Pakistani Candidate Asset Declaration Data - Progress and Update

Luke Sonnet

11/28/2018

This report outlines the goal of collecting data on Pakistani candidates, current progress on data entry and cleaning, and what can be expected in the final dataset in terms of the candidates included and the covariates available.

Introduction

In the run up to the 2018 General Elections in Pakistan, all prospective candidates must submit several forms to be approved as candidates in the general elections. Two of these forms, a candidate affidavit form and a statement of assets and liabilities were released by the Election Commission of Pakistan (ECP) in late June, 2018, just under a month before the elections.

These forms contain a rich set of data on self-declared candidate wealth, tax payments, foreign holdings, outstanding criminal charges, education, occupation, payments to and from political parties, and more. Every candidate is supposed to submit these documents. The central problem with these data is that they need not be released by the ECP, and thus there are gaps in which candidates/constituencies are covered, and that the data released are of questionable quality (e.g. some uploads are photographs of hand-written PDFs).

USIP, in a continued effort to release high-quality data around elections in Central and South Asia, has funded an effort to enter this data rigorously, and clean and prepare it so that it can be merged with other important datasets, such as the election results and the official tax payments reported by the tax authority of Pakistan.

Current status

The first wave of manual data entry by Research Solutions, a contracted firm based in Lahore, Pakistan, with experience working with survey and administrative economic and electoral data, was completed in September of 2018. In the ensuing months, I have gone back and forth with Research Solutions, asking them to reenter certain data, and to address certain forms that may not have been entered. This process is ongoing, but is approaching the final iteration.

In the meantime, I have begun cleaning and organizing the data, preparing to merge it with the election results data and make it amenable for analysis in a report and for future academic research.

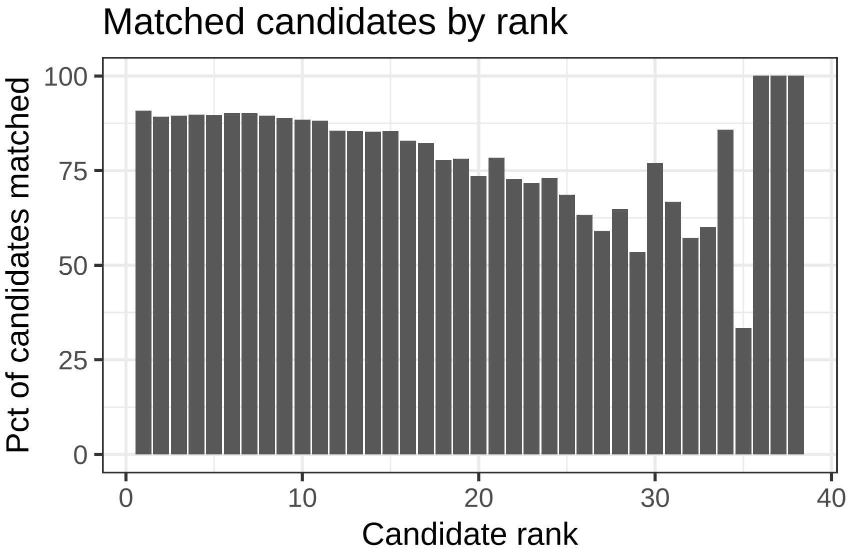
Data coverage

Data will not be available for all candidates, even in the best case scenario. Some of the PDFs that were released by the ECP are corrupt, some are unreadable, and some files are missing altogether, as are some constituencies.

In total, 17927 forms have been entered. There are only 11689 candidates that contested the 2018 General Elections in open seats, implying that many more forms were filed than candidates ended up contesting. This could be due to candidates being disqualified or dropping out.

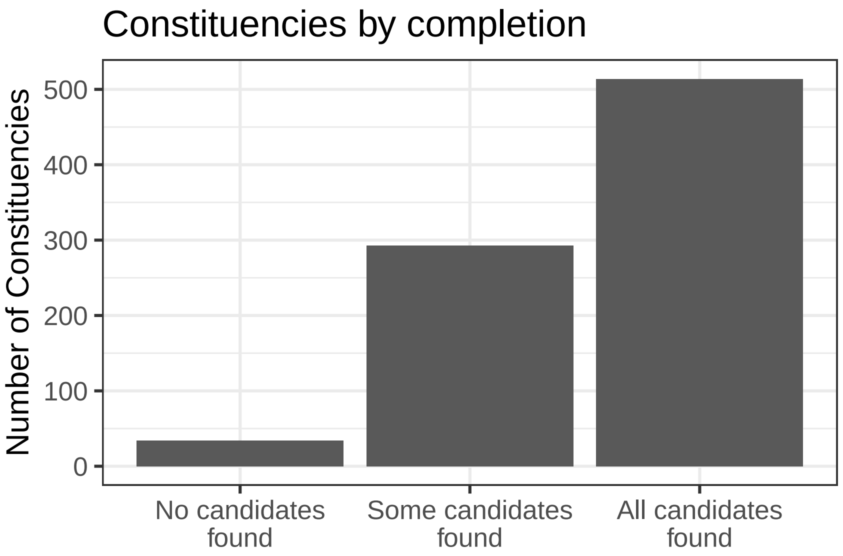
Of these 11689 candidates who contested the election, we have successfully matched 10214, or 87.4 percent of them. This number may go up as further efforts to plug gaps are being made. In particular, Research Consultants is currently investigating 624 forms that were released after the initial scrape of the ECP website that may shore up some of these gaps.

Let's look at how this missingness breaks down by candidate rank, constituency, and province. First, what percent of first place candidates, second place candidates, and so on have we matched? The next figure and table break it down, and shows that we generally recover higher ranked candidates at higher rates (although it should be said that after the fifth ranked candidate or so the number of total candidates begins to fall as well). Note in the table the one "NA" candidate ran unopposed.

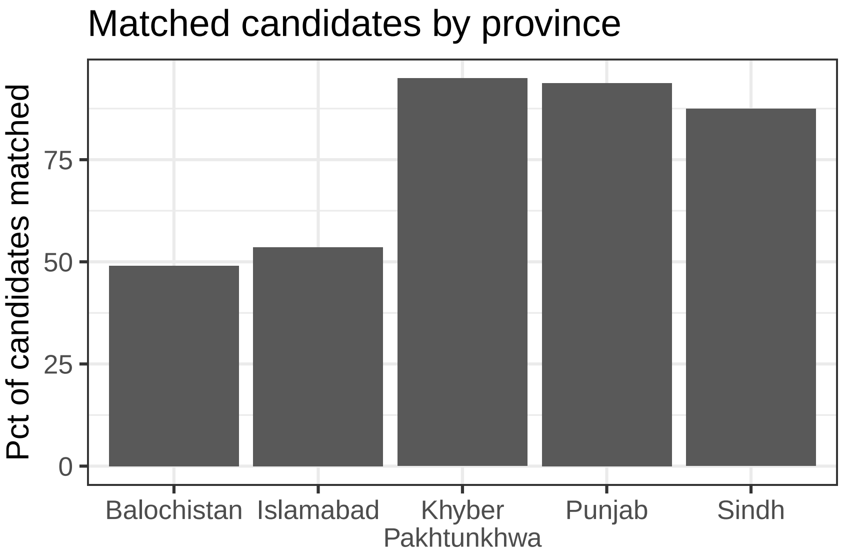


Candidate rank	Matched candidates	Total candidates	Percent matched
First	763	840	90.8
Second	749	840	89.2
Third	752	840	89.5
Fourth	754	840	89.8
Fifth	749	836	89.6
Sixth+	6447	7492	86.1
NA	0	1	0.0

The next figure shows the number of constituencies (out of 840 that contested this election), for which we have zero data, complete data, or partial data coverage.



Finally, we can see that much of the missingness comes from Balochistan and Islamabad (where again the total number of candidates is low as well). In Balochistan, some entire provinces are missing.



Province	Matched candidates	Total candidates	Percent matched
Balochistan	605	1233	49.1
Islamabad	37	69	53.6
Khyber Paktunkhwa	1771	1866	94.9
Punjab	5168	5510	93.8
Sindh	2633	3011	87.4

Lastly, there are both provincial and national assemblies. The national assemblies are much higher profile, and we can see our coverage there is also better:

Assembly	Matched candidates	Total candidates	Percent matched
National	3036	3431	88.5
Provincial	7178	8258	86.9

Available covariates

After all of the merging is done, a considerable amount of cleaning and estimating from the data will need to be done to create high quality data. However, the covariates included will be at least these below (although there may be missingness on some variables for some observations):

- Identifiers
  - National tax number
  - Computerized National Identity Card number
- Demographics
  - Education
  - Occupation
  - Number of children
  - Location registered to vote
  - Number and status of outstanding criminal cases
- Assets and liabilities
  - Total asset value this year
  - Total asset value last year
  - Total tax payments last three years
  - Total liabilities
  - Value of:
    - Domestic property
    - Foreign property
    - Capital holdings
    - Vehicles
    - Jewelry
    - Cash deposits
    - Other investments (by category)
- Other data
  - Foreign passports
  - Foreign trips number and cost
  - Elected previously