

Notes on the Research Design for the NERAP Evaluation

Justin Grimmer

January 24, 2009

In this note I outline how the sample was selected for the evaluation of the National Emergency Rural Access Program (NERAP).

1 Identifying Ignorable Treatment Assignment

Our research design attempts to mitigate the biases from self-selection into treatment. Ideally we would eliminate these biases through a randomized research design—but this was impossible for the NERAP or similar road construction projects. Instead, we outline here a research design that uses information collected about the areas in which road projects will be built, a large collection of control roads (roads not affected by NERAP), and a statistical procedure to limit the influence of statistical assumptions to identify the effects of NERAP on outcomes of interest: matching (Ho et al., 2007).

Evaluations of rural road projects present a unique set of challenges to generating an unbiased estimate of the effect of interventions. Roads are designed to connect clusters of villages: lowering the cost of travel, increasing exchange, and providing access to previously remote resources (van de Walle and Cratty, 2002). This introduces two new challenges for the assumptions necessary for matching. First, the selection of road projects will depend upon the characteristics of villages *around the road* not just one village. Second, any villages

response to a road rehabilitation will depend upon the response of surrounding villages. This interaction between units violates the standard assumptions usually postulated to justify matching to identify causal effects (Rosenbaum and Rubin, 1983).

In other words, the level of assignment—clusters of villages that lie near a treated or control road—is not the same as the level of analysis—the effect of the program on villages or households. Experimental studies that exploit this difference are called *cluster randomized experiments* (Imai, King and Nall, 2008). We outline here the assumptions necessary to identify causal effects for *clustered observational studies*. Specifically, we outline two new assumptions to identify causal effects: a new definition of the Stable Unit Treatment Value Assumption and ignorability to accommodate the clustered data. We show that these new assumptions are essential: if matching is done at the village level without considering the characteristics of villages around the road then the estimate of the causal effect is biased.

1.1 Notation

Suppose there are M roads, ($j = 1, \dots, M$) and that subset of these roads, M_t are affected by NERAP interventions. Along each road we say that there are n_j villages ($i = 1, \dots, n_j$), so that there are $N = \sum_{j=1}^M n_j$ total villages. We say that village i is near the j^{th} road project if $\|i - j\| < \alpha$ where $\|\cdot\|$ is some measure of cost to travel to the project j from village i and α is some cutoff value to define precisely the cutoff for treatment reception. For our evaluation, we use ‘as the crow flies’ as our distance measure and set $\alpha = 1$. We say village i and i' are in the same cluster j if both $\|i - j\| < \alpha$ and $\|i' - j\| < \alpha$. Within each village there are $h = 1, \dots, h_{j,m}$ households, with $H = \sum_{j=1}^m \sum_{i=1}^{n_j} i_{im}$ representing the number of total households included in the analysis.

1.2 SUTVA

The first step in defining the assumptions necessary to identify causal effects is to decide upon an appropriate version of the *Stable-Unit Treatment Value Assumption* (SUTVA). Here, we show why the usual definition of SUTVA is inappropriate for analyzing the effects of NERAP and then define a version of SUTVA that accounts for the unique characteristics of road level treatments.

1.2.1 The Problem with SUTVA at Household and Village Level

We first define the standard version of SUTVA, at the level of villages, which we will use to illustrate the shortcomings of the standard definition. Let t_{ij} represent the treatment status of the i^{th} village along the j^{th} road: if a village is near the road then $t_{ij} = 1$, otherwise $t_{ij} = 0$. Collect the treatment assignment of all N villages into the $N \times 1$ treatment assignment vector \mathbf{T} .

Suppose that we are interested in assessing the effect of some village-level response, which depends upon the treatment assignment. Call the potential outcome of village i along road j , $Y_{ij}(\mathbf{T})$. Assumption 1 defines SUTVA at the village level.

Assumption 1 (SUTVA at Village Level.). *Call $Y_{ij}(\mathbf{T})$ the potential outcome for village i along road j . Suppose that there are no hidden versions of the treatment. Then if $t_{ij} = t'_{ij}$ then $Y_{ij}(\mathbf{T}) = Y_{ij}(\mathbf{T}')$*

In words, Assumption 1 says that a village's outcome depends *only upon its treatment assignment*. Another version of Assumption 1 could be postulated at the household level: the response of the household depends only upon the household's treatment assignment, not the assignment of the surrounding houses.

This assumption is inappropriate for road projects. Roads are built to connect villages and households. In other words, one of the reasons to build a road is to induce *interaction*

between the treated units. Consider the following example. Suppose village a and b are connected by a road. Represent each villages treatment status with $\mathbf{t}_{a,b} = (t_a, t_b)$ and the villages potential outcomes with $Y_a(\mathbf{t}_{a,b})$ and $Y_b(\mathbf{t}_{a,b})$. If both villages are treated, then $\mathbf{t}_{a,b} = (1, 1)$. The treatment will likely increase the exchange of commodities and labor and open new markets in the two communities. But, if only one village is treated, say $\mathbf{t}_{a,b} = (1, 0)$ then exchange between the villages may still be limited—and it will certainly be less if both villages are treated. This dependence between villages seems plausible and intuitive *and violates SUTVA*. Specifically, $Y_a(1, 1) \neq Y_a(1, 0)$ and $Y_b(1, 1) \neq Y_b(0, 1)$, contradicting Assumption 1.

There is also an external validity question when SUTVA is defined at the village level. Once again, roads will affect groups of villages. Therefore, the effect of a rehabilitation on one village is not a substantive quantity of interest, because we will never be able to perform that intervention. Rather, it is the effect of the rehabilitation on a group of villages—the intervention that actually occurs—is the quantity we are interested in assessing.

1.2.2 SUTVA at the Road Level

The version of SUTVA outlined here allows for interaction among villages and households within the same road project, but assumes that the responses of villages that lie along different projects are independent. Define the treatment status of road j with T_j : if village j is treated, then $T_j = 1$, otherwise, $T_j = 0$. Collect the treatment status of all M roads into the $M \times 1$ vector \mathbf{T} . Notice that all villages that are near the same road will have the same treatment status, emphasizing that assignment to treatment occurs at a higher level in a hierarchy than the potential levels of analysis.

This assumption extends to clustered observational studies the SUTVA made in Imai, King and Nall (2008) for clustered randomized experiments.

Assumption 2 (SUTVA). *Suppose $Y_{hij}(\mathbf{T})$ is the potential outcome for household h in*

village i in cluster j Assume that there are no hidden versions of the treatment. If $T_j = T'_j$ then $Y_{hij}(\mathbf{T}) = Y_{hij}(\mathbf{T}')$.

Assumption 2 explicitly allows interaction between treated villages. The treatment in Assumption 2 is that a *road* receives treatment—then we observe the outcome of each village along that road. This definition of treatment assignment automatically allows villages that lie along the road to interact after the road is produced, which was our objection to the standard level of SUTVA.

Threats to Validity of Road-Level SUTVA Applying the SUTVA allows us to write $Y_{hij}(\mathbf{T})$ as $Y_{hij}(T_j)$. There are two primary threats to this assumption. If control roads are too close to treated road projects, they are likely to receive some of the benefits from the road rehabilitation. To limit the possibility of spillover we eliminate all control projects with villages within 10 km of any rehabilitated project. A second threat is that a village may reside within 1 km of more than one road project. This induces a dependence between clusters that share the village as a member. To cope with this problem, we collect all clusters with a village in common to one 'super-cluster' during analysis, but treat the clusters as separate during the matching stage.¹

1.3 Ignorability

The response of villages are likely to depend on the characteristics of the other villages that are connected on the same road. Selection of projects also occurred at the road level. That is, projects were included (or excluded) from NERAP based upon the characteristics of *all* the villages that lie close to the road (Report, 2005). It is insufficient to match

¹Another threat to the SUTVA used here is that there may be different versions of the treatment. In particular, it may be the case that the treatment is more 'potent' the closer a village is to the project. At the moment, we assume away this possibility (due to the absence of a reasonable instrument for distance from a roads projects we cannot model the different levels of 'dose')

upon the characteristics of villages. Rather, matching should be based upon the aggregate characteristics of the villages that lie along roads. Here, we provide a precise definition for covariates at the road level and state a version of *ignorability* appropriate at the road-level.

1.3.1 Road Project Covariates

Suppose that each village i has a $k \times 1$ vector of covariates, \mathbf{X}_{ij} and collect all the \mathbf{X}_{ij} from road project j to form an $n_j \times k$ matrix \mathbf{X}_j . Further, we suppose that \mathbf{R}_j is a $g \times 1$ vector of road covariates—such as road length, number of villages which lie upon the road, and the altitude of the road.

We assume that there is a function $f : \mathfrak{R}^{n \times k} \times \mathfrak{R}^g \rightarrow \mathfrak{R}^b$ which summarizes the characteristics of villages along the road and the road project into a vector. We then call $f(\mathbf{X}_j, \mathbf{R}_j)$ the Road Project covariates, which play the same role as covariates in the case of matching at the individual level. For example, we may take the mean poverty level within the villages, or use the median household income.

1.3.2 Selection on Observables for Assignment at the Road Project Level

We can now use the village level covariates to formulate an ignorability assumption appropriate for the road project level data. The version of ignorability stated here is directly analogous to ignorability when assignment and analysis is at the individual level.

Assumption 3. *For each road project j suppose that the village level covariates are collected in the $n_j \times k$ matrix \mathbf{X}_j and the road characteristics are collected in the $g \times 1$ vector \mathbf{R}_j , and that the cluster covariates are given by $f(\mathbf{X}_j, \mathbf{R}_j)$ for some function f .*

3.1 (Overlap) $0 < Pr(T_j = 1 | f(\mathbf{X}_j, \mathbf{R}_j)) < 1$ for all $j = 1, \dots, M$

3.2 $(Y_{hij}(1), Y_{hij}(0)) \perp T_j | f(\mathbf{X}_j, \mathbf{R}_j)$, for all $h = 1, \dots, i_h$, for all $i = 1, \dots, n_j$, for all $j = 1, \dots, M$.

Assumption 3.1 states that at each level of the cluster’s covariates there is a positive probability of a road project receiving treatment. Assumption 3.2 states that the distribution of potential outcomes for households is independent of the treatment assignment at each level of the cluster covariates, $f(\mathbf{X}_j, \mathbf{R}_j)$.

Threats to Validity of Ignorability Assumption This could be violated in two ways. Similar to matching at the village level, if we fail to condition upon a covariate that was used to determine treatment assignment and covaries with the outcome of interest, then Assumption 2.2 is violated. Second, if we use the wrong function to summarize the covariates, then it may be the case that there is still some dependence between the treatment assignment and the potential outcomes. Notice, that we do not need to include household level of information nor information summarizing information in covariates because we know assignment was based solely upon the characteristics of villages, rather than particular households within the village. Conditioning upon household characteristics during the analysis stage, however, allows us to increase the precision of our estimates.

1.4 Unbiased Estimation of Average Treatment Effects

We are interested in identifying the average treatment effect on the treated—or the effect of the road rehabilitation on villages and households that actually were near treated roads. Formally we define this quantity as

$$\text{ATT} = \text{E}[\mathbf{Y}_{hij}(1) - \mathbf{Y}_{hij}(0)|T = 1] \tag{1.1}$$

for household level treatment effects. Of course, a naive difference in means, $\text{E}[\mathbf{Y}_{hij}(1)|T = 1] - \text{E}[\mathbf{Y}_{hij}(0)|T = 0]$ does not necessarily provide an unbiased estimate of Equation 1.1. Road projects were targeted that are likely to provide the largest gains in social welfare (de Walle,

2002) and therefore the projects selected for rehabilitation are far from a random sample. But, we can use a subset of our control population to create a comparison group that can provide an unbiased estimate of Equation 1.1.

1.4.1 Matching for Causal Effects

One approach to identify Equation 1.1, given the SUTVA and ignorability assumptions, is to identify treated roads and control roads with identical covariates. Suppose that for each treated road j we identify a control road j' such that both have identical covariates: $f(\mathbf{X}_j, \mathbf{R}_j) = f(\mathbf{X}_{j'}, \mathbf{R}_{j'})$. Then SUTVA and ignorability imply that,

$$E[\mathbf{Y}_{hij}(1) - \mathbf{Y}_{hij}(0)|T = 1] = E[\mathbf{Y}_{hij}(1)|f(\mathbf{X}_j, \mathbf{R}_j)] - E[\mathbf{Y}_{hij}(0)|f(\mathbf{X}_{j'}, \mathbf{R}_{j'})].$$

which is often called *exact matching*. But exact matching is only a feasible strategy when there are a few, discrete covariates and a massive pool of potential control units (Rubin, 1980). In real-life problems (like the NERAP evaluation) we have several , potentially continuous covariates. In these cases the *curse of dimensionality* prevents us from exploiting exact matching.

Propensity-Score Matching If exact-matching is impossible, there are a number of different matching methods can be used to identify a causal effect: *genetic-matching* (Diamond and Sekhon, 2008), *synthetic matching* (Hainmueller, 2009), and *coarsened-exact matching* (Iacus, King and Porro, 2009), all of which could be applied to match at the road-level to obtain balance between the treatment and control groups. We focus on propensity score matching here, primarily because we obtain balance between the treatment and control groups using this method (Rosenbaum and Rubin, 1983).

Given Assumption 2 and Assumption 3, we can define the probability (propensity) that each road project is treated, *the propensity score*, as $e(f(\mathbf{X}_j, \mathbf{R}_j)) = \Pr(T_j = 1|f(\mathbf{X}_j, \mathbf{R}_j))$.

This leads to the most important theorems from Rosenbaum and Rubin (1983), which follow directly from our modified assumptions.

Theorem 1. *If treatment assignment is strongly ignorable given $f(\mathbf{X}_j, \mathbf{R}_j)$ then it is strongly ignorable given $e(f(\mathbf{X}_j, \mathbf{R}_j))$. Further, conditioning upon $e(f(\mathbf{X}_j, \mathbf{R}_j))$ provides an unbiased causal estimate of the ATT (Equation 1.1 at $e(f(\mathbf{X}_j, \mathbf{R}_j))$)*

Proof. Follows directly from Rosenbaum and Rubin (1983). □

The key to Theorem 1 is that we have a set of covariates at the cluster level are sufficient to identify Equation 1.1 therefore the same intuition from Rosenbaum and Rubin (1983) applies.

When examining road projects, a natural temptation might be to assume villages were selected for treatment based upon their individual characteristics and therefore match villages based upon their individual characteristics. Formally, we would condition upon each village's covariates \mathbf{X}_{ij} , without considering the characteristics of surrounding villages or the surrounding roads. In the following remark we show that under general conditions exact matching on individual covariates is insufficient to identify Equation 1.1.

Remark 1. *Suppose Assumptions 1 and 2 hold, but we use the village level covariates \mathbf{X}_{ij} . Suppose treatment assignment is at least in part based upon road characteristics \mathbf{R}_j or the characteristics of surrounding villages in the cluster. Then conditioning upon \mathbf{X}_{ij} is not sufficient to provide an unbiased estimate of Equation 1.1*

Proof. Suppose treatment assignment depends solely upon one road characteristic $f(\mathbf{X}_j, \mathbf{R}_j) = r_j$. Then,

$$E[\mathbf{Y}(1)|T = 1, \mathbf{X}_i] - E[\mathbf{Y}(0)|T = 0, \mathbf{X}_i] \neq E[\mathbf{Y}(1) - \mathbf{Y}(0)|\mathbf{X}_i]. \quad (1.2)$$

The case of surrounding villages follows similarly. □

Remark 1 shows that when assignment occurs at the road-project level, matching at the individual level can result in biased estimates of causal effects, even when exact matching is performed on the village level covariates

2 Identifying an Appropriate Population of Control Roads

In this section we detail the data used to create the matched samples. We describe the variables used in the matching procedure and the extent to which we were able to obtain balance along those control variables.

2.1 Data

2.1.1 Control Variables

To aid in matching, we have created a large database involving characteristics of Afghanistan's villages and the road projects: including data about the characteristics of roads, access to resources and demographics within villages, geographic characteristics of the area near the proposed projects, and the number of disbursements from other aid sources made prior to the implementation of the NERAP project. We will use this data to do nearest neighbor propensity score matching, as detailed below.

2.1.2 Road Projects

To remove bias and increase efficiency in the impact evaluation of NERAP, we generate a matched sample based upon a rich set of pre-treatment covariates. We focus upon road projects with at least one village within 1 km. This results in a set of 159 treated projects. Our control projects are selected from a Ministry of Rural Rehabilitation and Development

list which contains a set of road projects throughout the country. We then removed all roads that were previously improved or were within 10 km of any NERAP project. This leaves 1,925 control roads.

2.2 Matching

To select road projects—and subsequently villages that would be available for sampling—we used nearest neighbor propensity score matching as implemented in the `MatchIt` package for R (Ho et al., 2007). Table 1 enumerates the covariates, along with providing a summary statistic used to generate the ‘road’ level covariates. In addition to the variables listed in the table below, we also included squared, logged, and interaction terms to help ensure that we obtained the best balance possible on all moments of the data.

2.3 Balance Improvement

Using the covariates in Table 1, we used nearest neighbor propensity score matching to generate a matched sample. To measure the success of our matches, Figure 1 displays the standardized bias before and after matching. The standardized bias in the covariates—the average difference between treatment and control covariates scaled by the standard deviation of this difference—is a metric commonly used to assess the balance found in the data (Rubin, 2001). The black dots in the figure represent the pre-matching standardized bias, while the red dots represent the post-matching difference. Our goal is to move the red dots as close to zero as possible.

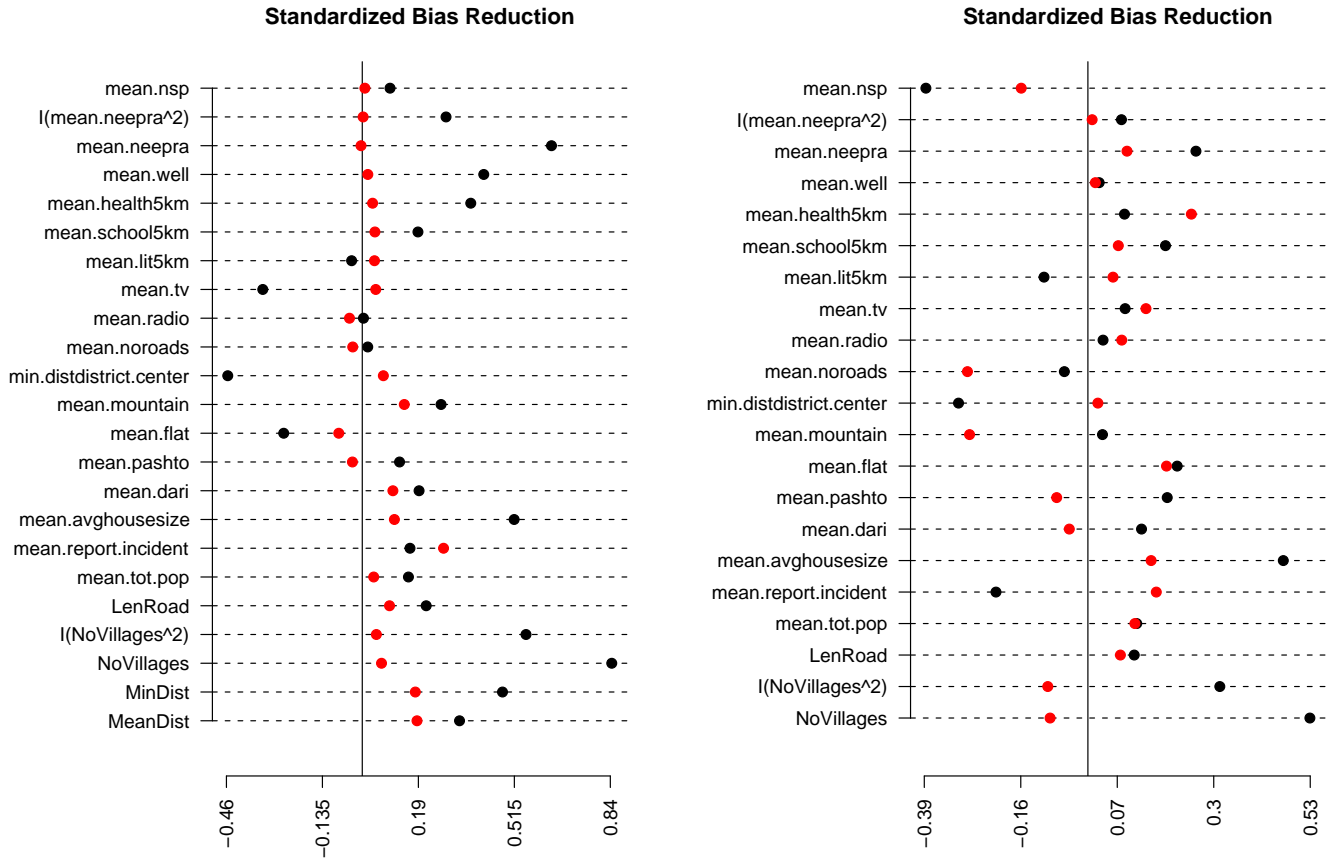
Figure 1 displays the substantial differences between treatment and control roads before matching. For example, treated roads were much more likely to have potato farming in villages along the projects and tended to be much longer than the control projects. Matching reduces the bias substantially—which is represented by the close proximity of the red dots

Table 1: Covariates Used in Matching

Covariate	Summary Function
No Villages Along Road	*
Mean Density Villages Along Road	*
Length of Project	*
Closest Hospital	*
No. Hospitals < 5 km	*
Total Population w/in 1 km of road	Sum of Villages' populations
Distance to district center	Closest Village on Project
Household size	Average over Villages
Dari	Proportion of Villages Speaking Dari
Pashto	Proportion of Villages Speaking Pashto
Flat	Proportion of Villages Near 'Flat' Terrain
Mountain	Proportion of Villages Near 'Mountainous' Terrain
Cars All Season	Proportion of Villages with Car Access All Season
No Roads	Proportion of Villages with No Roads
Literacy < 5km	Proportion of Villages with Lit Center W/in 5 km
Health < 5 km	Proportion of Villages with Health Center W/in 5 km
School < 5 km	Proportion of Villages with School W/in 5 km
Radio < 5 km	Proportion of Villages with radio access
TV < 5 km	Proportion of Villages with TV access
Altitude	Standard Deviation of Villages' Altitude
River	Proportion of Villages with River Village Access
Potato	Proportion of Villages Growing Potatoes
Rice	Proportion of Villages Growing Rice
Industrial	Proportion of Villages with other (non-farming)Industry
NSP	Mean No. NSP Projects Along Road
NEEPRA	Mean No. NEEPRA Projects Along Road

to the vertical line at zero. We suspect that matching is so successful here because the use of summary statistics, in particular the average of village characteristics, makes our data closer to normally distributed. It is well known that normally distributed data fall into an important class of covariates where matching is able to perform best (Rubin and Thomas, 1992).

Figure 1: Assessing Balance After Matching



This figure presents the standardized bias before and after matching. The standardized bias is the difference in means of the treatment and control covariates, scaled by their means. The black dots are the biases before matching, the red dots the biases after matching. Substantial differences existed before matching, but matching appears to ameliorate these differences.

3 Sampling

Using the identified control populations, we need to sample from the villages that lie along each road to obtain the outcome data that we will use to estimate treatment effects (as well as additional control data that we will use to increase efficiency). In this section, we outline how the samples of villages from the treated and control roads were collected.

3.1 Power-Calculations

Power-calculations are necessarily an approximate effort that is more guesswork than analysis, more art-form than science. In this section I describe some of the power-calculation work that we performed. While we set out to do a formal power calculation, we lacked the appropriate data to estimate the multitude of parameters necessary to do a power calculation for a road-evaluation. Therefore, we derived some parameters to describe differences in means, then used those parameters to generate approximate guidelines for the sampling.

3.2 Standard Error for Difference in Means

Suppose that we are interested in assessing the difference in means for some outcome between the treated \bar{y}_t and the control units \bar{y}_c . Suppose that $\sigma_{t,road}^2$ describes the attention in the dependent variable among the treated units across the roads, $\sigma_{t,village}^2$ is the variation among treated villages, $\sigma_{t,house}^2$ is the variation across treated households and that $\sigma_{c,road}^2$, $\sigma_{c,village}^2$, and $\sigma_{c,house}^2$ are the analogous quantities for the control units. Define the number of treated roads as M_t , treated villages as N_t and treated households as H_t , with M_c , N_c , and H_c as the analogous quantities for the control units. Then, the standard error for $\bar{y}_t - \bar{y}_c$ is

$$se(\bar{y}_t - \bar{y}_c) = \sqrt{\frac{\sigma_{t,road}^2}{M_t} + \frac{\sigma_{t,village}^2}{N_t} + \frac{\sigma_{t,house}^2}{H_t} + \frac{\sigma_{c,road}^2}{M_c} + \frac{\sigma_{c,village}^2}{N_c} + \frac{\sigma_{c,house}^2}{H_c}} \quad (3.1)$$

As Equation 3.1 shows, we need reasonable estimates for 6 different variance parameters to make accurate power calculations. Given the difficulty of this, we can look at Equation 3.1 to obtain intuition about where we should invest more resources.

3.3 A Heuristic For Sampling

We hypothesize that the variation between villages and households and villages along villages will be substantially smaller than the variance across road projects. Further, the variance among villages on a road project will likely increase as the length of the road increases. Therefore, we sampled one village for every 2.5 KM of roads. These villages were a simple random sample of the villages along the road.

References

- de Walle, Dominique Van. 2002. "Choosing Rural Road Investment to Help Reduce Poverty." *World Development* 30(4):575–589.
- Diamond, Alexis and Jasjeet Sekhon. 2008. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." University of California, Berkeley Mimeo.
- Hainmueller, Jens. 2009. "Synthetic Matching for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." Harvard University Mimeo.
- Ho, Daniel et al. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3):199–236.
- Iacus, Stefano, Gary King and Giuseppe Porro. 2009. "Matching for Causal Inference Without Balance Checking." Harvard University Mimeo.
- Imai, Kosuke, Gary King and Clayton Nall. 2008. "The Essential Role of Pair Matching in Cluster-Randomized Experiments, With Application to the Mexican Universal Health Insurance Evaluation."

- Report, Commission. 2005. National Emergency Employment Programme: Towards a National Rural Access Strategy. Technical report NEEP Joint Program Management Unit.
- Rosenbaum, Paul and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41–55.
- Rubin, D. 1980. "Bias Reduction in Mahalanobis-Metric Matching." *Biometrics* 36(2):293–298.
- Rubin, D. 2001. "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation." *Health Services and Outcomes Research Methodology* 2(1):169–188.
- Rubin, Donald and Neal Thomas. 1992. "Affinely Invariant Matching Methods with Ellipsoidal Distributions." *The Annals of Statistics* 20(2):1079–1093.
- van de Walle, Dominique and Dorothyjean Cratty. 2002. "Impact Evaluation of a Rural Road Rehabilitation Project."