Causal inference with small samples and incomplete baseline for the Millennium Villages Project

Shira Mitchell

Columbia University, New York, NY, USA.

Rebecca Ross

Columbia University, New York, NY, USA.

Susanna Makela

Columbia University, New York, NY, USA.

Elizabeth A. Stuart

Johns Hopkins University, Baltimore, MD, USA.

Avi Feller

University of California, Berkeley, CA, USA.

Alan M. Zaslavsky

Harvard University, Boston, MA, USA.

Andrew Gelman

Columbia University, New York, NY, USA.

Summary. The Millennium Villages Project (MVP) is a ten-year integrated rural development project implemented in ten sub-Saharan African sites. We describe the design for causal inference about the MVP's effect on a variety of development indicators. Causal inference for the MVP context presents many challenges: a nonrandomized design, limited baseline data for candidate controls, and the assignment of treatment to only ten sites, limiting effective sample sizes. We develop and carry out a matching procedure tailored to small samples and designed to facilitate communication with subject-matter experts. We propose hierarchical Bayesian causal models for multiple outcomes that account for uncertainty in baseline covariates and ameliorate the problem of "multiple comparisons." This paper provides a case study of the careful design of a non-randomized study, with clear pre-specification of the procedure and matches before outcome data are available.

1. Introduction

The Millennium Villages Project (MVP) is a ten-year economic development project that operates in ten clusters of rural sub-Saharan African villages in ten distinct countries. The MVP implements a multi-sector package of community-level interventions at each of the ten sites (Sachs and McArthur, 2005; Sanchez et al., 2007). See Mitchell et al. (2015a) for background on the project and our overall plan to evaluate it. This paper describes the design for causal inference about the MVP's effect on a variety of development indicators. Causal inference for the MVP context presents many challenges: a nonrandomized design, limited baseline data for candidate controls, and the assignment of treatment to only ten sites, limiting effective sample sizes. We develop and carry out a matching procedure tailored to small samples and designed to facilitate communication with subject-matter experts. We propose hierarchical Bayesian causal models for multiple outcomes that account for uncertainty in baseline covariates and ameliorate the problem of "multiple comparisons."

The MVP began in 2005, without designating control villages and only collecting data in the project sites, i.e. the "Millennium Villages" (MVs). At each MV, resources were concentrated in a core area of roughly 1000 households called the "MV1." (The remainder of each MV is called the "MV2," where a subset of interventions were implemented. We do not utilize the MV2 areas in this study.) Today, at the project's end-line, funding is available for surveying areas both inside and outside the MVs to conduct causal inference. Our causal design includes matching to select control villages, collection of outcome data in treatment and control villages, and then regression to estimate causal effects. Our outcomes are defined in Mitchell et al. (2015a), and include indicators of poverty, agriculture, education, gender equality, health, environmental sustainability, and infrastructure.

We define the causal effect of the MVP in terms of potential outcomes, outcomes that would have happened with the MVP or without. Focusing on a particular outcome from our list of development indicators, let $y(1)$ be the outcome for a unit (an individual, household, or village) that would have occurred had the unit been within a Millennium Village, and $y(0)$ the outcome that would have occurred had the unit not been within a Millennium Village. The causal estimand is then defined as a comparison between $y(1)$ and $y(0)$, usually as a difference, ratio, or odds ratio, averaged over a finite or superpopulation. Even with control data, estimation of causal estimands relies on untestable assumptions, whose justifications rely on context-specific knowledge.

One necessary assumption is the *stable unit treatment value assumption*, which requires that potential outcomes for any unit do not vary with the treatments assigned to other units (i.e. units do not interfere with one another), and for each unit there are no different versions of the treatment which lead to different potential outcomes (Imbens and Rubin, 2015, Chapter 1). Essentially, this assumption ensures that the potential outcomes introduced above are well-defined. In our evaluation, we consider only two levels of treatment: either a unit (an individual, household, or village) is within a Millennium Village,† or a unit is far enough away from any areas where the project operated that it cannot be affected by it. We aim to minimize interference by limiting our control pool to areas at least ten kilometers away from the MV, outside a "buffer zone" of very likely interference. The Millennium

†Here we ignore issues of migration and define treatment as being in a Millennium Village in 2015, regardless of the duration of stay in the Millennium Village.

Villages themselves are far apart from each other and situated in different countries, so we assume no interference among them.

A second assumption needed for estimating causal effects is *unconfoundedness* (Rubin, 1976, 1978, 2008; Imbens and Rubin, 2015; Gelman and Hill, 2007; Greenland et al., 1999; Bang and Robins, 2005; Angrist and Pischke, 2009). This assumption requires that the distribution of potential outcomes should be the same for the MVs and control areas, conditional on the observed pre-treatment variables. To make unconfoundedness plausible, we want to control for many variables that are not affected by treatment (Rosenbaum, 1984). These need not be temporally before treatment, as long as the project could not have affected them (e.g. temperature).

For our design, we follow matching with regression, since the combination of the two methods is more robust than each alone (Rubin, 1973; Rubin and Thomas, 2000; Ho et al., 2007; Kreif et al., 2011; Abadie and Imbens, 2011; Robins et al., 2000; Robins and Rotnitzky, 2001; Bang and Robins, 2005). Successful matching avoids extrapolation to areas of poor overlap, which would rely heavily on the correctness of the regression model. If the stable unit treatment value assumption holds, and we include enough pre-treatment variables to satisfy unconfoundedness, a combination of matching and regression should do well to approximate results from a randomized experiment (Dehejia and Wahba, 1999; Dehejia, 2005; Shadish et al., 2008).

We begin by discussing data sources for pre-treatment variables to use in both the matching and regression. Our search for relevant pre-treatment data was informed by researching the site-selection process, assembling documents and correspondences to learn about treatment assignment. Next, we describe the matching procedure to select controls for each MV, and propose models to be fit to the outcome data. We assess the unconfoundedness assumption using our pre-treatment variables. Finally, we present a design analysis (i.e. "power calculation") and the data collection plan.

2. Data sources in the control pool

We require pre-treatment variables in the ten countries, measured at a fine enough geographic resolution, to be able to identify matched controls and for regression adjustment in our causal models. Below we discuss identified sources of data.

Geographic data

For the ten countries, we collected geographic data from geographic information system (GIS) databases, including agroecological zone, travel time to nearest city of more than 100,000 population, soil composition, vegetation index, temperature, elevation, and population density (Dixon et al., 2001; Joint Research Centre: Land Resource Management Unit; ISRIC: World Soil Information; GPWv3; GADMv2, 2012; IRI/LDEO; The CGIAR Consortium for Spatial Information (CGIAR-CSI)). See Appendix B.1 for the list of geographic variables.

We need to be able to match the MV1s to controls of comparable geographic area. Given this requirement, and the scale of the geographic variables, the data were processed using fishnets with square grid cells approximately equal in area to each country's MV1 (ranging from 2km × 2km to 12km × 12km). These grid cells are a partition of the area in each country, making them a convenient choice for matching units. We consider the treatment units to be grid cells that overlap the MV1, and have either at least 40% area in MV1 and MV2 combined or have at least 20% area in MV1. These treatment units are two to four contiguous grid cells within each country. The set of candidate control grid cells excludes any grid cells that overlap the MV2 or a ten-kilometer buffer zone enveloping the MV.

Census data

Georeferenced census data and corresponding administrative boundary data is often difficult to procure and process, especially from pre-2005. We are working to resolve this issue, but due to time and resource constraints, census data was not usable in time for selection of matched controls.

Demographic and Health Surveys

Many of our outcomes of interest are measured by the Demographic and Health Surveys (DHS), using survey tools similar to ours (MVP, 2011; Rutstein and Rojas, 2006). The DHS employs two-stage cluster sampling, with census enumeration areas as primary sampling units (i.e. clusters) (Measure DHS/ICF International, 2012, p.4,15). To protect anonymity, the DHS reports the GPS locations of cluster points displaced by up to five kilometers in rural areas (Measure DHS/ICF International, 2012; DHS, 2014). Therefore, DHS data are not associated with grid cells, but rather, with *DHS buffers*, circles around DHS cluster points with a five kilometer radius. We approximate the enumeration area boundary with the DHS buffer (a reasonable approximation if there is spatial smoothness in the DHS variables). The disadvantage of DHS data for our purposes is that it is geographically sparse, with 350-900 out of 8000-600,000 enumeration areas sampled per country, and 20-30 households sampled within each enumeration area, see Figure 1.

3. Selecting control villages

As discussed above, our matching units are grid cells of equal size to the MV1 treatment areas. Associated with these grid cells are geographic variables, and associated with DHS buffers are wealth, education, and health variables. Below we describe how we handle these different spatial divisions (grid cells and DHS buffers). We wish to select the "best" subset of five grid cells per country (see Section 6 for justification of our choice of five).

For seven of the ten countries (all but Tanzania, Nigeria, and Ethiopia), the treatment grid cells overlap at least one DHS buffer. For these seven countries we restrict the set of candidate matches to grid cells overlapping DHS buffers, allowing us to have some pre-treatment data on many outcomes of interest in our matched control areas.‡ Restricting the matches to areas with DHS data does not worsen the match on geographic variables enough to cause concern for subject-matter experts. Thus, we prefer to know pre-treatment values of our outcome variables measured by the DHS, as this has been shown to reduce bias in observational studies (Cook et al., 2008; Steiner et al., 2010). Due to the geographic sparsity and anonymity displacements of the DHS data, we must assume spatial smoothness in the health, wealth, and education variables. We consider DHS data overlapping any of the treatment grid cells as relevant to all two to four contiguous treatment grid cells.

Our matching procedure is separate for each MV1 (i.e. for each country), with exact matching on categorical variables followed by non-exact matching on continuous variables.

3.1. Exact matching variables

We match exactly on country, and, given the project's emphasis on farming systems, we also match exactly on agroecological zone.

Each of the ten countries containing an MV is divided into administrative units, whose names and functionality differ from country to country. Each MV is contained within a district (or local equivalent of district). For survey administration logistics, we limit matched controls to the MV district, or any districts that border the MV. Furthermore, we suspect that areas closer to the MV are likely to be better matches than areas farther away.

For some countries, the district containing the MV is small enough that there are few grid cells within the district. Therefore, we follow Stuart and Rubin (2008) and choose matches both inside and outside the district, matching on continuous variables described below. There is a tradeoff between a preference for within-district matches (government programs are sometimes implemented at this level) and wanting close matches on the continuous variables. The literature does not offer much guidance beyond a suggestion to use prior knowledge and previous studies. We therefore defer to subject matter experts who recommend constraining at least two of the five matched grid cells to be within the district containing the MV.

‡The MVP collected baseline data in all ten treatment sites, using survey tools similar to the DHS. It can be argued that these data should be used in the matching, especially for Tanzania, Nigeria, and Ethiopia, which have no pre-treatment DHS data available in the treatment areas. However, MVP baseline data is of varying quality, and its comparability to DHS data can only be evaluated for countries with DHS data near project baseline, see Table **??**. We therefore omit project baseline data from consideration in our matching procedure.

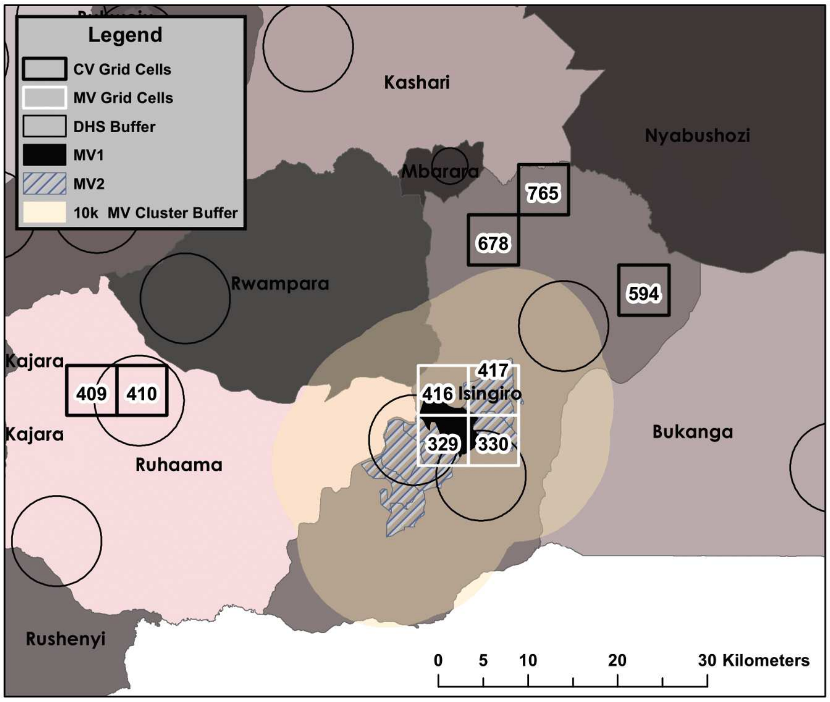


Fig. 1: A map of Uganda in the region surrounding the Millennium Village (MV). The core area that receives the full set of interventions, the MV1, is colored in black. The areas that received a subset of interventions, the MV2, are striped. A ten kilometer buffer is shaded in blonde. The DHS buffers, circles around DHS cluster points with a five kilometer radius, are drawn as circles. Treatment and comparison grid cells are in white and black, respectively, with comparison grid cells displayed for demonstration only. District boundaries are in different shades of gray (GADMv2, 2012). The MV is located in Isingiro district.

3.2. Non-exact matching variables

To make unconfoundedness as plausible as possible, we want to match on many pre-treatment variables. If assignment to treatment is unconfounded given covariates, then assignment is unconfounded given the *propensity score*, the average assignment probabilities for subpopulations with common values of the covariates (Rosenbaum and Rubin, 1983). It is simpler to find close matches using a scalar (the propensity score) rather than all covariates jointly. However, with so few treatment units (few grid cells, clustered into only ten MVs), it is difficult to fit propensity score models with many covariates. The models may vary substantially from country to country, increasing the number of parameters to estimate. We therefore choose to directly match on the variables of interest, and employ other methods of dimensionality reduction.

Our first form of dimensionality reduction involves selecting only the most relevant variables (Ho et al., 2007, p.217). For each continuous geographic variable, our data include the grid cells means and standard deviations (see Appendix B.1). We drop the standard deviation variables, except for the standard deviation associated with elevation, because it captures a ruggedness of terrain that is considered important. Among the thousands of variables available in the Demographic and Health Surveys (DHS), we choose a set closely resembling our outcomes of interest, see Appendix B.2.

Our second form of dimensionality reduction is creating indices of related variables. DHS computes a household asset index using the first principal component of a list of assets (similar to those measured by the MVP) (Filmer and Pritchett, 2001; Michelson et al., 2013). We use this as our measure of wealth. We create two additional indices for health and education. The variables used to create these indices are listed in Appendix B.2.

Our procedure to create indices involves the following. First, we aggregate variables measured within the household to a household-level variable $x_h^{(k)}$ for each household h and variable k. We then standardize each variable by its mean and standard deviation across all households in the country: $\tilde{x}_h^{(k)} = \frac{x_h^{(k)} - E(x_h^{(k)})}{\sqrt{Var(x_h^{(k)})}}$. Next, we "reorient" each variable so that larger values indicate higher economic development. Finally, for household h, its education (health) index is the mean of all $\tilde{x}_h^{(k)}$ where k is a variable that belongs to the education (health) index. If a variable is missing for a particular household, it contributes zero to the index (future work will explore more sophisticated methods of handling this missingness). Using the same procedure, we create a temperature index from four temperature geographic variables (see Appendix B.1).

We let S_{geo} denote the set of remaining geographic variables (with temperature combined into one index), and let S_{DHS} denote the set of three DHS indices.

3.3. Small area estimation

We fit small area models for each of the three DHS indices, using geographic data to improve our estimates (Ghosh and Rao, 1994; Ghosh and Natarajan, 1999; Nadram, 2000; Rao, 2003; Jiang and Lahiri, 2006). To account for design variables used in the DHS two-stage cluster sampling, our models include levels for clusters and regions within each country (Measure DHS/ICF International, 2012, p.4,15). Furthermore, we include the cluster sampling weights in the model, as recommended in the literature, since cluster sizes are not released by the DHS (Zheng and Little, 2003, 2004, 2005; Chen et al., 2010; Si et al., 2015).

We fit Fay and Herriot (1979) models, where the lowest level of the model is approximated by a non-Bayesian calculation without a complete model for the complex survey data structure (Zaslavsky, 2011):

$$\begin{aligned}
&\widehat{\overline{x}}_d \sim N(\overline{X}_d, v_d) \text{ for DHS clusters } d \\
&\overline{X}_d \sim N\left(\mathbf{e}_d^T\boldsymbol{\gamma} + \sum_{l=1}^{3} \beta_l I(w_d \geq \kappa_l) + \phi_{r[d]}, \sigma^2_{\text{cluster}}\right) \text{ for DHS clusters } d \\
&\phi_r \sim N\left(0, \sigma^2_\phi\right) \text{ for regions (/state/province) } r \text{ within a country}
\end{aligned} \tag{1}$$

where $\widehat{\overline{x}}_d$ is the standard design-based estimate of the mean DHS index in sampled cluster d and v_d its sampling variance. The $\mathbf{e}_d$ are the geographic variables, and w_d is the sampling weight for cluster d from the DHS. We use a "degree-0 spline" with knots κ_l chosen to be the sampling weights' quartiles (Zheng and Little, 2004). Where not otherwise specified, our priors are weakly informative.

Converting geographic data from grid cell level to DHS buffer level

In model 1, the geographic variables act as predictors at the DHS cluster level, which can be geographically identified as a DHS buffer (due to the displacement for anonymity). This requires us to convert the geographic data from grid cells to DHS buffers.

Let $\text{overlap}_{c,d}$ be the percent of grid cell c overlapped by DHS buffer d. Geographic variables are means across each grid cell area (see Section 3.2), except for the administrative units and the elevation standard deviation. For each such variable VAR, and for each DHS buffer d, compute:

$$\text{VAR}_d = \frac{\sum_c \text{overlap}_{c,d}\text{VAR}_c}{\sum_c \text{overlap}_{c,d}}.$$

Then, using a similar procedure for the second moment of the elevation variable, we compute the standard deviation of elevation within DHS buffer d as follows:

$$\text{ELEV\_STD}^2_d = \frac{\sum_c \text{overlap}_{c,d}(\text{ELEV\_STD}^2_c + \text{ELEV\_M}^2_c)}{\sum_c \text{overlap}_{c,d}} - \text{ELEV\_M}^2_d,$$

where ELEV_M_c and ELEV_STD_c are the mean and standard deviation of elevation within grid cell c (see Appendix B.1). In model 1, we use administrative units to partially pool across regions. To convert from grid cell administrative unit data to DHS buffers, we take the mode across grid cells overlapping the DHS buffer:

$$\text{ADMIN}_d = \text{mode}_{\text{overlap}_{c,d}>0}\{\text{ADMIN}_c\}.$$

We then take the first three principal components of the geographic variables (now at the DHS buffer level) to include as $\mathbf{e}_d$ in model 1.

Converting small area estimates from DHS buffer level to grid cell level

After fitting model 1, we have samples of $\overline{X}_d$ from the posterior distribution for each DHS cluster d. We convert to samples from the posterior at the grid cell level by computing, for each grid cell that overlaps at least one DHS buffer:

$$\overline{X}_c = \frac{\sum_d \text{overlap}_{c,d} \overline{X}_d}{\sum_d \text{overlap}_{c,d}}.$$

This procedure only computes DHS indices for grid cells that overlap DHS buffers. Though in the future we will improve our small area estimation procedure (via the inclusion of census variables, for example), due to time and data constraints we do not currently use our model to impute DHS data in grid cells not overlapping DHS buffers.

Millennium Village	Start date	DHS dates	census dates
Potou, Senegal	2006	2005, 2010-11	2002, 2013
Tiby, Mali	2006	2001, 2006	1998, 2009
Bonsaaso, Ghana	2006	2003, 2008	2000, 2010
Pampaida, Nigeria	2006	2003, 2008	1991, 2006
Koraro, Ethiopia	2005	2000, 2005, 2011	1994, 2007
Sauri, Kenya	2005	2003, 2008-9	1999, 2009
Ruhiira, Uganda	2006	2000-1, 2006, 2011	2002, 2013
Mayange, Rwanda	2006	2000, 2005, 2010	2002, 2012
Mbola, Tanzania	2006	2004-5, 2010	2002, 2012
Mwandama, Malawi	2006	2000, 2004, 2010	1998, 2008

3.4. Matching algorithms

After restricting to neighboring districts and the MV's agroecological zone (and for seven countries, to grid cells overlapping DHS buffers), our matching algorithm considers each possible set of five control grid cells to determine the set that best matches the treatment grid cells, with "best" defined below. Our search space is restricted to sets with at least two of the five matched controls lying within the district (see Section 3.1). For each set of control grid cells, we compute the match's "badness score," a measure of covariate imbalance described below.

After the exact-matching restrictions, let $N_{\text{in-district}}$ be the number of candidate control grid cells in the district containing the MV and $N_{\text{out-of-district}}$ the number in this district or any districts neighboring the MV. Thus, the number of possible matches is $\sum_{n_{in}=2}^{5} \binom{N_{\text{in-district}}}{n_{in}} * \binom{N_{\text{out-of-district}}}{5-n_{in}}$. If this number is greater than the number that can be considered in 48 hours, we instead first find the best two within-district matches, followed by the best three matches to complement these. This reduces the search space to $\sum_{n_{in}=2}^{5} \binom{N_{\text{in-district}}}{n_{in}} + \binom{N_{\text{out-of-district}}}{5-n_{in}}$. If this reduction is still insufficient to reduce the runtime to within 48 hours, we limit the search space using a variable thought by subject-matter experts to be highly correlated with the potential outcomes (e.g. the asset wealth index). We restrict control grid cells to be within an allowable margin of the mean of this particular variable amongst the treatment cells.

As mentioned above, in Tanzania, Nigeria, and Ethiopia, treatment cells do not overlap with DHS buffers. We therefore do not restrict control grid cells to overlap DHS buffers for those three countries. In Kenya and Uganda, the treatment cells do overlap DHS buffers, but Kenya only contains one grid cell within the district and agroecological zone that overlaps DHS buffers, and Uganda contains none. Therefore, for Kenya and Uganda we select two or three within-district matches using geographic data alone, but restrict the remaining matches to areas with DHS data.

3.5. Imbalance measures

Matching the joint distribution of the covariates between treatment and control implies that the simple difference in outcome means is unbiased for the treatment effect. However, with many covariates, estimates of the joint density are subject to the curse of dimensionality (Imai et al., 2008, p.498; Stuart, 2010, p.11). We follow the common procedure of working with lower-dimensional summaries (Ho et al., 2007, p.221), considering one matching variable at a time. For each variable k, let the (sample) means be $\overline{x}_t^{(k)}$, $\overline{x}_{fc}^{(k)}$, and $\overline{x}_{mc}^{(k)}$ for the treatment cells, the full set of candidate control cells, and the matched control cells, respectively. Let the standard deviations be $s^{(k)}$, $s_t^{(k)}$, $s_{fc}^{(k)}$, and $s_{mc}^{(k)}$ for all grid cells, the treatment cells, the full set of control cells, and the matched control cells, respectively. The standardized difference in means is widely recommended to check balance: $\frac{\overline{x}_t - \overline{x}_{mc}}{s^{(k)}}$ (see Stuart (2010, p.11), Imbens and Rubin (2015, Chapter 14, p.310-311), and Imai et al. (2008, p.498)). We also compare the differences in variance using the logarithm of the ratio of standard deviations between treatment and comparison groups, $\ln \frac{s_t^{(k)}}{s_{fc}^{(k)}}$ before matching, and $\ln \frac{s_t^{(k)}}{s_{mc}^{(k)}}$ after matching (Imbens and Rubin, 2015, Chapter 14, p.312).

Since we do not anticipate analyzing the MV1 grid cells separately, we do not examine within-pair statistics (Imbens and Rubin, 2015, Chapter 15, p.355-357). We combine the above scores into an overall "badness score" for a match, first by creating a badness score for the standardized difference in means:

$$\text{mean\_badness} = \frac{1}{|S_{geo}|} \sum_{k \in S_{geo}} \frac{|\overline{x}_t^{(k)} - \overline{x}_{mc}^{(k)}|}{\sigma^{(k)}} + w_{DHS} \frac{1}{|S_{DHS}|} \sum_{k \in S_{DHS}} \frac{|\overline{x}_t^{(k)} - \overline{x}_{mc}^{(k)}|}{\sigma^{(k)}}, \qquad (2)$$

where w_{DHS} is a weight used to increase the influence of DHS variables on the choice of matches. We also create a badness score for the differences in variance:

$$\text{var\_badness} = \frac{1}{|S_{geo}|} \sum_{k \in S_{geo}} \left| \ln \frac{s_t^{(k)}}{s_{mc}^{(k)}} \right| + w_{DHS} \frac{1}{|S_{DHS}|} \sum_{k \in S_{DHS}} \left| \ln \frac{s_t^{(k)}}{s_{mc}^{(k)}} \right|. \qquad (3)$$

We combine these two into a total badness score as follows:

$$\text{badness} = w_{mean} * \text{mean\_badness} + \text{var\_badness},$$

where w_{mean} is a weight that favors matching closely on means rather than variances.

Because the DHS variables are much closer to our outcomes of interest (they summarize pre-treatment values of the outcome variables), we set $w_{DHS} = 10$. We set $w_{mean} = 2$, assigning more importance to mean matching as opposed to variance matching.

As mentioned above, not all treatment grid cells overlap with DHS buffers, requiring modification of the above badness scores. For Tanzania, Nigeria, and Ethiopia (whose treatment cells do not overlap any DHS buffers) we drop the terms that measure the imbalance on DHS variables in expressions 2 and 3. For Kenya and Uganda we also consider matched control grid cells that do not overlap DHS buffers. In the above badness scores, this is handled by computing sample means and standard deviations using available cases. In future work, more sophisticated methods should be employed to handle the missingness of DHS data (we drop grid cells with missingness in the geographic variables, as this missingness is pre-treatment and the treatment grid cells have no missing geographic values).

Another complication with the above badness scores occurs when either $s_t^{(k)}$ or $s_{mc}^{(k)}$ is zero, making the variance badness infinite or undefined. When both $s_t^{(k)} = 0$ and $s_{mc}^{(k)} = 0$, we replace $\left| \ln \frac{s_t^{(k)}}{s_{mc}^{(k)}} \right|$ with zero, because this represents a good match (i.e. no badness). When $s_t^{(k)} = 0$ and $s_{mc}^{(k)} \neq 0$, we replace $\left| \ln \frac{s_t^{(k)}}{s_{mc}^{(k)}} \right|$ with $\left| \ln \frac{\frac{1}{10} s_c^{(k)}}{s_{mc}^{(k)}} \right|$. The idea here is that if the variance in the treatment group is zero, we want to enforce the variance in the matched control group to be small. The choice to aim to reduce the standard deviation to one tenth that of the full control group is ad hoc. When $s_{mc}^{(k)} = 0$ but $s_t^{(k)} \neq 0$ (which is much more rare), we simply allow the badness to be infinite, thereby eliminating these few matches from consideration.

3.6. Subject-matter experts' review

The above process included extensive dialogue with subject-matter experts, who can better determine whether differences between control and treatment are of concern. We presented plots (as shown in Figure 2) to development economists, public health practitioners, geographers, and agricultural scientists. These displays allowed the experts to see the differences in means and variances discussed above. If they voiced concerns about a particular variable, we reran the above algorithm with an adjusted badness score that gives more weight to the unbalanced variable. Alternatively, we began the procedure by restricting the control pool to grid cells within a range that corresponds more closely to the treatment cells.

3.7. Selecting villages

After the selection of matched control grid cells described above, our field teams listed all villages for which a majority of households fall within each grid cell's boundary. Village names may have changed since the start of the project, ten years ago. It is not uncommon for a village to split or for a few villages to merge. Though we do not want to use post-treatment information, we need our sampling frame to reflect the current villages, so we use the most recent village lists available.

Our field teams collected population data to determine the size (numbers of households or people) of each village listed. We restrict our sampling frame to villages within the range of the corresponding MV site village sizes. If for a particular grid cell no villages are within this range, we take as a control village the one closest to the range. After determining the sampling frame of villages, we randomly select one village per grid cell to serve as our control villages.

This procedure was pre-registered with The Registry for International Development Impact Evaluations (International Initiative for Impact Evaluation, 2013), including code to perform the randomization with the promise to use a specific future NASDAQ index as a random seed. This prevents alterations to control village selection once outcome data are available.

3.8. Case studies

Instead of describing the path to our final matches for all ten countries, we use Uganda and Ghana as case studies to show some of the most common issues that arose. For ease of notation, define $d_k = \frac{|\overline{x}_t^{(k)} - \overline{x}_{mc}^{(k)}|}{\sigma^{(k)}}$, the standardized absolute mean difference and $v_k = \left|\ln \frac{s_t^{(k)}}{s_{mc}^{(k)}}\right|$ the absolute log variance ratio. Define a mean operator, $\mathop{\mathrm{M}}_{i \in S} x_i \equiv \frac{1}{|S|} \sum_{i \in S} x_i$. The original proposed badness score was

$$\begin{aligned} \text{mean\_badness} &= 10 \mathop{\mathrm{M}}_{k \in S_{DHS}} d_k + \mathop{\mathrm{M}}_{k \in S_{geo}} d_k \\ \text{var\_badness} &= 10 \mathop{\mathrm{M}}_{k \in S_{DHS}} v_k + \mathop{\mathrm{M}}_{k \in S_{geo}} v_k, \\ \text{badness} &= 2 * \text{mean\_badness} + \text{var\_badness}. \end{aligned} \tag{4}$$

We always weight the mean badness twice as much as the variance badness, so henceforward we drop this last line from our specification of badness scores.

Ghana

Optimization of the badness score resulted in a successful match in Ghana, representing our experience in eight (out of ten) countries. Unlike other countries, when we restrict selection to areas with DHS data and to the agroecological zone of the Ghana MV (tree crop), there are only two candidate control grid cells within the same district as the MV (Amansie West). Therefore, the matching procedure optimizes over the remaining three matches, which must come from outside Amansie West.

Optimizing with the original proposed badness score (4), we obtained a reasonably good match, but with population density roughly 30% higher in one control grid cell than in the treatment areas. Subject-matter experts suggested we improve the match on population. In response, we increased the weight on the population variable to match the collective importance of the DHS variables, see badness score (5). The grid cell with the highest population density was replaced with a grid cell with population roughly equal to that of the MV. After consultation with subject-matter experts, this was determined to be the final match, see Figure 2.

$$\begin{aligned} \text{mean\_badness} &= 10 \mathop{\mathrm{M}}_{k\in S_{DHS}} d_k + \mathop{\mathrm{M}}_{\substack{k\in S_{geo}\\ k\neq \mathrm{POPD}}} d_k + 10 d_{\mathrm{POPD}} \\ \text{var\_badness} &= 10 \mathop{\mathrm{M}}_{k\in S_{DHS}} v_k + \mathop{\mathrm{M}}_{\substack{k\in S_{geo}\\ k\neq \mathrm{POPD}}} v_k + 10 v_{\mathrm{POPD}} \end{aligned} \tag{5}$$

Uganda

Uganda demonstrates a case in which the badness score does not afford us a semi-automated procedure largely free from human input. In fact, the badness score itself did not drive the selection in Uganda. Instead, we used visualization and input from experts to arrive at our final matches. Our experience with the matching process in Tanzania was similar.

Optimizing with the original proposed badness score, subject matter experts were unhappy with the match on population density. We increased the weight on population density, but this made the match on travel time to major cities very poor, with little overlap. This tradeoff is easily seen via a two-dimensional plot of the two variables, see Figure 3. Controls with population density similar to the MV are closer to major cities and controls with access to major cities similar to the MV have lower population density. Both access to major cities and population density are correlated with access to health and education services, and therefore to our outcomes (Balk et al., 2004; Roberts et al., 2006; Gage, 2007; Linard et al., 2012). Thus, we take some matches that are a good match on population density, and some that are a good match on access to major cities. To do this, we dropped the variance contributions to the badness score for each of these variables to allow the matches to have higher variance than the treatment areas.

Using a badness score with only population and access to major cities,

$$\begin{aligned} \text{mean\_badness} &= 10 d_{\mathrm{POPD}} + 2 d_{\mathrm{ACCESS}} \\ \text{var\_badness} &= 0, \end{aligned} \tag{6}$$

we examined two-dimensional plots to find the relative weights we wanted to give each variable (see Figure 3b). We used this ratio of weights including other variables in the badness score,

$$\begin{aligned}
\text{mean\_badness} &= 10 \mathop{\mathrm{M}}\limits_{k \in S_{DHS}} d_k + \mathop{\mathrm{M}}\limits_{\substack{k \in S_{geo} \\ k \neq \text{POPD, ACCESS}}} d_k + 25 d_{\text{POPD}} + 5 d_{\text{ACCESS}} \qquad (7) \\
\text{var\_badness} &= 10 \mathop{\mathrm{M}}\limits_{k \in S_{DHS}} v_k + \mathop{\mathrm{M}}\limits_{\substack{k \in S_{geo} \\ k \neq \text{POPD, ACCESS}}} v_k + 0 * v_{\text{POPD}} + 0 * v_{\text{ACCESS}},
\end{aligned}$$

but the match on population density was unsatisfactory, see Figure 3c. We increased the weights on the population and access, using the other variables only as tie breakers,

$$\begin{aligned}
\text{mean\_badness} &= 10 \mathop{\mathrm{M}}\limits_{k \in S_{DHS}} d_k + \mathop{\mathrm{M}}\limits_{\substack{k \in S_{geo} \\ k \neq \text{POPD, ACCESS}}} d_k + 100000 d_{\text{POPD}} + 20000 d_{\text{ACCESS}} \qquad (8) \\
\text{var\_badness} &= 10 \mathop{\mathrm{M}}\limits_{k \in S_{DHS}} v_k + \mathop{\mathrm{M}}\limits_{\substack{k \in S_{geo} \\ k \neq \text{POPD, ACCESS}}} v_k + 0 * v_{\text{POPD}} + 0 * v_{\text{ACCESS}}.
\end{aligned}$$

The resulting match is shown in Figures 3d and 4. We also plot the final matches with each variable on a scale from the minimum value in Uganda to the maximum value in Uganda, see Figure 5.

In addition to seeing the limitations of the badness score, our experience with Uganda's matching points to the challenge of how to prioritize matching variables. The relative importance of matching variables was unclear prior to receiving feedback on candidate matches from subject-matter experts. For example, we were encouraged to include the standard deviation of elevation as a matching variable, as it captures a ruggedness of terrain that agricultural and food security experts deemed important. However, when presented with the matches, the improved match on population far outweighed the worsened match on standard deviation of elevation. The literature does not present a way to easily compare the two variables' prognostic value for our outcomes, nor does our information regarding the treatment assignment (i.e. the selection of Millennium Village sites).

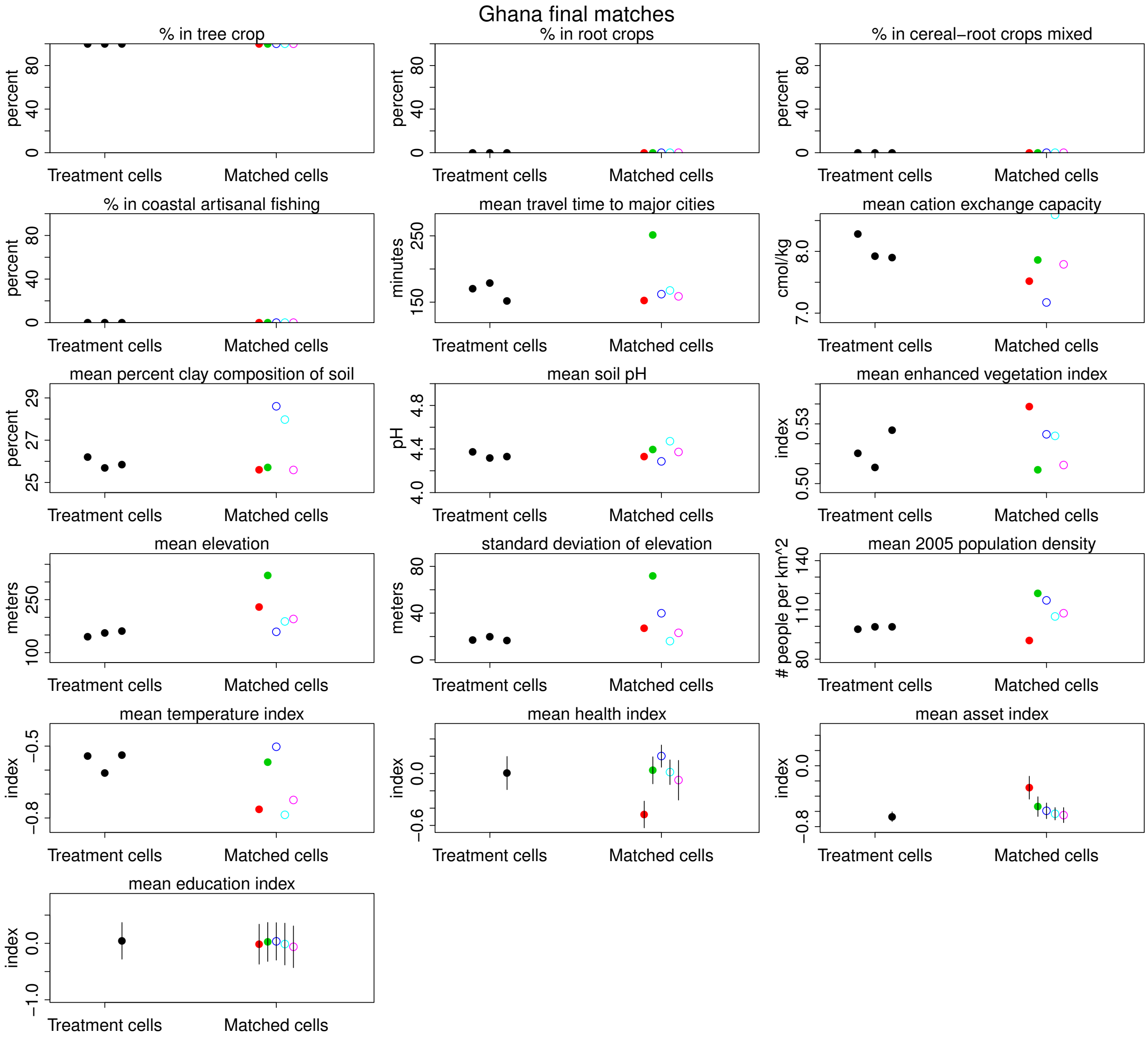
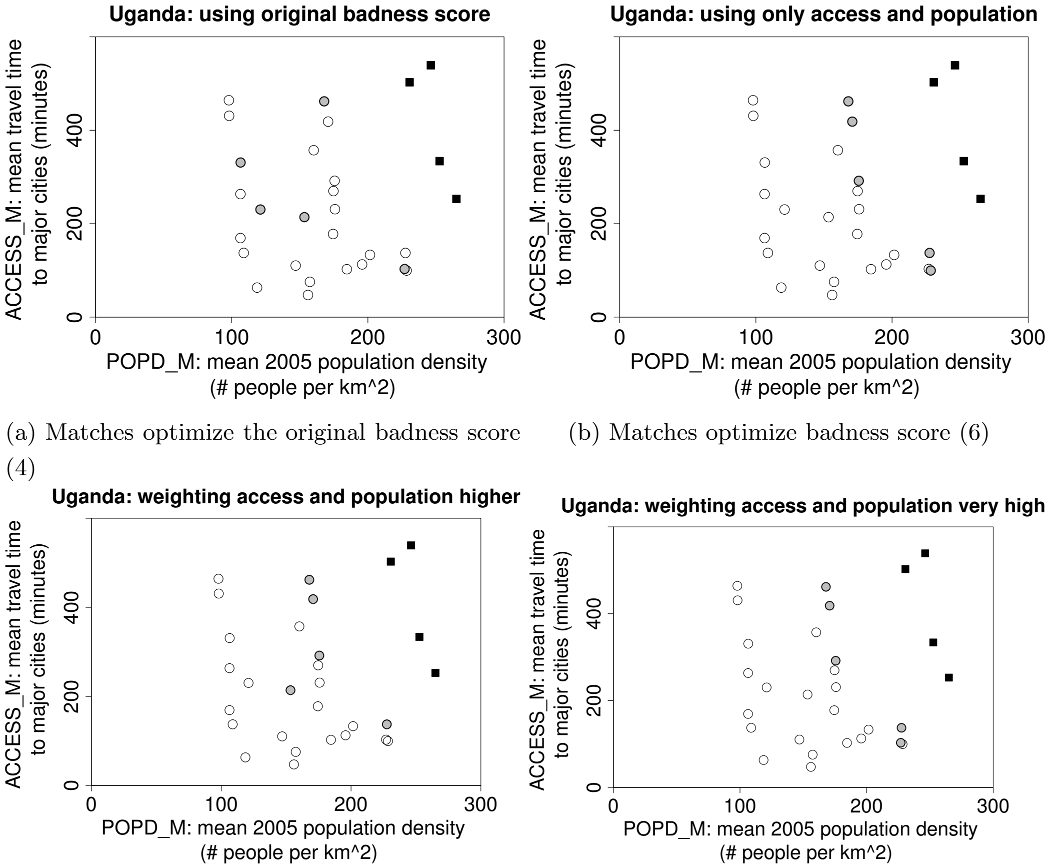


Fig. 2: This plot displays values for the matching variables in both the treatment grid cells and matched control grid cells in Ghana when the matches are found by optimizing the badness score in equation (5). Each circle corresponds to a grid cell. Black circles are treatment grid cells, while the colorful circles are the matched controls. We use the colors to identify each matched control cell, to allow for inspecting across variables (e.g. one cell/color may do well on one variable, and badly on another). Filled-in circles represent within-district grid cells, and empty circles the out-of-district grid cells. For the DHS indices (education, assets, and health), we also present the 95% posterior intervals, to represent the uncertainty from our small area estimation procedure. There are fewer black circles for these indices because only a subset of the treatment grid cells overlap DHS buffers.



(a) Matches optimize the original badness score (4)

(b) Matches optimize badness score (6)

(c) Matches optimize badness score (7)

(d) Matches optimize badness score (8)

Fig. 3: For Uganda's matching, we first exact match on agroecological zone (highland perennial), and restrict candidate matches to either Isingiro or Ruhaama districts. We restrict to areas with Demographic and Health Surveys (DHS) data outside of the district containing the Millennium Village (Isingiro), but allow non-DHS areas inside of the district, i.e. in Ruhaama. After these restrictions, the remaining candidate control grid cells are displayed as circles in this figure. We compare population density in 2005 (the average number of people per square kilometer) versus travel time to major cities (in minutes). In filled-in black squares are the four treatment grid cells. We fill in the chosen control grid cells in gray.

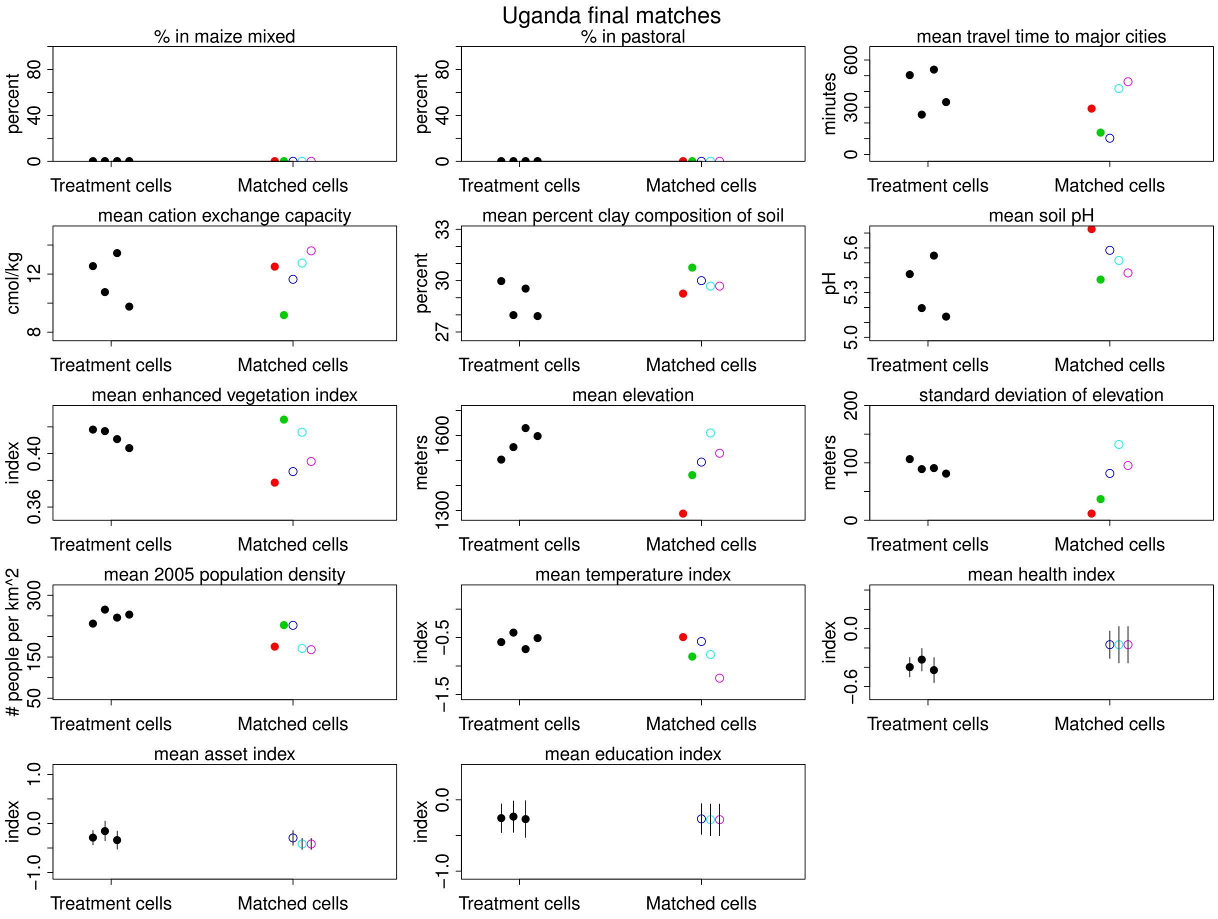


Fig. 4: This plot displays values for the matching variables in both the treatment grid cells and matched control grid cells in Uganda when the matches are found by optimizing the badness score in equation (8). Each circle corresponds to a grid cell. Black circles are treatment grid cells, while the colorful circles are the matched controls. We use the colors to identify each matched control cell, to allow for inspecting across variables (e.g. one cell/color may do well on one variable, and badly on another). Filled-in circles represent within-district grid cells, and empty circles the out-of-district grid cells. For the DHS indices (education, assets, and health), we also present the 95% posterior intervals, to represent the uncertainty from our small area estimation procedure. There are fewer black circles for these indices because only a subset of the treatment grid cells overlap DHS buffers.

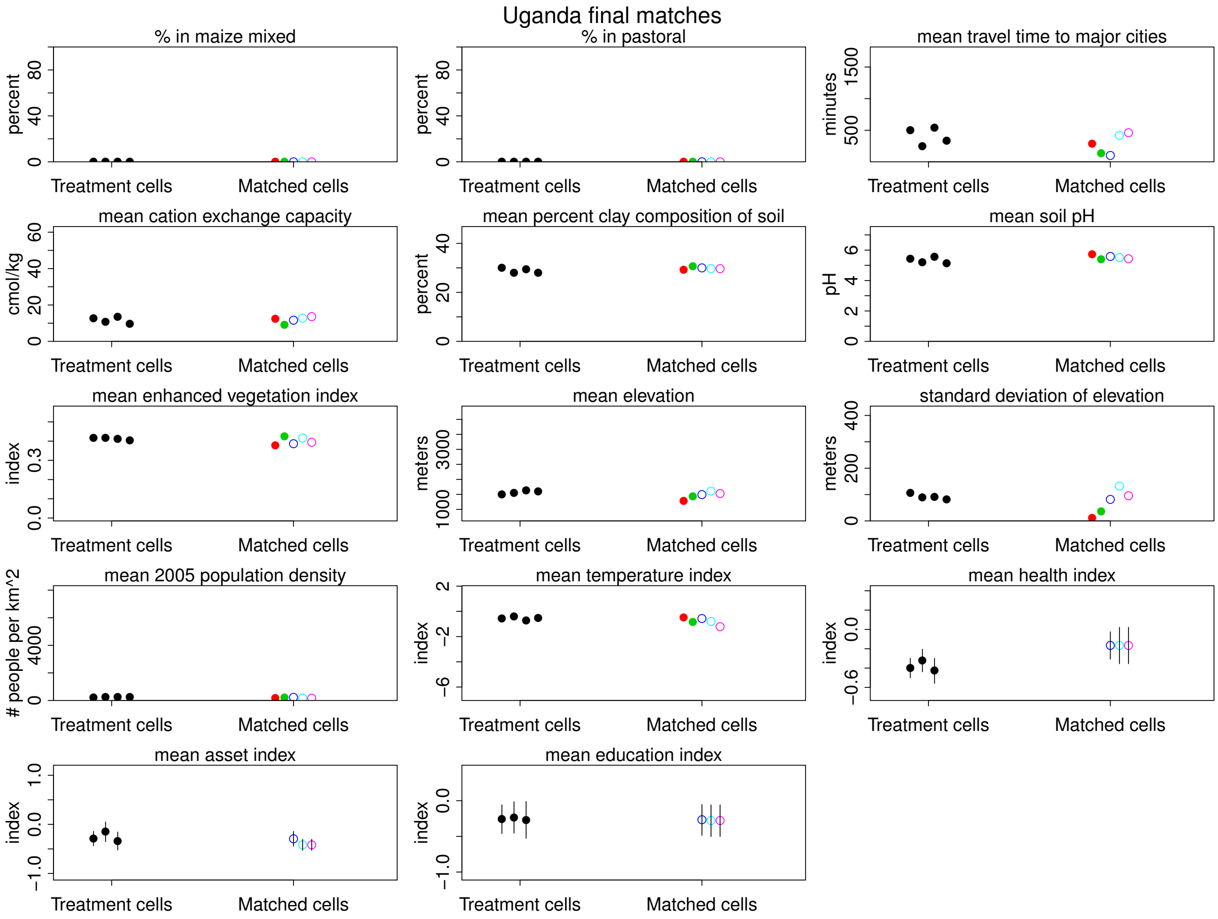


Fig. 5: This plot displays values for the matching variables in both the treatment grid cells and matched control grid cells in Uganda when the matches are found by optimizing the badness score in equation (8). Each circle corresponds to a grid cell. Black circles are treatment grid cells, while the colorful circles are the matched controls. We use the colors to identify each matched control cell, to allow for inspecting across variables (e.g. one cell/color may do well on one variable, and badly on another). Filled-in circles represent within-district grid cells, and empty circles the out-of-district grid cells. For the DHS indices (education, assets, and health), we also present the 95% posterior intervals, to represent the uncertainty from our small area estimation procedure. There are fewer black circles for these indices because only a subset of the treatment grid cells overlap DHS buffers. The axes for these three indices extend from the minimum value in Uganda to the maximum value in Uganda, in order to provide context.

4. Candidate models for causal inference

We have many outcomes of interest, defined in Mitchell et al. (2015a), including a subset of Millennium Development Goal (MDG) indicators and a number of indicators that are relevant to systems delivery, which we refer to as 'MVP indicators.' In this section, we suggest some causal models that we will fit to the end-line outcome data. The analysis will fork in many ways, with different modeling choices. In the end-line evaluation, we will report and compare all results to reduce the scope for fishing (i.e. deciding to report a model based on the realization of the conclusion, see Humphreys et al. (2013); Gelman and Loken (2013)).

Our strategy will begin with single-outcome models whose results will serve as a type of data summary. From the single outcome models we will build up to a multi-outcome model that includes all outcomes and will allow the treatment effects on related indicators to inform each other via partial pooling, as recommended in Gelman and Tuerlinckx (2000) and Gelman et al. (2012). We define groups of related indicators based on the domains that they address, as follows:

- poverty indicators: composed of our MDG 1 indicators, MVP agriculture indicators (a.1 to a.4), and MDG indicator 8.15 (access to mobile phones);
- education indicators: composed of our MDG 2 and 3 indicators, and MVP education indicators (b.1 to b.3);
- child health indicators: composed of our MDG 4 and 7 indicators, and MVP health indicator c.1;
- maternal health indicators: composed of our MDG 5 indicators; and
- HIV-malaria indicators: composed of our MDG 6 indicators.

Our data summary begins by fitting single-outcome models separately to each indicator from Mitchell et al. (2015a), and reporting all results. Next, for each of the above groups of indicators, we will create a summary measure using the treatment effect estimates from the single-outcome model regressions. With many separate analyses, there may be concerns about multiple comparisons: the idea that testing many hypotheses makes it very likely that at there will be at least one assertion of statistical significance (i.e. an uncertainty interval for the treatment effect not including zero), even when all null hypotheses are true (i.e. all treatment effects are exactly zero). As one way to alleviate these concerns, we will reduce the number of comparisons by creating two overall summary measures: one of all the indicators and one limited to the Millennium Development Goal indicators and proxies. Later in this section we give a more complete perspective on multiple comparisons.

We will create these summary measures as follows: we standardize country- and outcome-specific treatment effects using the "divide by 4 rule" for binary outcomes and dividing continuous outcomes by twice their respective standard deviations in the control group (see Section 6, and Gelman (2008); Clingingsmith et al. (2009)). Next, we will reorient the treatment effects so that larger is better. We will then average the transformed treatment effects, weighting all outcomes equally. Lastly, we will average across countries q and outcomes k:

$$\frac{1}{\text{number of countries}}\sum_q \frac{1}{\text{number of outcomes}}\sum_k \frac{\tau_q^{(k)}}{2\sigma_q^{(k)} I(k \text{ is continuous}) + 4I(k \text{ is binary})},$$

where the average is either across all outcomes (to create the overall summary measure) or across all outcomes within a group (to create group-level summaries). These summary measures are sometimes referred to as average effect size estimates (O'Brien, 1984; Clingingsmith et al., 2009). The multi-outcome model will have parameters that correspond to these group-level and overall summary measures. We expect that the average treatment effect estimate will be similar for the multi-outcome model and the above constructed summary. However, the group-level treatment effects will be shrunk towards each other in the multi-outcome model.

Before introducing our causal models, we describe their common structure and notation. Causal inferences can be biased if we adjust for variables affected by treatment (Rosenbaum, 1984), so we restrict to adjusting for pre-treatment variables. Additionally, we are limited to adjusting for aggregate baselines. We only have panel (i.e. longitudinal) data in the MVs (see Section 7), but not in control villages, since deidentification of the data from external surveys (e.g. Demographic and Health Surveys, as well as country censuses) prevents us from identifying the villages and individuals surveyed in the past.

After the selection of comparison grid cells described in Section 3, we will (randomly) select one village per grid cell. Though many of our pre-treatment variables are measured at the grid cell level, for clarity of exposition we do not present our models with a grid cell level included. However, we propose to include a grid cell level as a diagnostic during model assessment. Let j index a village, z_j be the indicator of treatment, and $\mathbf{x}_j$ be a vector of pre-treatment covariates (indexed by l), including small area estimates of wealth, education, and health indices, see Section 3.3. These pre-treatment covariates are mostly measured at the grid cell level, and not at the village level. Lastly, let $y_i^{(k,t)}$ denote the individual-level outcome k at time t (similarly let $y_j^{(k,t)}$ denote the village-level outcome).

Where not otherwise specified, priors on parameters are weakly informative. Our estimands are superpopulation average treatment effects, conditional on covariates (Gelman et al., 2014, Chapter 8). Thus, we imagine that the villages were "sampled" from a population of villages with similar covariates, with high levels of political buy-in, where MVP treatment would not have been disrupted by financing shortages or political instability (Mitchell et al., 2015a). We will perform posterior predictive checks on the models proposed below in order to iteratively adjust them, expanding when necessary (Gelman et al., 2014, Chapters 6 and 7).

4.1. Single-outcome models

We consider a ladder of models, starting with simple models and building to more complex models. The first few rungs of the ladder include only one outcome at a time, and treatment effects that do not vary across countries. The first rung of the ladder includes no covariates, and pools across countries. For each outcome k that is continuous we will fit a linear model,

$$y_i^{(k,2015)} \sim N\left(\delta_0^{(k)} + \tau^{(k)} z_{j[i]}, \sigma_y^2\right) \text{ for individuals } i,$$

where $\tau^{(k)}$ is the treatment effect for outcome k. The second rung of the ladder includes no covariates but does include partial pooling over villages and countries. While the third rung of the ladder includes covariates $(\mathbf{x}_j)$ as well,

$$\begin{aligned} \widehat{y}_j^{(k,2015)} &\sim N\left(\mathbf{x}_j^T \boldsymbol{\delta}_{q[j]}^{(k)} + \tau^{(k)} z_j, \sigma_{\text{village}}^2 + v_j\right) \text{ for villages } j \\ \boldsymbol{\delta}_q^{(k)} &\sim N\left(\boldsymbol{\delta}^{(k)}, \Sigma^{(k)}\right) \text{ for countries } q \\ \Sigma^{(k)} &= \text{diag}(\boldsymbol{\sigma}_\delta)\Omega\text{diag}(\boldsymbol{\sigma}_\delta) \\ \sigma_{\delta,l} &\sim \text{Cauchy}(0, 2.5) \text{ for covariates } l \\ \Omega &\sim \text{LKJcorr}(2), \end{aligned} \tag{9}$$

where $\widehat{y}_j^{(k,2015)}$ is the estimated village-level indicator, and v_j its variance. We use a separation-strategy prior that decomposes the variance-covariance matrix into variance and correlation components, specifying separate priors for each component (McCulloch and Meng, 2000). We place weakly informative priors on the variances recommended by Gelman (2006), and a weakly informative prior on the correlation matrix whose probabilities are inversely proportional to its determinant (Lewandowski et al., 2009). For binary outcomes, we will fit analogous logistic models, and in the next section we describe models for the mortality outcomes.

We may relax the exchangeability of villages within country via additional levels to the models, or a conditional autoregressive (CAR) spatial model. We propose to add interactions between the treatment indicator and covariates to assess sensitivity to the assumption that the coefficients of the pre-treatment covariates do not vary by treatment group. However, we may not have the precision to estimate these interactions without strong regularization via prior distributions. For example, one model we propose to fit will interact the linear predictor with the treatment indicator, replacing the village level of the above model with $\widehat{y}_j^{(k,2015)} \sim N\left(\mathbf{x}_j^T \boldsymbol{\delta}_{q[j]}^{(k)} + \tau^{(k)} z_j + \gamma \mathbf{x}_j^T \boldsymbol{\delta}_{q[j]}^{(k)} z_j, \sigma_{\text{village}}^2 + v_j\right)$ for villages j. We center the $\mathbf{x}_j$ so that $\tau^{(k)}$ can be interpreted as a superpopulation average treatment effect, $E[y_j(1) - y_j(0)]$.

Mortality outcomes - survival models

For mortality indicators, standard methods used by the DHS are described in Rutstein and Rojas (2006, p.92-94). We can use these methods to compute village-level mortality rates, and fit a village-level model. Alternatively, we can fit a survival model. For the under-5 mortality rate, the end-line study period is 2010-2015, following the conventions in Rutstein and Rojas (2006); UN Millennium Project (2014). With women's birth histories collected in 2015, we will have birth and death dates (if the child died) for any child age 0-5 years alive during this study period. The complications with considering under-5 mortality in 2010-2015 are: we want a child born before 2010 to contribute to the analysis only during the study period, and we want only ages 0-5 to contribute to the analysis. To accomplish this we propose the following method:

Let J_{0i} be child i's *joining time*, which equals 2010 for children born before 2010, and equals the calendar year of birth for children born after 2010. Let A_{0i} be child i's *age adjustment*, which equals the child's age in 2010 for children born before 2010, and equals zero for children born after 2010. Let T_i^* be child i's survival time, i.e. how many years child i lives in total. Then $T_i = T_i^* - A_{0i}$ is the survival time since the joining time J_{0i}. The censoring time in years since the joining time is $C_i = \min(5 - A_{0i}, 2015 - J_{0i})$ because children born before 2010 are censored when they reach age 5 and those born after 2010 are censored in 2015. The observed data are $(U_i, \delta_i, \mathbf{x}_{j[i]})$ where $U_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$ indicates whether the child died, and $\mathbf{x}_{j[i]}$ are covariates, including treatment indicator, country effect, and other variables.

C_i may depend on T_i because both may depend on A_{0i}: for children born before 2010, $C_i = 5 - A_{0i}$ while $T_i = T_i^* - A_{0i}$. Thus, we want to condition on A_{0i} in our analysis so that C_i and T_i are more plausibly independent. We also want to condition on J_{0i} because otherwise $(U_i, \delta_i, \mathbf{x}_{j[i]})$ may not be i.i.d. (independent and identically distributed): for a child with a smaller value of J_{0i}, the observation $(U_i, \delta_i, \mathbf{x}_{j[i]})$ is more likely to be $(T_i, 1, \mathbf{x}_{j[i]})$, while for a child with a larger value of J_{0i} (but same value of covariates $\mathbf{x}_{j[i]}$), the observation $(C_i, 0, \mathbf{x}_{j[i]})$ is more likely. In addition to including A_{0i} and J_{0i} as covariates, we need to include the interaction of A_{0i} and treatment in order to account for the possible benefits that children born before 2010 may have had from getting the treatment for a few years prior to joining the study period.

Finally, we fit a survival analysis model (Cox, 1972; Ibrahim et al., 2001) adjusting for the variables mentioned above, analogous to the proposed models in the previous section. The coefficient of treatment, $\tau^{(k)}$, represents a log hazard ratio, comparing the hazard of death among children in a treatment village to those in a control village, among children with the same covariates adjusted for in the model , ages 0-5 during 2010-2015. We can also use the model to compute other summaries of the treatment effect (including the difference or ratio of the probability of a child surviving to age 5 in treatment versus control villages) by estimating the baseline survivor function.

Difference-in-differences methods

Previous evaluations of the MVP, Clemens and Demombynes (2011) and Pronyk et al. (2012), as well at the proposal for the new northern Ghana MV evaluation, ITAD evaluation for Northern Ghana (2013), use *difference-in-differences* methods. Difference-in-differences uses measurements at two time points, baseline and end-line (and possibly also time points in between), and an assumption of *additivity* to difference out time-invariant effects and identify the effect of treatment. Additivity requires the potential gains over time to be the same across treatment and comparison groups, adjusted for covariates.

Instead of additivity, our models above, often known as *ANCOVA models*, assume unconfoundedness given the baseline outcome variables and other covariates. Difference-in-differences and ANCOVA models each make different assumptions, neither makes strictly fewer assumptions than the other (Imbens and Wooldridge, 2009, p.70). Imbens and Wooldridge (2009) suggest that unconfoundedness given baseline is, in general, more attractive. To test the sensitivity to these assumptions, we propose to fit difference-in-differences models analogous to our above models, for any outcome k for which we have an estimate of baseline. If there are large discrepancies between the two types of models, we will have to conclude that we are uncertain which to trust.

Without individual-level data at baseline in the control villages, we cannot fit an individual-level difference-in-differences model. In a difference-in-differences model analogous to model 9, the estimated baseline is subtracted from the 2015 outcome and not included in covariates $\mathbf{x}_j$. This enables us to difference out any time-invariant grid cell-level effects.

Varying treatment effects

We propose to fit our above models allowing for treatment effects to vary by Millennium Village (i.e. country), with partial pooling (Hill and Scott, 2009; Feller and Gelman, 2014). Extending our model 9, we will fit a model that allows the treatment effects to vary by country, correlated with the pre-treatment covariates linear predictor:

$$\begin{aligned}
\widehat{y}_j^{(k,2015)} &\sim N\left(\mathbf{x}_j^T \boldsymbol{\delta}_{q[j]}^{(k)} + \tau_{q[j]}^{(k)} z_j + \gamma \mathbf{x}_j^T \boldsymbol{\delta}_{q[j]}^{(k)} z_j, \sigma^2_{\text{village}} + v_j\right) \text{ for villages } j \\
\tau_q^{(k)} &\sim N\left(\tau_0^{(k)} + \omega \overline{\mathbf{x}}_q^T \boldsymbol{\delta}_q^{(k)}, \sigma_\tau^2\right) \text{ for countries } q \\
\boldsymbol{\delta}_q^{(k)} &\sim N\left(\boldsymbol{\delta}^{(k)}, \Sigma^{(k)}\right) \text{ for countries } q \\
\Sigma^{(k)} &= \text{diag}(\boldsymbol{\sigma}_\delta)\Omega\text{diag}(\boldsymbol{\sigma}_\delta) \\
\sigma_{\delta,l} &\sim \text{Cauchy}(0, 2.5) \text{ for covariates } l \\
\Omega &\sim \text{LKJcorr}(2).
\end{aligned}$$

Due to the small numbers of villages in each country, estimates of $\tau_q^{(k)}$ will have high variance unless $\sigma^2_{\text{village}}$, the unexplained variance between villages, is small. In this model we center the $\mathbf{x}_j$ by country means so that $\tau_q^{(k)}$ can be interpreted as a superpopulation average treatment effects for each country.

Accounting for uncertainty from small area estimation

A subset of variables in $\mathbf{x}_j$ will be small area estimates and we want our intervals for the treatment effects to honestly reflect the uncertainty in our procedure. To account for uncertainty in each such x_j, we add a level to the hierarchical causal models: $x_j \sim N\left(\widehat{x}_j, var_j\right)$, where var_j is the posterior variance from the small area estimation procedures described in 3.3 (see Gelman et al. (2014, p.474) for a similar example). We may transform x_j to make normality a better approximation to the posterior distribution.

4.2. Joint-outcome models

Our outcomes $k = 1, ..., K$ (where $K = 51$, see Mitchell et al. (2015a)) target different populations (e.g. infants, women, etc.). These outcomes are grouped into five related groups of outcomes (poverty, education, child health, maternal health, HIV-malaria) indexed by $g = 1, ..., 5$.

We standardize the outcomes so that they are on the same scale, and the positive direction is better (higher standard of living). To avoid the issue of different populations, the joint-outcome model we propose uses estimates of the outcomes at the village level, $\widehat{y}_j^{(k,2015)}$, with estimated variance-covariance matrix $\widehat{\Sigma}_y$:

$$\begin{bmatrix} \vdots \\ \widehat{y}_j^{(k,2015)} \\ \vdots \end{bmatrix} \sim N\left(\begin{bmatrix} \vdots \\ \theta_j^{(k)} \\ \vdots \end{bmatrix}, \widehat{\Sigma}_y\right) \text{ for villages } j$$

$$\begin{aligned}
\theta_j^{(k)} &\sim N(\theta_j^{(g[k])} + \mathbf{x}_j^T \boldsymbol{\delta}_{q[j]}^{(k)} + \tau_{q[j]}^{(k)} z_j, \sigma^2_{\text{village,within-group}}) \text{ for villages } j \text{ and outcomes } k \\
\theta_j^{(g)} &\sim N(0, \sigma^2_{\text{village,between-group}}) \text{ for villages } j \text{ and outcome groups } g \\
\tau_q^{(k)} &\sim N(\omega \overline{\mathbf{x}}_q^T \boldsymbol{\delta}_q^{(k)} + \tau_q^{(g[k])}, \sigma^2_{\tau,\text{within-group}}) \text{ for countries } q \text{ and outcomes } k \\
\tau_q^{(g)} &\sim N(\tau_q, \sigma^2_{\tau,\text{between-group}}) \text{ for countries } q \text{ and outcome groups } g \\
\tau_q &\sim N(\tau, \sigma^2_{\tau,\text{between-country}}) \text{ for countries } q \\
\boldsymbol{\delta}_q^{(k)} &\sim N\left(\boldsymbol{\delta}^{(k)}, \Sigma^{(k)}\right) \text{ for countries } q \\
\Sigma^{(k)} &= \text{diag}(\boldsymbol{\sigma}_\delta)\Omega\text{diag}(\boldsymbol{\sigma}_\delta) \\
\sigma_{\delta,l} &\sim \text{Cauchy}(0, 2.5) \text{ for covariates } l \\
\Omega &\sim \text{LKJcorr}(2).
\end{aligned}$$

4.3. Multiple comparisons

The probability of making at least one error is larger with estimation of many outcomes than with a single outcome. This is the concern of "multiple comparisons," and it is usually framed in terms of Type I errors: asserting statistical significance (i.e. an uncertainty interval for the treatment effect not including zero) even when the null hypothesis is true (i.e. the treatment effect is exactly zero). One way to circumvent this concern is to consider only our overall summary measure proposed above or the corresponding parameter τ in the joint model.

Another perspective views the null hypothesis as an uninteresting event to condition on, because we do not believe that any treatment effect is exactly zero. This is unrelated to the Millennium Villages Project, but is a general statement about continuous parameters: they equal zero with probability zero. However, it is plausible that all the treatment effects are close to zero relative to the error in the data. In this situation, claims that the treatment effects are nonzero (i.e. statistical significance) are not errors, but two different errors can arise. *Type M* (magnitude) error is the expected absolute value of the estimate divided by the true effect size, if it is statistically significant (Gelman and Carlin, 2013). *Type S* (sign) error is the probability that the estimated treatment effect has the incorrect sign, if it is statistically significant. Gelman and Tuerlinckx (2000) and Gelman et al. (2012) show that hierarchical modeling can reduce both these types of errors.

5. Assessing unconfoundedness and sensitivity analysis

Although unconfoundedness cannot be tested directly, there are analyses that can assess its plausibility (Altonji et al., 2005; Imbens and Rubin, 2015, Chap.21). Imbens and Rubin (2015) describe three methods, one of which can be done before outcome data are available. Sensitivity analyses relax (rather than assess) unconfoundedness, obtaining ranges of plausible values for the treatment effects. We conduct the one analysis which can be done at the design stage (i.e. before outcome data are available): assessment of unconfoundedness using pseudo-outcomes. We also outline the remaining two ways to assess unconfoundedness and our proposed sensitivity analyses.

Considering outcomes at the grid-cell level, the unconfoundedness assumption is:

$$y_c(0), y_c(1) \perp z_c \mid \mathbf{x}_c \text{ (unconfoundedness).} \tag{10}$$

A related assumption is subset unconfoundedness, which leaves out the pth pre-treatment variable from the conditioning set:

$$y_c(0), y_c(1) \perp z_c \mid \mathbf{x}_c^{(-p)} \text{ (subset unconfoundedness).} \tag{11}$$

This assumption cannot be tested for the same reason unconfoundedness cannot be tested: we do not observe $y_c(1)$ if $z_c = 0$ and we do not observe $y_c(0)$ if $z_c = 1$ (Imbens and Rubin, 2015). Suppose, however, that one of our pre-treatment variables is a good proxy for one of the potential outcomes (e.g. $y_c(0)$). This variable, $x_c^{(p)}$, can serve as a pseudo-outcome in a testable version of unconfoundedness:

$$x_c^{(p)} \perp z_c \mid \mathbf{x}_c^{(-p)} \text{ (pseudo-outcome unconfoundedness).} \tag{12}$$

The link between the unconfoundedness assumption (10) and the testable assumption (12) depends on two steps: linking assumptions (10) and (11) and linking assumptions (11) and (12). Both links are based on heuristic arguments that rely on subject-matter knowledge, neither are probabilistic theorems.

While it is theoretically possible that subset unconfoundedness (11) holds but unconfoundedness (10) does not, in practice it is rare if all the $\mathbf{x}_c$ are pre-treatment variables. More concerning is the more plausible scenario that unconfoundedness (10) holds but subset unconfoundedness (11) does not, because conditioning on $x_c^{(p)}$ is critical.

Subset unconfoundedness (11) and pseudo-outcome unconfoundedness (12) are most closely related when $x_c^{(p)}$ serves as a good proxy for $y_c(0)$ or $y_c(1)$. This is most plausible when $x_c^{(p)}$ is a lagged version of the outcome (Imbens and Rubin, 2015). In our analysis, the DHS variables are composites of outcome measures and are therefore some of the best pseudo-outcomes. However, for this same reason they might be critical to condition on, calling into question the subset unconfoundedness assumption.

For pseudo-outcomes, we consider only continuous (not categorical) variables. We consider nine geographic variables: access to major cities, cation exchange capacity of the soil, percent clay composition of the soil, soil pH, enhanced vegetation index, land surface temperature, elevation, elevation standard deviation (i.e. roughness of terrain), and population density. From the DHS we consider three variables: an asset wealth index, education index, and health index, see Appendix B.

We always include categorical variables (agroecological zone and district or neighboring districts) in $\mathbf{x}_c^{(-p)}$, and perform exact-matching as described in Section 3.1. We then use the continuous variables (except for the pseudo-outcome) in matching procedures described in Section 3.4. Lastly, we fit a simple hierarchical model,

$$\begin{aligned} x_c^{(p)} &\sim N\left(\delta_{0,q[c]} + \mathbf{x}_c^{(-p)^T}\boldsymbol{\delta} + \tau z_c, \sigma^2_{\text{grid-cell}}\right) \text{ for grid cells } c \qquad (13)\\ \delta_{0,q} &\sim N(\delta_0, \sigma^2_\delta) \text{ for countries } q. \end{aligned}$$

In addition to this model, we also conduct t-tests of the pseudo-outcome between treatment and matched control groups.

We perform the above procedure (matching, fitting model 13, and conducting t-tests) for each pseudo-outcome, recording each treatment effect interval of uncertainty. We only have DHS data for both treatment and control groups in seven of the ten countries (see Section 3). Therefore, we split our assessment of unconfoundedness into two parts. In one part, we drop the DHS variables from the pre-treatment covariates and pseudo-outcomes and perform the procedure using data from all ten countries. In the other part, we include DHS variables and limit our analysis to data from the seven countries with DHS data.

Without DHS variables, using data from all ten countries, we assess the nine possible geographic pseudo-outcomes by examining the treatment effect interval of uncertainty from fitting model 13. Only enhanced vegetation index had an interval that did not contain zero. With DHS variables, using data from only seven countries, we assess all twelve possible pseudo-outcomes by examining the treatment effect interval of uncertainty from fitting model 13. Cation exchange capacity of the soil, elevation, and population density all had intervals that did not contain zero. None of the t-test results were significant. These four variables are not particularly compelling pseudo-outcomes, and neither our matching procedure nor modeling strategy included inspections or checks (e.g. examining plots such as Figure 2 to inspect the matching, or posterior predictive checks of model fit). Therefore, we do not (yet) abandon attempts at causal inference.

A second method of assessing unconfoundedness splits the comparisons into two groups and estimates the treatment effect with "treatment" equal to the group variable (Imbens and Rubin, 2015, Chap.21). In the MVP setting, splitting the few comparison areas (five per country) in two may result in poor balance on pre-treatment variables. Therefore, this pseudo-treatment may be found to be significant, even if unconfoundedness holds. A third method looks at robustness to the set of pre-treatment variables, comparing treatment effects based on different versions of subset unconfoundedness (11). We propose to implement these two approaches once outcome data are available.

Additionally, in our final evaluation report, we will conduct a variety of analyses to assess sensitivity to the unconfoundedness assumption. In particular, we will use the assumption-free results of Ding and VanderWeele (2015) to produce a bound on the treatment effect, creating plots similar to their Figure 1 on p.15, showing the extent of confounding required to explain away estimated treatment effects. The results of Ding and VanderWeele (2015) handle binary and nonnegative§ outcomes, on the odds ratio, risk ratio, or difference scales. We also propose parametric sensitivity analysis that assumes a particular model, using ideas and software from Carnegie et al. (2015a,b). We will extend models (4.2)-(4.4) in Carnegie et al. (2015b) to include a hierarchical structure, and create plots similar to Figure 1 on p.16, showing true treatment effects given the observed data and an assumed level of confounding.

6. Design analysis

We perform design analysis (i.e. power calculations) to recommend the number of control villages and magnitude of sampling within each (Gelman and Hill, 2007). We examine four outcomes: annualized consumption (a measure of income), weight for age z-score, measles immunization, and bednet usage. We simplify the simulations by considering each MV as a

§Outcomes can be made nonnegative if they are bounded from either below or above and transformed. For example, with a lower-bound on weight for age z-scores of -10 (presumably no one can be alive below such a z-score), all values can be shifted by 10 and the sensitivity bounds derived in Ding and VanderWeele (2015) can be applied.

single village. This is justifiable because the villages compromising each MV are contiguous, and are plausibly more highly correlated than the control villages, which are more spatially scattered. We consider the intra-household correlation to be zero (equivalently, that we sample one person per household) and assume simple random sampling of individuals and households, no nonresponse, and that treatment effects and coefficients of pre-treatment covariates are constant across countries. For continuous outcomes our model to generate data and estimate treatment effects is

$$\begin{aligned} \overline{y}_j^{(k)} &\sim N\left(\delta_{0,q[j]}^{(k)} + \mathbf{x}_j^T \boldsymbol{\delta}^{(k)} + \tau^{(k)} z_j, \sigma_{\text{village}}^2 + \sigma_y^2/n_j\right) \text{ for villages } j \\ \delta_{0,q}^{(k)} &\sim N(\delta_0^{(k)}, \sigma_\delta^2), \\ \mathbf{x}_j &\sim N\left(\widehat{\mathbf{x}}_j, \Sigma_{SAE}\right), \end{aligned} \tag{14}$$

where z_j is an indicator of treatment for village j, $\mathbf{x}_j$ is a vector of the true village-level covariates, and $\widehat{\mathbf{x}}_j$ are small area estimates. We account for small area estimation uncertainty as described in Section 4 with Σ_{SAE}, a diagonal matrix whose elements are the posterior variances from the small area estimation procedures. For binary outcomes (measles immunization and bednet usage), we fit an analogous model.

We simulate imperfect matching by drawing pre-treatment variables from a Normal distribution centered at the MVP baselines with a ten percent coefficient of variation. We consider these generated values to be estimates from a small area estimation model, the $\widehat{\mathbf{x}}_j$ in the model above. We compare power with large (50% coefficient of variation) posterior variance from small area estimation (the diagonal of Σ_{SAE}) to zero baseline uncertainty, see Figure 7.

We use MV data from years 0 and 5 (2005 and 2010) to obtain reasonable values for $\boldsymbol{\delta}^{(k)}$, σ_δ and σ_y by taking posterior medians from fitting the following model,

$$\begin{aligned} y_i^{(k,2010)} &\sim N\left(\delta_{0,q[i]}^{(k)} + \mathbf{x}_{j[i]} \boldsymbol{\delta}^{(k)}, \sigma_y^2\right) \text{ for individuals } i \\ \delta_{0,q}^{(k)} &\sim N(\delta_0^{(k)}, \sigma_\delta^2), \end{aligned} \tag{15}$$

with the analogous logistic regression for binary outcomes. To set reasonable values for σ_{village} we fit a basic hierarchical model to DHS data, whose clusters are of similar order of magnitude to the MVs.

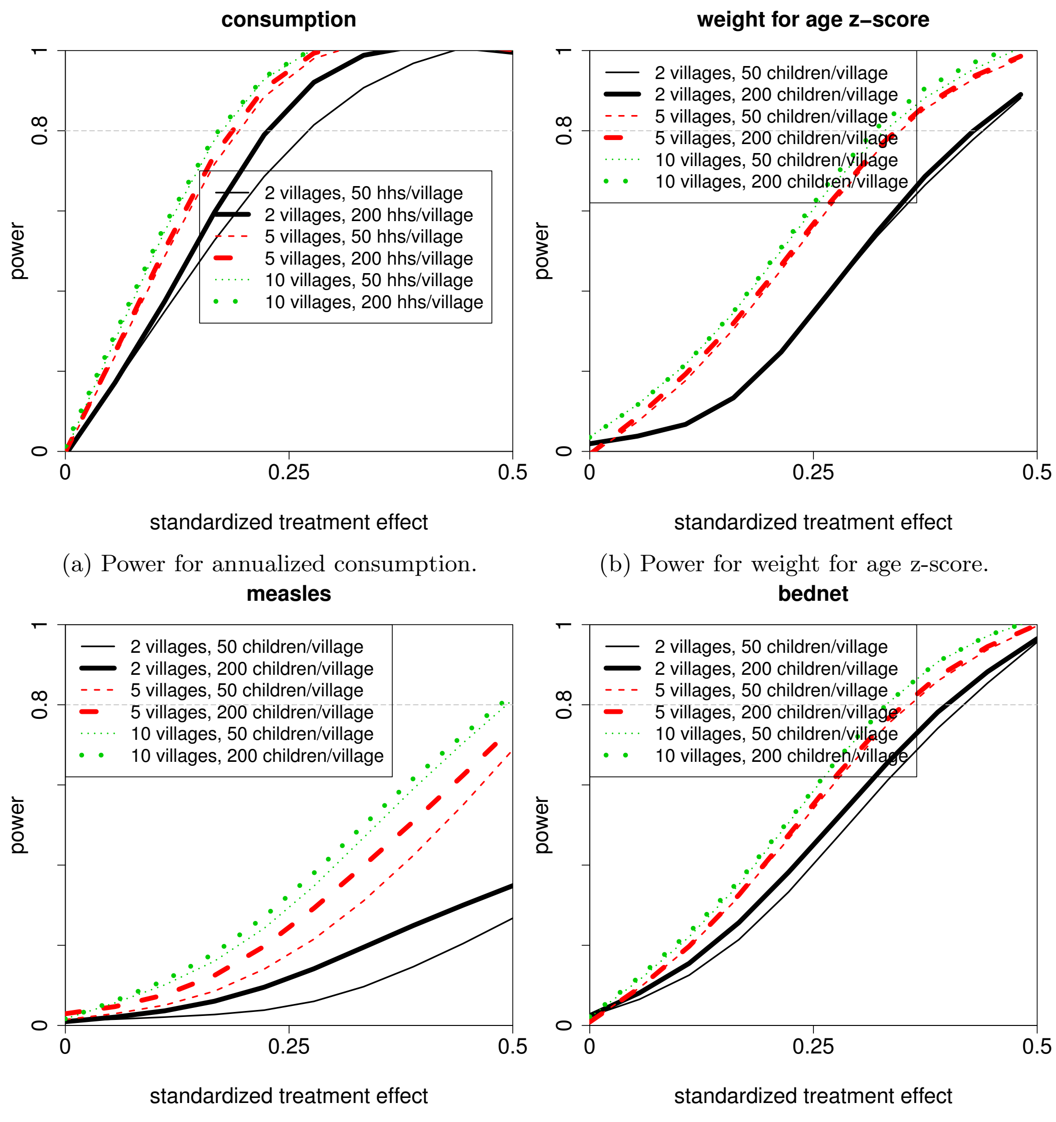
Using these values for the parameters and baselines, we generate data from model 14 (and the analogous logistic model) taking $\tau^{(k)}$ to be a range of values (see our standardization described in the next paragraph). We fit these same models to obtain estimates of treatment effect $\tau^{(k)}$. We compute, via simulation, the power (the probability that the estimated treatment effect is statistically significant) for each value of $\tau^{(k)}$, for 50 or 200 individuals per control village, and for either 2, 5, or 10 control villages matched to each MV. In each MV we sample 300 individuals due to recommendations for the adequacy component of the evaluation (Mitchell et al., 2015a).

We standardize the treatment effects across outcomes, dividing continuous outcomes by twice their standard deviations (Gelman, 2008), and dividing logistic regression coefficients by four (Gelman and Hill, 2007, p.82). We take treatment values ranging from zero to one half on this scale, i.e. zero to one standard deviation in the outcome.

In Figure 6 we plot power as a function of treatment effect for the four outcomes and in Figure 7 we examine different simulation conditions for the weight for age z-score outcome. Results for the difference-in-differences versions of the models yielded similar results. The usual gains in efficiency from ANCOVA models (as compared to difference-in-differences, see McKenzie (2012)) were not seen here, perhaps because the baselines are not at the individual level, but rather, at the village level.

The plots in Figure 6 show that increasing the number of households (or individuals) sampled per control village from 50 to 200 does not improve the power substantially. Similarly, increasing the number of control villages per Millennium Village from five to ten does not result in large gains in power. These patterns are due to the fixed number of treatment clusters and the intra-village correlation, as can be seen by examining Figure 7d, which shows results when intra-village correlation is set to zero. In Figure 7b we see that increasing the sample size in the MVs from 300 to 600 households (or individuals) does not improve the power substantially. In contrast, we see in Figure 7c that lowering the baseline uncertainty to zero does appreciably increase the power. These conclusions led us to recommend sampling five control villages, with 50 households sampled per control village. Furthermore, we will work to improve our small area estimates in parallel with data collection.

In Figure 8, in Appendix C, we plot the *Type M* (magnitude) error, the expected absolute value of the estimate divided by the true effect size, if it is statistically significant (Gelman and Carlin, 2013). We see that when the true treatment effect is small, this exaggeration factor is large. We obtain similar results for the *Type S* error, the probability that the estimated treatment effect has the incorrect *sign*, if it is statistically significant. The models we fit in this design analysis use flat priors for the treatment effects, so when the true treatment effect is small, the rate of Type S errors is near 50%, dropping off as the true treatment effect gets larger (Gelman and Tuerlinckx, 2000). We propose to reduce these errors through partial pooling, as in the joint model proposed in Section 4.2.



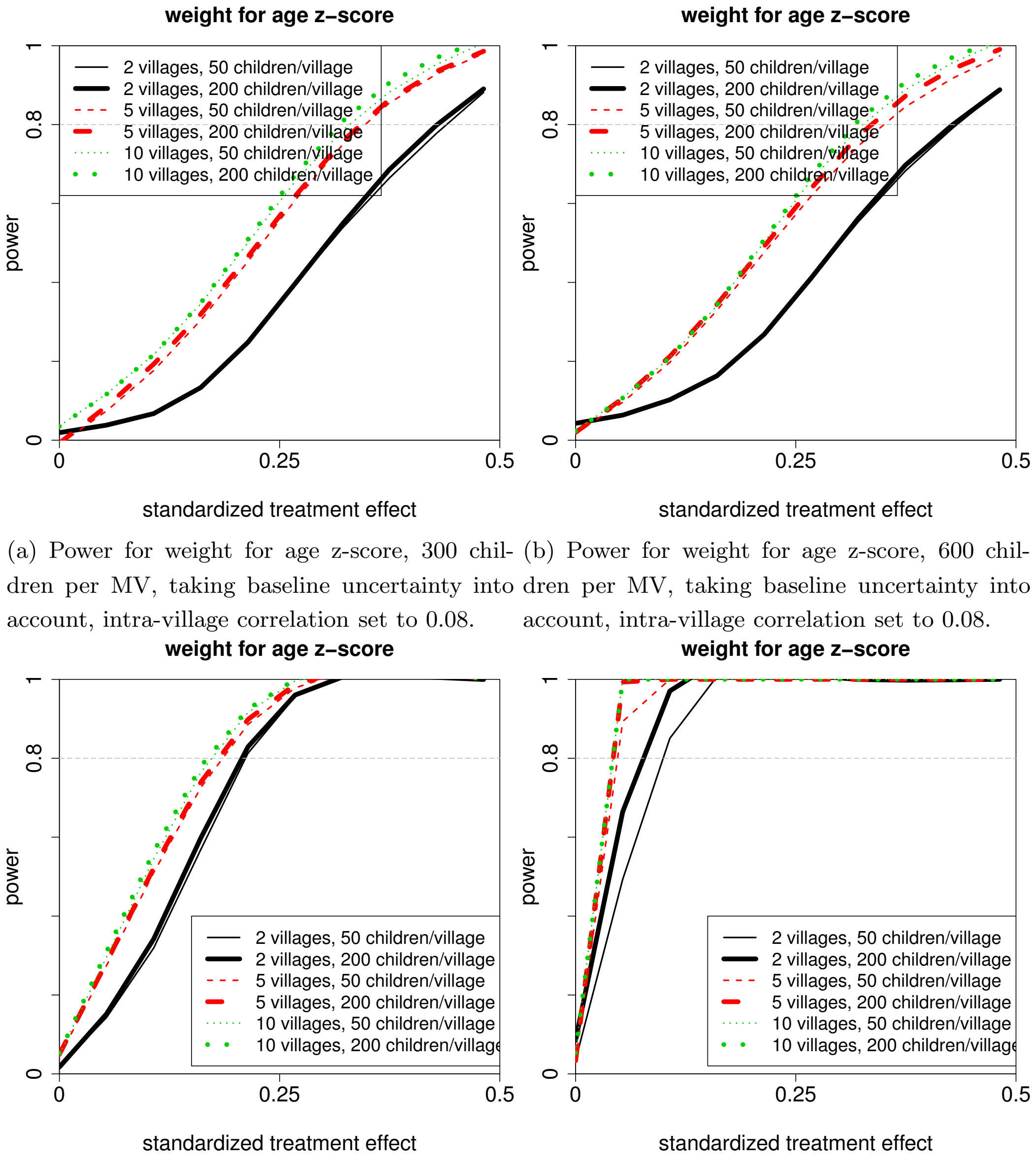
(a) Power for annualized consumption.

(b) Power for weight for age z-score.

(c) Power for measles immunization.

(d) Power for bednet usage.

Fig. 6: Power (the probability that the estimated treatment effect is statistically significant) as a function of treatment effect for four different outcomes: (a) annualized consumption, in USD (PPP 2005), (b) weight for age z-score, (c) measles immunization, (d) bednet usage; and different sample sizes: 50 or 200 children or households (hhs) per control village, 300 children or hhs per Millennium Village, and 2, 5, or 10 control villages per Millennium Village. We fit a model that assumes unconfoundedness given baseline outcomes.



(a) Power for weight for age z-score, 300 children per MV, taking baseline uncertainty into account, intra-village correlation set to 0.08.

(b) Power for weight for age z-score, 600 children per MV, taking baseline uncertainty into account, intra-village correlation set to 0.08.

(c) Power for weight for age z-score, 300 children per MV, no baseline uncertainty, intra-village correlation set to 0.08.

(d) Power for weight for age z-score, 300 children per MV, no baseline uncertainty, intra-village correlation set to zero.

Fig. 7: Power (the probability that the estimated treatment effect is statistically significant) as a function of treatment effect for weight for age z-score, with either: 300 or 600 children per Millennium Village (MV); taking baseline uncertainty or not; intra-village correlation set to 0.08 or zero.

7. Survey Data Collection: Sampling Design

Within each MV1 and control village, we will collect cross-sectional survey data using a two-stage design: households will be sampled in stage I, followed by people within households in stage II (Lohr, 2010; Särndal et al., 1992). Mitchell et al. (2015b) conduct a simulation study to determine the optimal sampling plan for our multi-module survey, choosing simple random sampling rather than probability proportional to size sampling of households.

For each of the ten MV1s, we sample 300 households out of roughly 1000 households. As described above, we randomly selected one village per matched grid cell, giving $C = 5$ distinct control villages for each MV1. Sampling in each control village is identical to sampling in the MV1 cross-sectional sample, with sample sizes divided by C (e.g. $300/C$ households sampled in each control village). Full demographic censuses are conducted in the MV1s but not the control villages, since the project has not operated in those areas. Thus, before sampling in each control village we must collect a *household list*: a list of all non-abandoned households with GPS coordinates identifying their locations. After choosing a random sample from each control village's household list, we will conduct a demographic census in those sampled households to form the sampling frame for the within-household sampling.

7.1. Post-treatment design variables

The survey sampling design variables are characteristics of the population in 2015, and are therefore post-treatment: the number of household members (of different age-sex categories) per household in 2015. To achieve ignorability of the sampling scheme, these design variables will be included in the candidate causal models proposed in Section 4. Alternatively, design-based weighting can be used, avoiding the adjustment for post-treatment variables. We propose to try both techniques and assess the sensitivity of the results.

8. Future work

Survey data collection in the treatment and control villages across the ten sites will be completed by December, 2015. After data are processed, outcome analyses will proceed as proposed in Section 4.

In parallel with data collection, we will improve on the small area models from Section 3.3 so that the results can be used in the candidate causal models. We plan to finish acquiring and processing census data, whose variables may serve as good predictors of the Demographic and Health Surveys (DHS) variables (pre-treatment measurements of the outcomes). We may fit multivariate small area models (DeSouza, 1992; Datta et al., 1998; Raghunathan et al., 2007; Li and Zaslavsky, 2010), and allow the slope parameters to vary by cluster or region. We will perform posterior predictive checks in order to iteratively adjust these small area models (Gelman et al., 2014, Chapters 6 and 7), and compare estimates to the project baseline data in countries with DHS data near project baseline (see Table **??**).

The above proposal, together with Mitchell et al. (2015a), functions to reduce the number of measurement and causal modeling choices made after outcome data are available. This work represents our pre-registration to increase credibility and transparency. Though our proposal is uniquely tailored to the Millennium Villages Project design, many aspects address features of other social policy evaluations, e.g. retrospective design, area-level interventions, and sparsity of baseline data.

Acknowledgements

The authors would like to thank the following people for very valuable feedback and ideas: Jeffrey D. Sachs, Joseph K. Blitzstein, Qixuan Chen, Jennifer Hill, Macartan Humphreys, Michael Clemens, Alberto Abadie, Marc Levy, Linda Pistolesi, Keli Liu, Peng Ding, Natalie Exner, Abhishek Chakrabortty, Rachael Meager, and Natalie Bau.

All mistakes are our own.

A. Software

For fitting multilevel models we use Stan in R, (Stan Development Team, 2013; R Development Core Team, 2014). For geographic data processing we use ArcGIS (ERSI, 2013).

B. Data in candidate comparison areas: available variables

B.1. Geographic data variables

Below we list geographic variables compiled from from geographic information system (GIS) databases and used in the matching procedure described in Section 6.4Selecting comparison villagessubsection.6.4 of the main paper.

- GADMv2 and GPWv4 - administrative boundaries (these do not correspond exactly with the census administrative names) (GPWv3; GADMv2, 2012)
- PCT_MV1, PCT_MV2, PCT_BUFF - Percent of total grid cell area in the MV1, MV2, and ten kilometer buffer zone around the MV
- Agroecological zones (Dixon et al., 2001)
  - AEZ_1 - Percent of total grid cell area in the "Irrigated" agroecological zone
  - AEZ_2 - Percent of total grid cell area in the "Tree crop" agroecological zone
  - AEZ_5 - Percent of total grid cell area in the "Highland perennial" agroecological zone
  - AEZ_6 - Percent of total grid cell area in the "Highland temperate mixed" agroecological zone
  - AEZ_7 - Percent of total grid cell area in the "Root crops" agroecological zone
  - AEZ_8 - Percent of total grid cell area in the "Cereal-root crops mixed" agroecological zone
  - AEZ_9 - Percent of total grid cell area in the "Maize mixed" agroecological zone
  - AEZ_11 - Percent of total grid cell area in the "Agro-pastoral millet/sorghum" agroecological zone
  - AEZ_12 - Percent of total grid cell area in the "Pastoral" agroecological zone
  - AEZ_13 - Percent of total grid cell area in the "Sparse (arid)" agroecological zone
  - AEZ_14 - Percent of total grid cell area in the "Coastal artisanal fishing" agroecological zone

  - AEZ_COAST - Percent of total grid cell area that is either water, coastal land, or island and is not captured by the agroecological zone data set. Calculated as $100 - \sum_j$ AEZ_j
- Access to major cities (Joint Research Centre: Land Resource Management Unit)
  - ACCESS_M - Mean (across the grid cell area) of the travel time in minutes to major cities of more than 100,000 population (1987-2004)
  - ACCESS_STD - Standard deviation (across the grid cell area) of the travel time in minutes to major cities of more than 100,000 population (1987-2004)
- Soil variables (ISRIC: World Soil Information)
  - CEC_M - Mean cation exchange capacity in cmol/kg (1950-2005)
  - CEC_STD - Standard deviation of cation exchange capacity (1950-2005)
  - CLY_M - Mean percent clay composition of the soil (1950-2005)
  - CLY_STD - Standard deviation of percent clay composition of the soil (1950-2005)
  - PH_M - Mean soil pH (1950-2005)
  - PH_STD - Standard deviation of soil pH (1950-2005)
- Vegetation index (IRI/LDEO)
  - EVI_1_M - Mean Enhanced Vegetation Index (2000-2005)
  - EVI_1_STD - Standard deviation of Enhanced Vegetation Index (2000-2005)
- Temperature (IRI/LDEO)
  - LST_D1_M - Mean MODIS Land Surface Temperature Day (2002-2005)
  - LST_D1_STD - Standard deviation of MODIS Land Surface Temperature Day (2002-2005)
  - LST_D2_M - Mean MODIS Land Surface Temperature Day (2005-2010)
  - LST_D2_STD - Standard deviation of MODIS Land Surface Temperature Day (2005-2010)
  - LST_N1_M - Mean MODIS Land Surface Temperature Night (2002-2005)
  - LST_N1_STD - Standard deviation of MODIS Land Surface Temperature Night (2002-2005)
  - LST_N2_M - Mean MODIS Land Surface Temperature Night (2005-2010)
  - LST_N2_STD - Standard deviation of MODIS Land Surface Temperature Night (2005-2010)
- Elevation (The CGIAR Consortium for Spatial Information (CGIAR-CSI))
  - ELEV_M - Mean elevation in meters (2008)
  - ELEV_STD - Standard deviation of elevation (2008)
- Population density (GPWv3)
  - POPD_M - Mean 2005 population density in persons per square kilometer. Grid cells which fall within the GPWv4 water mask have been assigned a value of zero. (1990-2000 data, projected to 2005)
  - POPD_STD - Standard deviation of 2005 population density. Grid cells which fall within the GPWv4 water mask have been assigned a value of zero. (1990-2000 data,

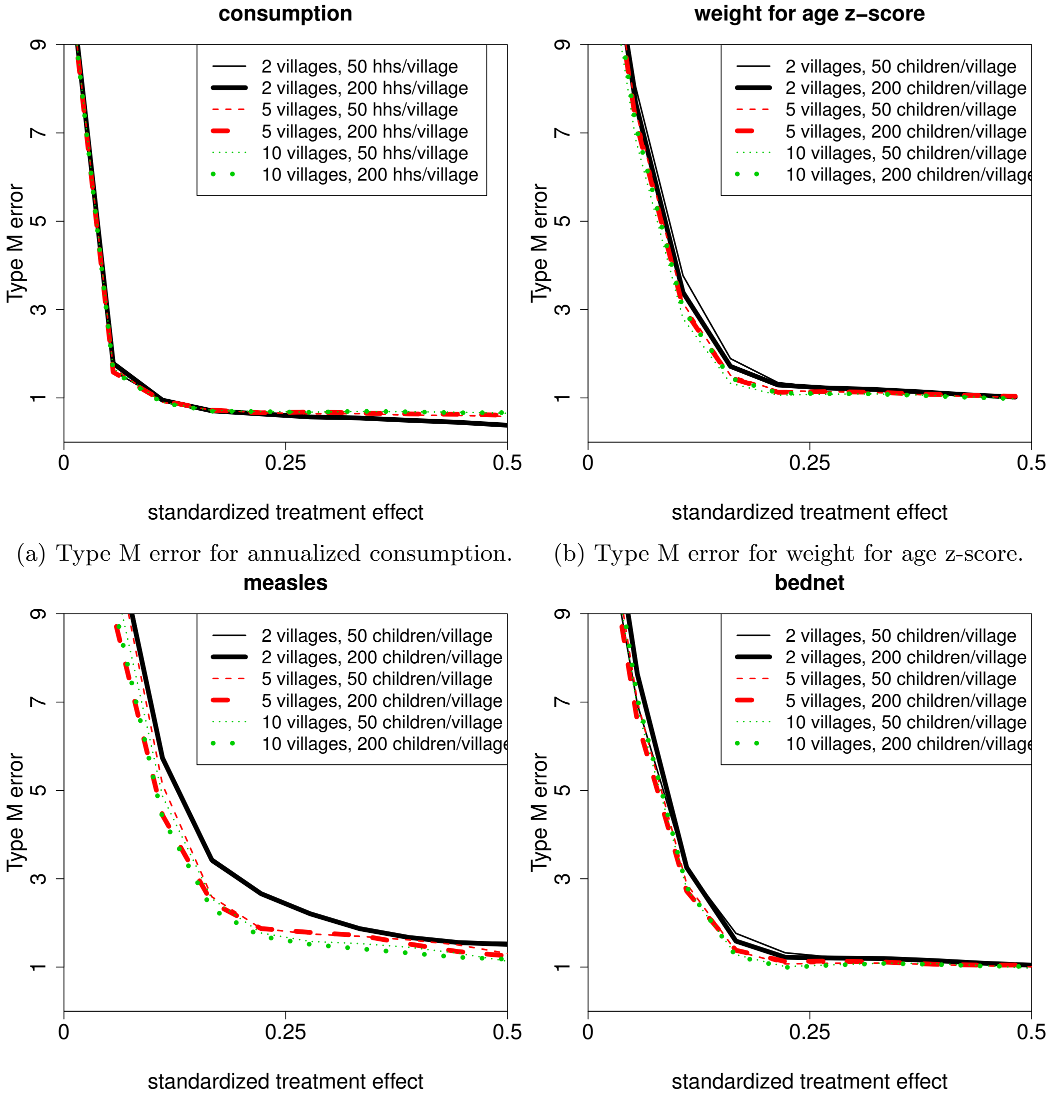
projected to 2005)

B.2. Demographic and Health Surveys variables

Below we list the variables from the Demographic and Health Surveys (DHS) that were used to construct three indices: an asset wealth index (created by the DHS), and education and health indices created via the procedure described in Section 3.2 of the main paper.

- **Asset wealth index**: the first principal component of a list of assets (source of water, type of toilet facility, materials used for housing construction, ownership of TVs, radios, bicycles, land, livestock)
- **Education index**: head of household's education level in years, children's school attendance status in the previous year
- **Health index**:
    - household-level variables:
      ownership of a bed net
      place for hand-washing observed
      soap/cleaning agent observed
    - within-household measurements, aggregated to household level:
      sought treatment for childs last diarrhea episode
      vaccinated children for measles
      women took malaria medicine during pregnancy (all births in last 5 years)
      average length for age z-score for children within a household
      proportion of children under-5 who slept under bednet the night before
      average number of ANC visits per woman per birth (in last 5 years)
      proportion of children in household with fever who get treatment
      proportion of women in household who are aware of AIDs
      proportion of women in household who use any method of contraception

C. Type M errors



(a) Type M error for annualized consumption.

(b) Type M error for weight for age z-score.

(c) Type M error for measles immunization.

(d) Type M error for bednet usage.

Fig. 8: Type M errors (the expected absolute value of the estimate divided by the true effect size, if it is statistically significant) as a function of treatment effect for four different outcomes: (a) annualized consumption, in USD (PPP 2005), (b) weight for age z-score, (c) measles immunization, (d) bednet usage; and different sample sizes: 50 or 200 children or households (hhs) per control village, 300 children or hhs per Millennium Village, and 2, 5, or 10 control villages per Millennium Village. We fit a model that assumes unconfoundedness given baseline outcomes.

References

(2014) URLhttp://www.measuredhs.com/faq.cfm.

Abadie, A. and Imbens, G. W. (2011) Bias-corrected matching estimators for average treatment effects. *Journal of Business and Economic Statistics*, **29**, 1–11.

Altonji, J. G., Elder, T. E. and Taber, C. R. (2005) Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, **113**, 151–184.

Angrist, J. D. and Pischke, J. S. (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Balk, D., Pullum, T., Storeygard, A., Greenwell, R. and Neuman, M. (2004) A spatial analysis of childhood mortality in West Africa. *Population, Space and Place*, **10**, 175–216.

Bang, H. and Robins, J. M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**, 962–972.

Carnegie, N. B., Harada, M. and Hill, J. (2015a) *treatSens: A Package to Assess Sensitivity of Causal Analyses to Unmeasured Confounding*. New York, NY. URLhttp://CRAN.R-project.org/package=treatSens. Version R package version 1.1.

Carnegie, N. B., Hill, J. L. and Harada, M. (2015b) Assessing sensitivity to unmeasured confounding using a simulated potential confounder.

Chen, Q., Elliott, M. R. and Little, R. J. A. (2010) Bayesian penalized spline model-based inference for finite population propotion in unequal probability sampling. *Survey Methodology*, **36**, 23–34.

Clemens, M. A. and Demombynes, G. (2011) When does rigorous impact evaluation make a difference? the case of the millennium villages. *Journal of Development Effectiveness*, **3**, 305–339.

Clingingsmith, D., Khwaja, A. I. and Kremer, M. (2009) Estimating the impact of the hajj: Religion and tolerance in islam's global gathering. *The Quarterly Journal of Economics*, **124**, 1133–1170. URLhttp://scholar.harvard.edu/files/kremer/files/hajj_qje_2009_august.pdf.

Cook, T. D., Shadish, W. R. and Wong, V. C. (2008) Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. *Journal of Policy Analysis and Management*, **27**, 724–750.

Cox, D. R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

Datta, G., Day, B. and Maiti, T. (1998) Multivariate Bayesian small area estimation: application to survey and satellite data. *Sankhya, Series A*, **60**, 1–19.

Datta, G., Ghosh, M., Nangia, N. and Natarajan, K. () Estimation of median income of four-person families: A bayesian approach. In *Bayesian Analysis in Statistics and Econometrics* (eds. W. Berry, K. Chaloner and J. Geweke), 129–140. New York, NY: Wiley.

Dehejia, R. H. (2005) Practical propensity score matching: a reply to Smith and Todd. *Journal of Econometrics*, **125**, 355–364.

Dehejia, R. H. and Wahba, S. (1999) Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, **94**, 1053–1062.

DeSouza, C. M. (1992) An appropriate bivariate bayesian method for analysing small frequencies. *Biometrics*, **48**, 1113–1130.

Ding, P. and VanderWeele, T. J. (2015) Sensitivity analysis without assumptions.

Dixon, J., Gulliver, A. and Gibbon, D. (2001) Farming systems and poverty: Improving farmers' livelihoods in a changing world. *Tech. rep.*, FAO and the World Bank, Rome and Washington DC.

ERSI (2013) *ArcGIS Desktop: Release 10.2*. ERSI, Redlands, CA: Environmental Systems Research Institute.

Fay, R. and Herriot, R. (1979) Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, **366**, 269–277.

Feller, A. and Gelman, A. (2014) Hierarchical models for causal effects. URL`http://www.stat.columbia.edu/ gelman/research/published/HierarchicalCausal.pdf`. Working paper.

Filmer, D. and Pritchett, L. H. (2001) Estimating wealth effects without expenditure data - or tears: An application to educational enrollments in states of india. *Demography*, **38**, 115–132.

GADMv2 (2012) Global Administrative Areas Databse (GADMv2). URL`http://www.gadm.org`.

Gage, A. J. (2007) Barriers to the utilization of maternal health care in rural Mali. *Social Science & Medicine*, **65**, 1666–1682.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–533.

— (2008) Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, **27**, 2865–2873.

Gelman, A. and Carlin, J. (2013) Beyond power calculations to a broader design analysis, prospective or retrospective, using external information. Working paper.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2014) *Bayesian Data Analysis*. Chapman & Hall/CRC texts in statistical science, third edn.

Gelman, A. and Hill, J. L. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Gelman, A., Hill, J. L. and Yajima, M. (2012) Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, **5**, 189–211.

Gelman, A. and Loken, E. (2013) The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time. URL`http://www.stat.columbia.edu/ gelman/research/unpublished/p_hacking.pdf`.

Gelman, A. and Tuerlinckx, F. (2000) Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, **15**, 373–390.

Ghosh, M. and Natarajan, K. (1999) Small area estimation: A bayesian perspective. In *Multivariate Analysis, Design of Experiments and Survey Sampling* (eds. S. Ghosh and M. Dekker), 69–92. New York, NY: Wiley.

Ghosh, M. and Rao, J. N. K. (1994) Small area estimation: An appraisal. *Statistical Science*, **9**, 55–76.

GPWv3 () Socioeconomic Data and Applications Center (SEDAC): Gridded Population of the World (GPW), v3. URL`http://sedac.ciesin.columbia.edu/data/collection/gpw-v3`.

Greenland, S., Robins, J. M. and Pearl, J. (1999) Confounding and collapsibility in causal inference. *Statistical Science*, **14**, 29–46.

Hill, J. L. and Scott, M. (2009) Comment: The essential role of pair matching. *Statistical Science*, **24**, 54–58.

Ho, D. E., Imai, K. and King, G. (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, **15**, 199–236.

Humphreys, M., de la Sierra, R. S. and van der Windt, P. (2013) Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis*, **21**, 1–20.

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001) *Bayesian survival analysis*. Springer series in statistics. 233 Spring St., New York, NY 10013, USA: Springer Science+Business Media, Inc.

Imai, K., King, G. and Stuart, E. A. (2008) Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society A*, **171**, 481–502.

Imbens, G. W. and Rubin, D. B. (2015) *Causal Inference in Statistics and Social Sciences: An Introduction*. New York, NY: Cambridge University Press.

Imbens, G. W. and Wooldridge, J. M. (2009) Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, **47**, 5–86.

International Initiative for Impact Evaluation (2013) The registry for international development impact evaluations. URL`http://ridie.3ieimpact.org/`.

IRI/LDEO () IRI/LDEO Climate Data Library. URL`http://iridl.ldeo.columbia.edu/`.

ISRIC: World Soil Information () Soil property maps of Africa at 1 km. URL`http://www.isric.org/data/soil-property-maps-africa-1-km`.

ITAD evaluation for Northern Ghana (2013) Impact Evaluation of a New Millennium Village in Northern Ghana: Initial Design Document. *Tech. rep.*, UK Department for International Development.

Jiang, J. and Lahiri, P. (2006) Mixed model prediction and small area estimation. *Test*, **15**, 1–96.

Joint Research Centre: Land Resource Management Unit () Travel time to major citis: A global map of accessibility. URL`http://bioval.jrc.ec.europa.eu/products/gam/sources.htm`.

Kreif, N., Grieve, R., Radice, R. and Sekhon, J. S. (2011) Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation. URL`http://www.lshtm.ac.uk/php/hsrp/reducing-selection-bias/output/regression_adjusted_matchi`

Paper presented at the Causal Inference Group Meeting at the Harvard School of Public Health.

Lewandowski, D., Kurowicka, D. and Joe, H. (2009) Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, **100**, 1989–2001.

Li, F. and Zaslavsky, A. M. (2010) Using a short screening scale for small-area estimation of mental illness prevalence for schools. *Journal of the American Statistical Association*, **105**, 1323–1332.

Linard, C., Gilbert, M., Snow, R. W., Noor, A. M. and Tatem, A. J. (2012) Population distribution, settlement patterns and accessibility across africa in 2010. *PLoS One*, **7**. URLhttp://journals.plos.org/plosone/article?id=10.1371/journal.pone.0031743.

Lohr, S. L. (2010) *Sampling: Design and Analysis.* Cengage Learning, 2 edn.

McCulloch, J. B. R. and Meng, X. L. (2000) Modeling covariance matrices in terms of standard deviations and correlation, with application to shrinkage. *Statistica Sinica*, **10**, 1281–1311.

McKenzie, D. (2012) Beyond baseline and follow-up: The case for more t in experiments. *Journal of Development Economics*, **99**, 210–221.

Measure DHS/ICF International (2012) Sampling and household listing manual: Demographic and health surveys methodology. *Tech. rep.*, Measure DHS. URLhttp://www.measuredhs.com/pubs/pdf/DHSM4/DHS6_Sampling_Manual_Sept2012_DHSM4.pdf.

Michelson, H., Muniz, M. and DeRosa, K. (2013) Measuring socio-economic status in the millennium villages: The role of asset index choice. *The Journal of Development Studies*, **49**, 917–935.

Mitchell, S., Gelman, A., Ross, R., Huynh, U. K., McClellan, L., Harris, M., Bari, S., Chen, J., Ohemeng-Dapaah, S., Namakula, P., Palm, S. E. S. C. and Sachs, J. D. (2015a) The Millennium Villages Project: A protocol for the final evaluation. *Submitted to The Lancet.*

Mitchell, S., Gelman, A., Ross, R., Stuart, E. A., Feller, A., Makela, S. and Zaslavsky, A. M. (2015b) Design of the Millennium Villages Project sampling plan: a simulation study for a multi-module survey. *Working Paper.*

MVP (2011) Survey enumeration manual: Guidelines for enumerators, field supervisors, and data managers. *Tech. rep.*, Millennium Villages Project, New York, NY. URLhttps://ciesin.columbia.edu/confluence/download/attachments/91488269/MVP_Y5_Enumeration_Ma

Nadram, B. (2000) Bayesian generalized linear models for inference about small areas. In *Generalized Linear Models* (eds. D. Rey, S. K. Ghosh and B. K. Mallick), 89–109. Boca Raton: CRC Press.

O'Brien, P. C. (1984) Procedures for comparing samples with multiple endpoints. *Biometrics*, **40**, 1079–1087.

Pronyk, P. M., Muniz, M., Nemser, B., Somers, M. A., McClellan, L., Palm, C. A., Huynh, U. K., Amor, Y. B., Begashaw, B., McArthur, J. W., Niang, A., Sachs, S. E., Singh, P., Teklehaimanot, A. and Sachs, J. D. (2012) The effect of an integrated multisector model for achieving the millennium development goals and improving child survival in rural sub-saharan africa: a non-randomised controlled assessment. *The Lancet*, **379**, 2179–2188.

R Development Core Team (2014) The R project for statistical computing. URL`http://www.r-project.org/`.

Raghunathan, T. E., Xie, D., Schenker, N., Parsons, V. L., Davis, W. W., Feuer, E. J. and Dodd, K. W. (2007) Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association*, **102**, 474–486.

Rao, J. N. K. (2003) *Small Area Estimation.* Hoboken, New Jersey: John Wiley and Sons.

Roberts, P., Shyam, K. C. and Rastogi, C. (2006) Rural access index: A key development indicator. *Tech. rep.*, World Bank. URL`http://www.worldbank.org/transport/transportresults/headline/rural-access/tp-10-final.pd`

Robins, J. M. and Rotnitzky, A. (2001) Comment on the Bickel and Kwon article, "On double robustness.". *Statistica Sinica*, **11**, 920–936.

Robins, J. M., Rotnitzky, A. and der Laan, M. J. V. (2000) Comment on the Murphy and Van der Vaart article, "On profile likelihood.". *Journal of the American Statistical Association*, **95**, 431–435.

Rosenbaum, P. R. (1984) The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society A*, **147**, 656–666.

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

Rubin, D. (2008) For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, **2**, 808–840.

Rubin, D. B. (1973) The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, **29**, 185–203.

— (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

— (1978) Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, **6**, 34–58.

Rubin, D. B. and Thomas, N. (2000) Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, **95**, 573–585.

Rutstein, S. O. and Rojas, G. (2006) Guide to DHS Statistics. *DHS toolkit*, Demographic and Health Surveys, Demographic and Health Surveys, Calverton, Maryland. URLhttp://dhsprogram.com/pubs/pdf/DHSG1/Guide_to_DHS_Statistics_29Oct2012_DHSG1.pdf.

Sachs, J. D. and McArthur, J. W. (2005) The millennium project: a plan for meeting the millennium development goals. *The Lancet*, **365**, 347–353.

Sanchez, P., Palm, C., Sachs, J. D., Denning, G., Flor, R., Harawa, R., Jama, B., Kiflemariam, T., Konecky, B., Kozar, R., Lelerai, E., Malik, A., Mutuo, P., Niang, A., Okoth, H., Place, F., Sachs, S. E., Said, A., Siriri, D., Teklehaimanot, A., Wang, K., Wangila, J. and Zamba, C. (2007) The african millennium villages. *Proceedings of the National Academy of Sciences*, **104**, 6775–80.

Särndal, C. E., Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Shadish, W. R., Clark, M. H. and Steiner, P. M. (2008) Can nonrandomised experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, **103**, 1334–1343.

Si, Y., Pillai, N. S. and Gelman, A. (2015) Bayesian nonparametric weighted sampling inference. *Bayesian Analysis*, **10**.

Stan Development Team (2013) Stan: A c++ library for probability and sampling, version 1.3. URLhttp://mc-stan.org/.

Steiner, P. M., Cook, T. D., Shadish, W. R. and Clark, M. H. (2010) The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, **15**, 250–267.

Stuart, E. A. (2010) Matching methods for causal inference: A review and a look forward. *Statistical Science*, **25**, 1–21.

Stuart, E. A. and Rubin, D. B. (2008) Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics*, **33**, 279–306.

The CGIAR Consortium for Spatial Information (CGIAR-CSI) () SRTM 90m Digital Elevation Data. URL`http://srtm.csi.cgiar.org/`.

UN Millennium Project (2014) URL`http://mdgs.un.org/unsd/mdg/Metadata.aspx`.

Zaslavsky, A. M. (2011) Sampling from a bayesian menu. *Statistical Science*, **26**, 235–237.

Zheng, H. and Little, R. J. A. (2003) Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. *Journal of Official Statistics*, **19**, 99–117.

— (2004) Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Survey Methodology*, **30**, 209–218.

— (2005) Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, **21**, 1–20.