Coordinates Assessment

Tyler Brown

2017-11-28

We need to verify that imputed coordinates are making sense. We start with the "combined" data which includes mostly unique `pid-year` combinations.

Description of Datasets and Imputations

Table 1: 'Combined' dataset dimensions

row_count	col_count
672315	81

Given these dimensions, the `Latitude` and `Longitude` columns are both filled out about `65%` of the time.

Table 2: Coordinate Pair Completion Rate

complete	incomplete	percent_incomplete	total
434048	238267	35.44	672315

Table 3: Count of unique PID with Incomplete Coordinate Pairs

incomplete	PID
238267	168343

The second file, `fixed_loc.csv`, includes imputed values for the missing coordinate pairs. We imputed coordinates from openaddresses.io for the missing data. The number of imputed values do not account for all of the incomplete cases.

Table 4: Percent Imputed by `Fixed Locations`

table3_PID	fixed_loc_row_count	percent_imputed
168343	110395	65.58

Quality of Imputed Coordinates

Let's take a closer look at `fixed_loc.csv` to see if coordinate pairs were successfully imputed.

Table 5: Check: Is each row a unique PID?

row_count	unique_PID	percent_unique_pid
110395	57803	52.36

The number of unique PID values is almost double. This is most likely caused by a one-to-many join when a one-to-one join was intended.

Table 6: Top 10 Multiple PID-Coordinate Combinations

n	unique_row_count
1	62336
2	2263
3	1261
5	667
6	387
9	283
7	255
11	197
4	158
10	111

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    20.0    62.0   200.4   118.0  1584.0
```

The fixed_loc.csv file clearly needs some work. This work will need to be done in a later version of the application. Let's try to put together a minimal version of fixed_loc.csv that may be usable for an initial imputation.

We have multiple coordinate pairs matched to PID values. I'm going to group by PID and take the first match.

Table 7: Comparing First Group Row Count to Table 5

first_match_rows	table5_row_count
57803	110395

Now that we have a file without duplicate PID values, let's see how unique imputed coordinates are compared to the Combined file.

```
## [1] "Quantiles of an Aggregate on the `first_match` table"
```

```
##     0%    25%    50%    75%   100%
##    1.0    1.0    2.0    8.5 2813.0
```

We can clearly see here that some PID values will share a coordinate pair. This is to be expected on some level because we ignored the unit number when matching coordinates. It wouldn't make sense for the unit number within a property to have a meaningfully different coordinate value.

The last thing to do for this version is to go through and see if a coordinate pair with multiple PID values is substantively valid.

```
## Warning: Grouping rowwise data frame strips rowwise nature
```

Table 8: Number of unique addresses imputed.

unique_pid_addr_in_combined	imputed_coord_count
172292	58855

Now we need to see if the number of coordinates match up with the number of units for each address. A PID

will correspond to each unit within a property.

Table 9: Granularity of Coordinates

st_address_cpair_rows	st_address_rows	cpair_rows	st_addr_cpair_to_cpair_ratio
11402	11392	8670	1.31511

In the above table, it's clear to see that the granularity of imputed coordinates is an order of magnitude less than the granularity provided by each street address.

Comparing Non-imputed Granularity of Coordinates

Let's see how granular non-imputed coordinates are for the city of Boston.

Table 10: Granularity of Non-Imputed Coordinates for Combined

st_address_cpair_rows	st_address_rows	cpair_rows	st_addr_cpair_to_cpair_ratio
294025	98318	99144	2.965636

It appears that imputed coordinates are (`6.5:1`)considerably less granular than the given coordinates (`4:1`).

Summary

- The initial `fixed_loc` file had some issues with multiple coordinates being matched to a `PID` in several cases.
- As a first version fix, I grouped on each `PID` within `fixed_loc` and kept the first coordinate pair.
- The ratio of `PID` address and coordinate pair over `PID` coordinate pair is similar for the imputed and non-imputed coordinates.

To Do

- Given the `6.5:1` and `4:1` ratio of street addresses to unique coordinate pairs for both imputed and non-imputed values, we will probably need to reconsider our mapping strategy.