Pakistan Project Report

Using a fuzzy set, I originally tried to match the names on the shape files to the survey data. However, given the variability of the names and the significant difference between “chah 10/5h” and “chah 90/5h,” even though the Levenshtein distance is only off by one, using a fuzzy set was not very effective. Rob manually went through the survey data to verify the Tehsil and District of each and every data entry.

Next, after removing the perfect matches between the survey data and the shape file, I entered the address from the shape file to the Google API Geocoder. This returned the address it thought I was asking for and the corresponding latitude and longitude coordinates. I then cross-matched the coordinates to that of the polygons in the shape files, where I then pulled the union council, tehsil, and district from the polygon. I then checked to see if the union council in the survey data was somewhere in the returned address from the Google API Geocoder. Next, I checked to see if the tehsil and district from the survey data matched the tehsil and district from the shape files we pulled. And finally, for each entry in the survey data, I checked to see if the tehsil and district from the shape files we pulled matched the tehsil and district that we expected, since Rob manually matched the tehsil and district for each entry in the survey data. For each match, I did a direct comparison after removing spaces and ignoring case. In light of this, I also did a fuzzy match for every direct match, after removing spaces and ignoring case.