Targeting social audits to improve service delivery: A tool for policymakers in India

Topic revision since proposal:
In my proposal, my project title was "Predicting corruption hot-spots in social welfare programs: A tool for policymakers in India." I am still using data from the same program (MGNREGS) in the same setting (northern India) with roughly the same objective (create a tool for policymakers that helps them identify where they should conduct social audits of the program). However, rather than using "ghost workers" as a local estimate of corruption in the program, I am using surveys of program beneficiaries to construct local measurements of the quality of program management. The reason for this shift is data availability: the data that I received on ghost workers was very messy and would have required a lot more work in order to put it into a usable format, with no guarantee of interesting findings (i.e. there may be no variation in the estimated number of ghost workers across villages, or no evidence of ghost workers at all). Instead, I am using different data that I received from the same researchers to explore perceptions of program management from a random sample of 10,000 beneficiaries across 250 villages in three districts of Rajasthan, India. I was able to clean this data, incorporate it into a prediction model, and create some preliminary visualizations in time for the Milestone 1 deadline.

Overview and Motivation:
The Mahatma Gandhi National Rural Employment Guarantee Scheme (MGNREGS) is the largest pay-for-work program in the world: last year, it provided 2.3 billion person-days of work to 50 million households in rural India (MGNREGS website, FY 2012-2013, URL). Under MGNREGS, any rural household that requests work is legally guaranteed a job card and 100 days per year of unskilled manual labor at a minimum wage determined by the state. MGNREGS requires huge public expenditures, and spending on MGNREGS projects constitutes over 75% of the budgets of local governments (MGNREGS website, Fund Flow Statement for FY 2012-2013, URL).

MGNREGS implementation is decentralized: local governments are responsible for identifying program needs, lobbying for funds for projects, and managing all aspects of the program in the village. As a result, the quality of local management of MGNREGS varies across villages. The line department that is responsible for the program, Ministry of Rural Development, has an ongoing interest in determining the extent to which quality varies across villages, and to improve implementation in the lowest-performing villages.

My final project uses survey data from a random sample of MGNREGS beneficiaries in three districts of the state of Rajasthan to estimate and predict the quality of program management at the village-level. The primary objective of the project is to develop a tool that policymakers can use to identify priority areas for social audits. In order to communicate the results of the model to policymakers effectively, I have created a visualization that shows area-specific estimates and predictions of performance, with further details about specific villages on demand. This project will build on prior research that I have conducted on the impact of specific aspects of MGNREGS, and on the interaction of MGNREGS with local political institutions in rural India.

Related Work:
The motivation for my final project comes from previous studies that estimate quality of public service delivery in local village governments in India, including:

Besley, Timothy, Rohini Pande and Vijayendra Rao (2012). Just Rewards? Local Politics and Public Resource Allocation in South India. World Bank Economic Review 26(2): 191-216.

Niehaus, Paul and Sandip Sukhtankar (2012),"Corruption Dynamics: The Golden Goose Effect", Working Paper, Dartmouth College.

Niehaus, Paul and Sandip Sukhtankar (2012),"The Marginal Rate of Corruption in Public Programs: Evidence from India", Working Paper, Dartmouth College.

My final project is a natural extension to these studies in that it uses estimates of public service delivery to visualize spatial variation in public service delivery across local village governments. The objective of my final project differs from these studies, in that the final product is not academic (i.e. a theoretical model supported by empirical analysis), but rather policy-focused (i.e. a tool that policymakers can use to target social audits).

Questions:

- Where is MGNREGS being managed well, and where is MGNREGS being managed poorly? This visualization enables policymakers to quickly identify the villages, and clusters of villages, where the program is poorly, and to direct auditing resources appropriately.

- How precise are these predictions? It may be that some or all of the MGNREGS indicators have limited ability to predict the quality of local management. Communicating error bounds around performance predictions is an important part of the visualization.

- What specific indicators contribute to each prediction? Providing information that will allow policymakers to dissect predictions for specific areas will be an important part of the visualization. For instance, one village may be assigned a low performance score by the model since a large fraction of beneficiaries report paying bribes for program benefits, while another village may be assigned a low performance score by the model since the average wait time for program benefits is very long. This details-on-demand feature of the visualization enables policymakers to focus on specifics aspects of the program in each village, which may increase the effectiveness of audits.

Data:

Geographic data
Geographic data comes from the Census of India (2001), which I obtained from the online Harvard Geospatial Library. I relied on two datasets:

- A village-level dataset that includes geographic centroids of villages (points), with the names of the block and district that each village is found in.
- District-level shapefiles

In order to generate village-level polygons, I imported the village-level dataset into ArcGIS and constructed "nearness polygons" (also "Thiessen polygons") around each village centroid. I clipped this map using the boundaries of the three study districts from the district-level shapefiles. Block boundaries were obtained from aggregating ("dissolving") the village polygons by block. Hence, village and block boundaries are not exact but depend on the Thiessen approximations, whereas district boundaries are exact.

I exported the village, block and district maps from ArcGIS as shapefiles, then converted them to GeoJSON files using this website.

Surveys of program beneficiaries
Survey data was generously provided to me by researchers at the Abdul Latif Jameel Poverty Action Lab at MIT (J-PAL). The survey was conducted in February 2010 in 250 villages that are representative of all 750 villages in the three study districts in Rajasthan. The sample consisted of approximately 40 MGNREGS beneficiaries per village, for a total of 10,088 beneficiaries.

I cleaned the survey data in STATA, which included constructing indicators of performance based on survey responses, aggregating indicators to the village level, and adding geographic identifiers so that the data could be merged with the GeoJSON files. My .do files are available upon request.

MGNREGS administrative data
MGNREGS administrative data at the village-level is available on the program's website. I used Python to scrape data for the study area from block-level tables like this one. However, researchers at the Harvard Kennedy School have recently developed a data portal that facilitates downloading data from the MGNREGS website.

I cleaned the administrative data in STATA, which included merging datasets by village (different indicators were provided in different tables), filtering the data by year, and adding geographic identifiers so that the data could be merged with the GeoJSON files. My .do files are available upon request.

Other data
Village-level data from the most recent Census of India (2011) was used in the prediction model. This data is available here.

I considered using another J-PAL dataset to construct performance indicators, which consisted of reports on visual inspections of MGNREGS worksites. However, I decided not to use this dataset since it included fewer indicators that were clearly related to the quality of local management of the program.

Exploratory Data Analysis:

Behind the performance scores: Who are program beneficiaries?
(TO ADD: Summary statistics of demographics, representativeness of population)

Components of performance scores
The village-level performance scores were constructed from aggregating responses to the following survey questions:

1. Performance indicator: Unmet demand.
   Calculated from the following survey questions:

1.D.1	Did you work for NREGA in the last 12 months?	1. Yes **(Go 1.D.6)** 2. No
1.D.2	Did you want to do some NREGA work in the past 12 months?	Yes No **>> Go to next section**

2. Performance indicator: Low wages (less than Rs 100 per day).
   Calculated from the following survey question:

1.D.3	How much money were you getting per day?	_____ Rupees

3. Performance indicator: MGNREGS ranking
   Calculated from the following survey question:

1.E.1	How would you rate NREGA implementation in your village?	1. Very bad 2. Bad 3. Neither bad nor good 4. Good 5. Very Good

4. Performance indicator: Pay a bribe for job card.
   Calculated from the following survey question:

1.E.2	What do you need to do to get a job card?	1. Get registered at the Panchayat office 2. Give a family picture 3. Give money 98. Any other answer 99. Don't know

5. Performance indicator: Wait for job.

Calculated from the following survey question:

1.E.3	How long should it take to get a job after you apply?	1. ____days 2. ____months 99. Don't know

(TO ADD: Distributions of indicator variables across villages)
(TO ADD: Distribution of performance score across villages)

Predicting performance scores in non-surveyed villages
(TO ADD: description of process: stepwise regression to construct a linear prediction model)
(TO ADD: specifications of linear prediction model)
(TO ADD: results of model: significance of individual coefficients, R^2, etc)

Design Evolution:

Implementation:

Evaluation: