential outcomes. In practice, it is often the case that treated units are sufficiently spread apart such that $Y_i(1,1)$ units are never realized. Such is the case for peacekeeping in Liberia as operationalized in this study. When this is the case, two types of conditional mean differences between treated and non-treated units can be defined: $\mathrm{E}[Y_i(1,0) - Y_i(0,0)|X_i]$ and $\mathrm{E}[Y_i(1,0) - Y_i(0,1)|X_i]$, where X_i is the set of conditioning covariates that ensures exhangeability over the support of X_i for treated units. Our measure of spill-over is given by the difference in these conditional mean differences—i.e. $\mathrm{E}[Y_i(0,1) - Y_i(0,0)|X_i]$. The causal interpretation of these quantities is complicated, but together, they provide measures that allow for a certain type of average effect of the treatment on the treated to be estimated. See Rosenbaum (2007) for a relevant discussion.

[3] This value was .125 in standard x-y latitude/longitude coordinate space.

Exposure to a conflict event was measured as $\exp(-\alpha \cdot d_{cv}^2)$, where α is a scale parameter moderating the effect of distance on exposure, d_{cv} is the distance of the population center of cluster c from the geographic center of conflict event v. The α level was set to our own judgment of what makes for a plausible relationship between distance and exposure. We examined sensitivity of rankings of exposure levels to different α values and found that within a reasonable range, sensitivity was low. The population center of cluster c is simply the average of latitudinal and longitudinal coordinates of villages in that cluster. Conflict event data were taken from the ACLED dataset.

- Geographic location, operationalized in terms of longitudinal and latitudinal coordinates of the population center of the cluster.

- Pre-intervention local social infrastructure, including numbers of schools and health facilities as of 2005 (the earliest data available). These data are from the UN-OCHA PCodes dataset.

- The number of household in the cluster as of 2005 (the earliest data available). These data are from the UN-OCHA PCodes dataset.

The cluster selection process was as follows. Ten treated clusters were drawn randomly from the set of 44 treated clusters in the country. These 10 treated clusters were first matched with control clusters. The ten treated clusters were then matched with intermediate clusters. Together, these two sets of pairs resulted in a set of 30 clusters with maximal balance over covariates. We will be able to use the matched trios to test for spill-over effects; this can be done by examining whether differences in treated-intermediate outcomes differ significantly from differences in treated-control outcomes. Because all matching was done to balance relative to treated clusters, the quantity of interest that we are estimating is a form of the "average effect of the treatment on the treated."

The results of the matching process are displayed in Figures 2 and 3. The results show great improvement in balance over covariates in the matched relative to the unmatched samples. The resulting sample of clusters is shown in the map in Figure X [to be included].

IV. Household and individual sampling within clusters

Household sampling proceeded in two stages. For the first stage, we used the 2008 census enumeration areas created by the Liberia Institute of Statistics and Geo-Information Services. Two enumeration areas were selected at random from the list of enumeration areas in each clan. For the second stage, the enumeration teams visited the selected enumeration areas, and listed all households. The target number of households were selected at random from this list. As described in the power analysis, the target was to select 30 households within each cluster. In practice, the enumeration teams sometimes selected more households; this seemed to be due to miscommunication among enumeration staff about both target numbers and the need to interview replacement households. In other clans, non-response problems made it so that the 30 household target could not be realized in some of the clusters. As a result, the number of households per clan was uneven. The weighting scheme below addresses this problem to a certain extent.

V. Resulting Sample

The distribution of households in the resulting sample is shown in Table 1. Non-response explains the cases where the number of households is less than the intended 30 per clan. Supplemental households were taken in some other clans to compensate for such non-response, which resulted in some clan-level samples exceeding the intended 30 per clan.

Table 1: **Distribution of Household Sample Over Clans**

Distant Communities

Clan	County	Household Sample Size
Gbor-B	Bomi	30
Lower Mecca	Bomi	25
Lower Zor	Bomi	36
Lorla	Bong	29
Waytuah	Bong	25
Gborbo	Grand Gedeh	33
Tarleh	Grand Gedeh	30
Tchien Menyea	Grand Gedeh	39
Z400 Central West Point	Montserrado	38
Gbehlay	Nimba	30
Ding Rural	Montserrado	23
Mehn Rural	Montserrado	17

Proximate Communities

Clan	County	Household Sample Size
Deygbo	Bomi	31
Manna	Bomi	40
Tehr	Bomi	25
Famazette Ward	Grand Bassa	24
Gowingbo Grand	Bassa	27
Own Your Own	Grand Bassa	29
Gwenee	Grand Gedeh	37
Quardu	Lofa	24
Z300 Central Clara Town I	Montserrado	23
Gborplay	Nimba	44
Leepea1	Nimba	15
Fahn-Seh	Rural Montserrado	22

Deployment Communities

Clan	County	Household Sample Size
Lower Togay	Bomi	28
Garyea	Bong	20
Nyaforquellie	Bong	27
Suakoko	Bong	22
Bexley Ward	Grand Bassa	17
Kannah	Grand Gedeh	33
Upper Buchanan Community	Grand Bassa	36
Gizima	Lofa	21
Harbel	Margibi	28
Z100 Central New Kru Town	Montserrado	19
Z200 Central Logan Town	Montserrado	14
Sango-Zao	Nimba	21
Sehyi	Nimba	38

The distribution of sampled excombatants is given in Table 2. Excombatant interviews were arranged in coordination with representatives of the national reintegration program in areas outside Monrovia nearby where the household sampling was conducted.

Table 2: **Distribution of Excombatant Sample Over Clans**

Clan	County	Number of Sampled Ex-combatants
Manna	Bomi	48
Nyaforquellie	Bong	1
Garyea	Bong	22
Famazette Ward	Grand Bassa	4
Gowingbo	Grand Bassa	1
Upper Buchanan Community	Grand Bassa	16
Gborbo	Grand Gedeh	4
Gwenee	Grand Gedeh	17
Kannah	Grand Gedeh	22
Tarleh	Grand Gedeh	10
Tchien Menyea	Grand Gedeh	13
Gizima	Lofa	28
Quardu	Lofa	39
Leepea	Nimba	21
Sehyi	Nimba	29

VI. Field implementation

The survey was implemented by a team of enumerators recruited through the Liberia Institute of Statistics and Geo-Information Services. Data was entered by a team of data entry personnel also from the Liberia Institute of Statistics and Geo-Information Services.

VII. Missing Data

Variables in the final dataset exhibited some missingness due to enumerator mistakes in following skip patterns and typographical errors in the data entry process. For some important key variables—e.g. demographic variables and income—we used multiple imputation to deal with these problems. Specifically, we used the chained regression imputation method implemented in `mice`, using the predictive mean matching option (van Buuren and Groothuis-Oudshoon 2009). In the data analysis, we used standard methods for combining estimates from the multiply imputed data.

VIII. Matching

As described above, "genetic matching" was used to construct the sample of clans. In the data analysis, a more intuitive method was employed (e.g. in the analysis of economic reintegration outcomes)—namely, the "coarsened exact matching" (CEM) method developed by Iacus et al (2009). CEM operates by matching observations exactly on coarsened covariates (e.g., age in years becomes a set of age categories—"15-25", "26-35", and so on—and then the algorithm matches exactly on these age categories). We would have used this to construct the sample as well, however the software to do so was not yet available.

IX. References

Diamond A, Sekhon JS. 2005. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." Typescript. University of California–Berkeley.

Gelman A, Hill J. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press.

Iacus SM, King G, Porro G. 2009. "Causal Inference Without Balance Checking: Coarsened Exact Matching." Typescript, University of Milan, Harvard University, and University of Trieste.

Imai K, King G, Nall C. 2008. "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation." Typescript. Princeton University and Harvard University.

Mvukiyehe E, Samii C. 2010. "Quantitative Impact Evaluation of the United Nations Mission in Liberia: Final Report." Typescript, Columbia University.

Rosenbaum PR. 1987. "The Role of a Second Control Group in an Observational Study." *Statistical Science.* 2:292-306.

Rosenbaum PR. 2007. "Interference Between Units in Randomized Experiments." *Journal of the American Statistical Association.* 102:191-200.

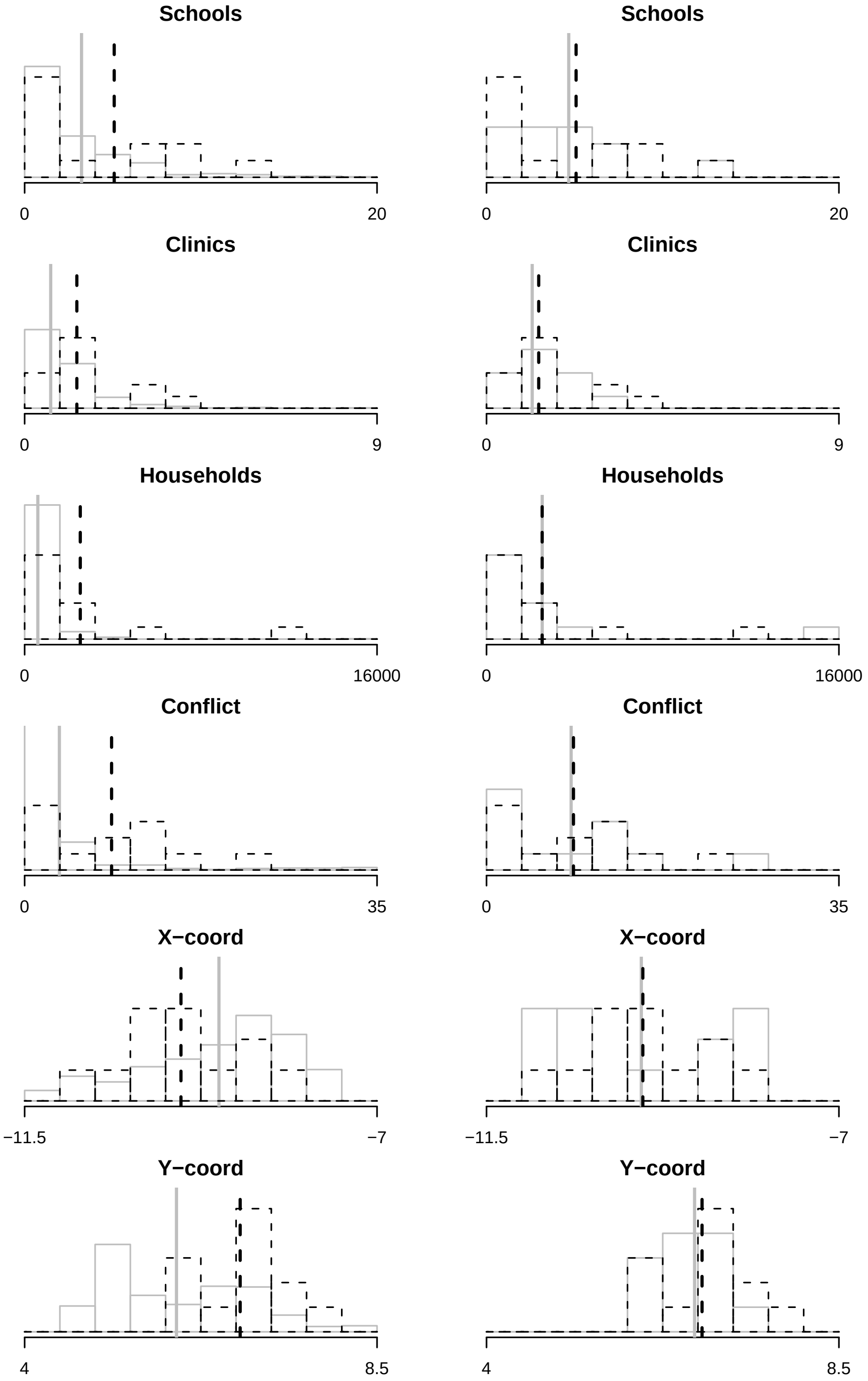
Rubin DB. 2001. "Using Propensity Scores to Help Design Observational Studies: Application to tobacco litigation." *Health Services and Outcomes Reseach Methodology.* 2:169-188.

Rubin DB. 2008. "For Objective Causal Inference, Design Trumps Analysis." *The Annals of Applied Statistics.* 2:808-840.

Snijders T, Bosker R. 1999. *Multilevel Analysis.* Thousand Oaks: Sage Publishers.

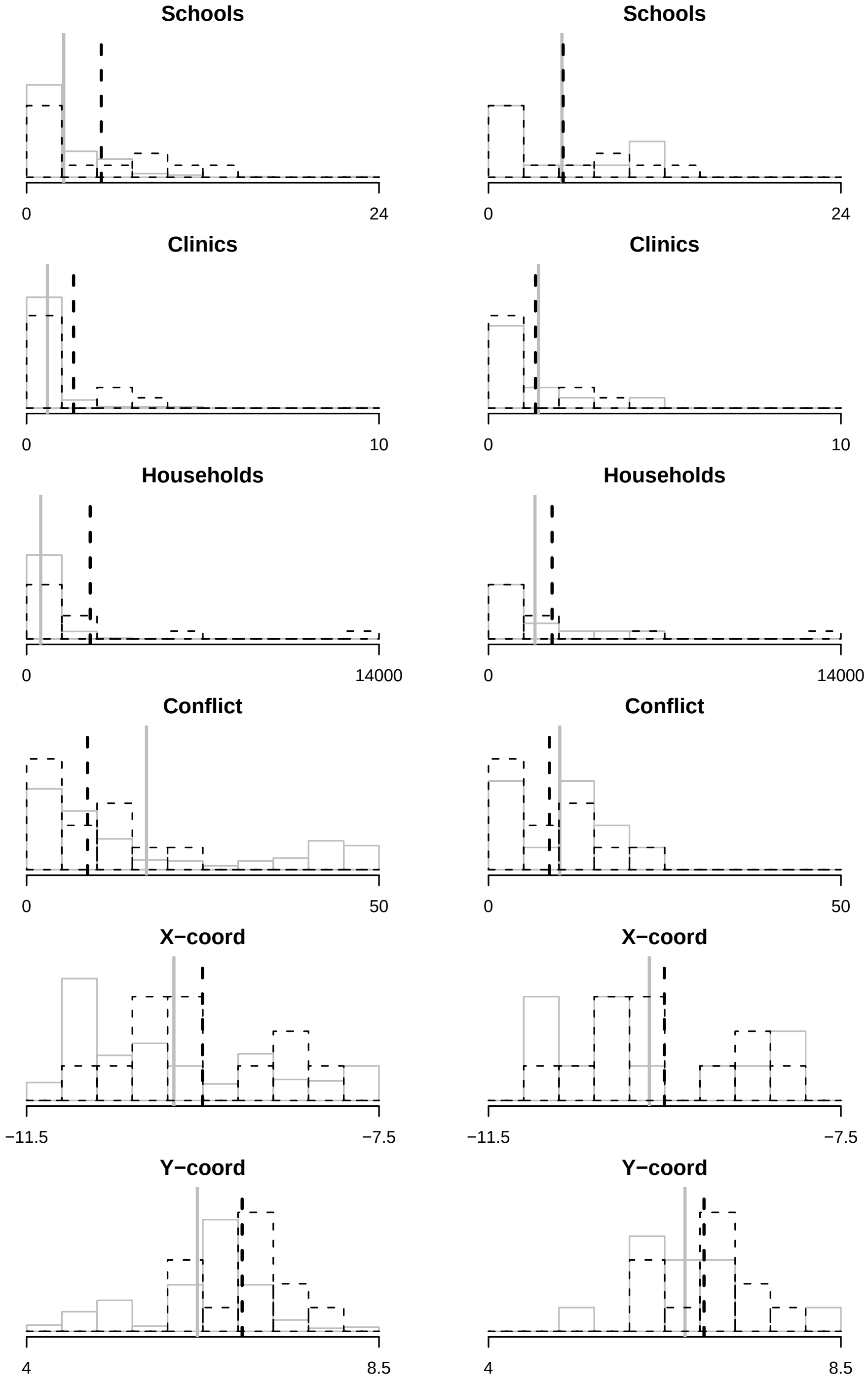
van Buuren S, Groothuis-Oudshoon K. 2009. "MICE: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software.* Forthcoming.

Figure 2: **Matching Results for Constructing "Control" Cluster Sample**



The histograms show covariate distributions before (left) and after matching. Covariate distributions for non-treated clusters are indicated by grey solid lines; dashed black lines are used for treated clusters. Mean values are indicated by the thick vertical grey line for non-treated and dashed black line for treated clusters.

Figure 3: **Matching Results for Constructing "Intermediate" Cluster Sample**



The histograms show covariate distributions before (left) and after matching. Covariate distributions for non-treated clusters are indicated by grey solid lines; dashed black lines are used for treated clusters. Mean values are indicated by the thick vertical grey line for non-treated and dashed black line for treated clusters.