Geolocation and Interpolation of Polling Places Using GIS Census Data

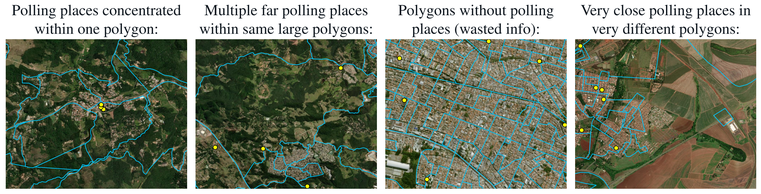
Fabricio Vasselai
Political Science & Scientific Computing (vasselai@umich.edu)

Motivation

Geolocation
Automagic geocoding services like those from Google Maps, ESRI and Bing Maps are convenient, but they:

- have accuracy kwnon to be much worse in rural and poorer areas.
- have inferior (and understudied) performance in developing countries
- are black boxes, with under-the-hood reference data and algorithm unkown

Interpolation
Point-in-polygon (data from overlapping polygon) has several drawbacks:



Point-to-point options like Kriging and Splines are undefined for compositional outputs (typical census data) and/or prohibitive for large N (like census blocks).

Geolocating with census data

Only way to improve over automagic services, manual geocoding requires:

Finding reference data. We propose using the GIS info on addresses and facilities used to organize countries' censuses. To illustrate the wide feasibility of this, we gathered names and links of such data for around 50 countries.

Matching to reference data. We derive a novel sub-string matching algorithm that sweeps through a name or address, each time fuzzy-searching for the correspondence of continuous sub-sentences with one less word than before (see algorithm below and example to the right).

'St. John Neuman School of Science' matches 'Saint John Neumann School' up to 1 typo?
St. John Neuman School of Science ✗
St. John Neuman School of Science ✗
St. John Neuman School of Science ✗
St. John Neuman School of Science ✗
St. John Neuman School of Science ✗
St. John Neuman School of Science ✗
St. John Neuman School of Science ✓

Algorithm Sub-Sentence Sweep Matching (simplified)

```
function S3M(P_{1×n}, T_{1×m}, rev, minw, maxd)
    range ← {minw : n} if rev else {1 : n−minw+1}
    if n ≥ minw then
        matched ← true
        j ← minw if rev else n − minw + 1
        while matched and j ∈ range do
            s ← 1 if rev else j
            e ← j if rev else n
            matched ← agrep(P_{[s:e]}, T, maxd)
            if matched then
                result ← j if reverse else n − j + 1
            end if
            j ← j + 1 if rev else j − 1
        end while
    else
        result ← n if strdist(P, T) ≤ maxd else result
    end if
    return result
end function
```

- $P_{1\times n}$: sequence of n words to be searched
- $T_{1\times m}$: seq. of m words where to search
- rev: whether to iterate backwards
- minw: min. number of words for a sub-sentence of P to be considered
- maxd: max. edit distance accepted
- agrep: whether a string is contained within another up to a maxd edit distance
- strdist: calculates edit distance

Validating the geolocation

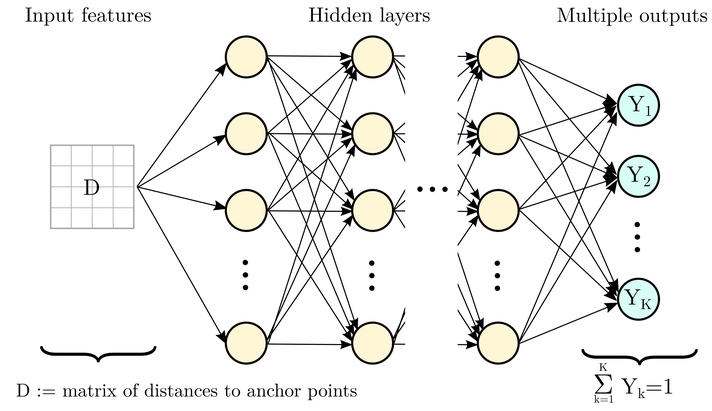
We geolocate the 136,000+ polling places used in Brazilian elections 2006-2020. Around 38% have ground-truth coordinates given by the official authority:

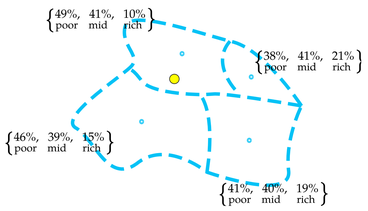
Median geolocation error in meters (1m ≈ .0006 miles), according to type of census block the geolocated polling place is located at.

source	overall	urban	rural	low inc.	mid inc.	high inc.
our method	**39**	**34**	**147**	**39**	**39**	**43**
Google Map (by address)	388	155	8642	738	88	53
Google Map (by facility name)	509	84	10558	1102	58	58
ArcGIS Map (by address)	514	188	10493	1009	149	69
Bing Map (by address)	1005	303	12037	3229	249	123
ArcGIS Map (by facility name)	2323	704	12259	3476	1104	1319
Bing Map (by facility name)	4109	1823	11945	5555	2064	3290

Interpolating from census data

Next, to interpolate census data to the geolocated polling places, we propose a Deep Neural Network defined for compositional outputs and whose only features are the distances to chosen anchor points:

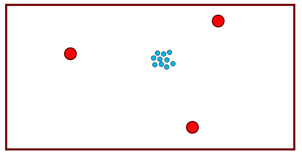




Compositional outputs. The DNN is trained to predict, from census points (e.g. census block centroids) to each polling place, not just a single value but a vector of K values that sum up to 1 (e.g. % of poor, mid and rich persons, in which case $K = 3$).

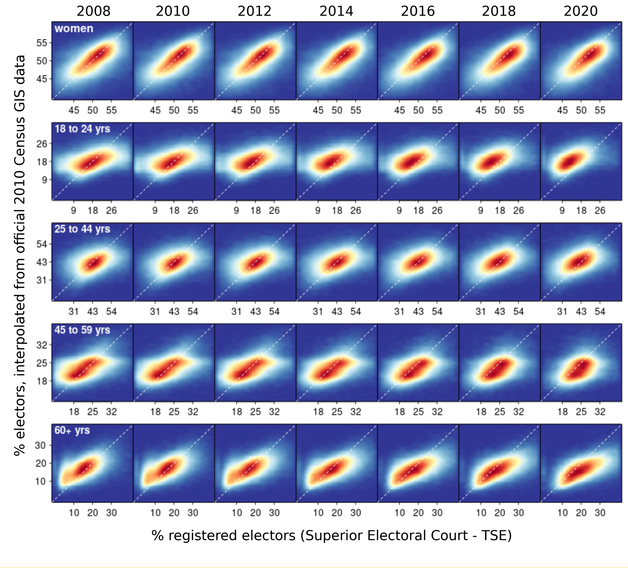
Anchors. Features are distances from training points (e.g. census block centroids) or test points (e.g. polling places) to chosen anchor points (bounding box corners, centroids of cities, etc). Since anchors don't vary, distances to them work as spatial embeddings, accounting for location and proximity between training/test points.

We know ● are close if we know their distances to ● (anchors) are similar:



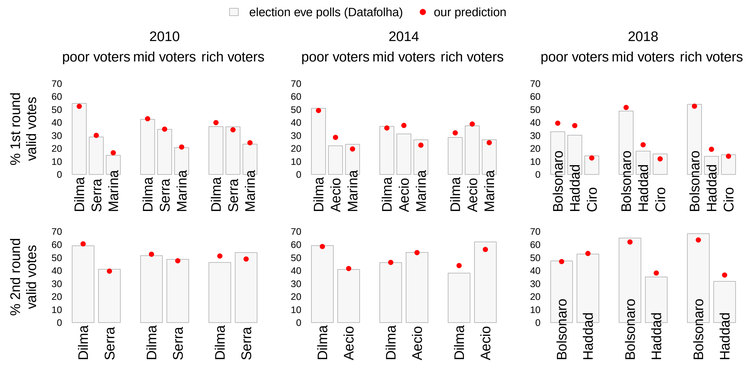
Validating the interpolation

We interpolate shares of electors by age and sex to the Brazilian polling places and compare to ground truth shares provided by the official electoral authority:

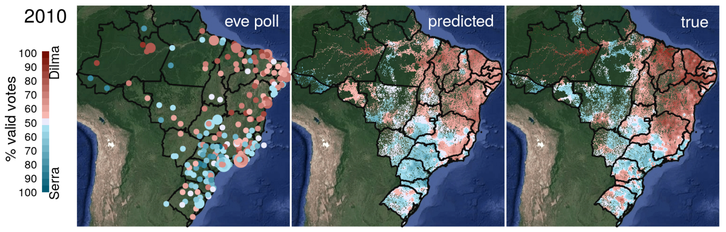


Initial applications

Income-group vote estimation using income info interpolated to polling places with our method, compared to results from large N (>10,000) election eve polls:



Voting-map forecasting. Using interpolated polling place demographics and geocoded eve poll responses to forecast the voting-map:



Effect of weather on turnout. Comming soon (in the paper!)