Language diversity and neighborhood characteristics in DC

Data

The data required for this project will come from two different sources. Basic data required to address the questions outlined include:

(1) Geospatial data about the city's neighborhoods

(2) Information about languages spoken in households in each neighborhood

(3) Foursquare data about the venues in each neighborhood.

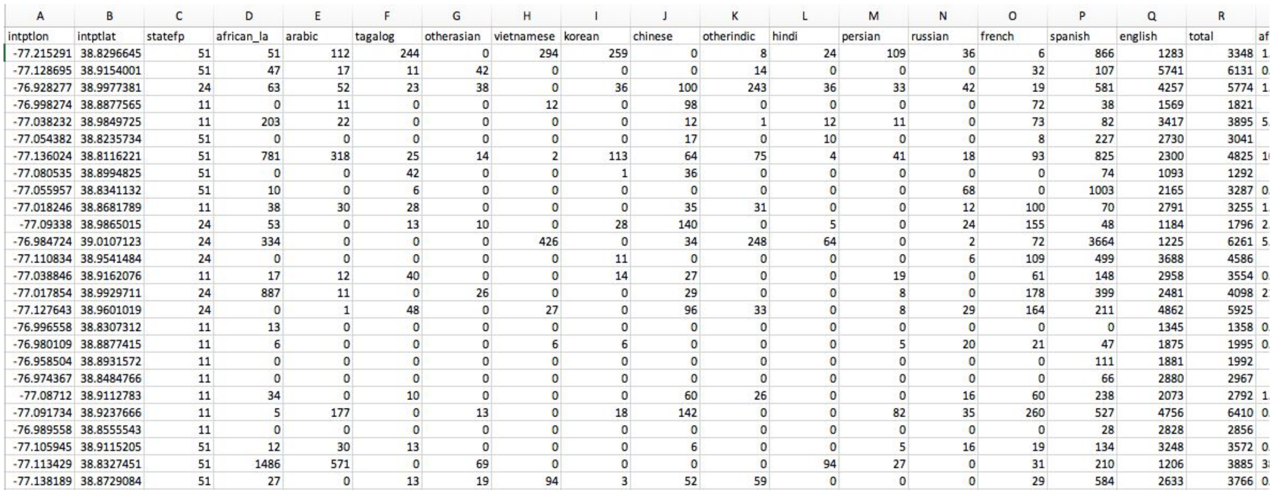
Data sources

American Community Survey dataset (via Randy Smith)
Geospatial and language data (combined) is obtained from a dataset downloaded from the site of Randy Smith, a GIS specialist at Hood College: https://randyhsmithjr.carto.com/tables/dcmetro/public

These data are in turn extracted from the American Community Survey, which is conducted periodically in between national censuses and is geographically grouped by census tract. Census tracts may not line up perfectly with named resident-identified neighborhoods, but are roughly neighborhood-sized (typically consisting of several thousand people). The ACS survey question underlying the dataset is a question about languages spoken at home.

Before importing the data, I have reduced the dataset to include only languages or language groups with more than 5000 speakers in the DC metro area. This is to make the number of languages more tractable and to ensure that our model focuses on languages with a significant community of speakers in the area.

Here's what a small segment of the data looks like:



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	
intptlon	intptlat	statefp	african_la	arabic	tagalog	otherasian	vietnamese	korean	chinese	otherindic	hindi	persian	russian	french	spanish	english	total	af
-77.215291	38.8296645	51	51	112	244	0	294	259	0	8	24	109	36	6	866	1283	3348	1.
-77.128695	38.9154001	51	47	17	11	42	0	0	0	14	0	0	0	32	107	5741	6131	0.
-76.928277	38.9977381	24	63	52	23	38	0	36	100	243	36	33	42	19	581	4257	5774	1.
-76.998274	38.8877565	11	0	11	0	0	12	0	98	0	0	0	0	72	38	1569	1821	
-77.038232	38.9849725	11	203	22	0	0	0	0	12	1	12	11	0	73	82	3417	3895	5.
-77.054382	38.8235734	51	0	0	0	0	0	0	17	0	10	0	0	8	227	2730	3041	
-77.136024	38.8116221	51	781	318	25	14	2	113	64	75	4	41	18	93	825	2300	4825	1
-77.080535	38.8994825	51	0	0	42	0	0	1	36	0	0	0	0	0	74	1093	1292	
-77.055957	38.8341132	51	10	0	6	0	0	0	0	0	0	0	68	0	1003	2165	3287	0.
-77.018246	38.8681789	11	38	30	28	0	0	0	35	31	0	0	12	100	70	2791	3255	1.
-77.09338	38.9865015	24	53	0	13	10	0	28	140	0	5	0	24	155	48	1184	1796	2.
-76.984724	39.0107123	24	334	0	0	0	426	0	34	248	64	0	2	72	3664	1225	6261	5.
-77.110834	38.9541484	24	0	0	0	0	0	11	0	0	0	0	6	109	499	3688	4586	
-77.038846	38.9162076	11	17	12	40	0	0	14	27	0	0	19	0	61	148	2958	3554	0.
-77.017854	38.9929711	24	887	11	0	26	0	0	29	0	0	8	0	178	399	2481	4098	2
-77.127643	38.9601019	24	0	1	48	0	27	0	96	33	0	8	29	164	211	4862	5925	
-76.996558	38.8307312	11	13	0	0	0	0	0	0	0	0	0	0	0	0	1345	1358	0.
-76.980109	38.8877415	11	6	0	0	0	6	6	0	0	0	5	20	21	47	1875	1995	0.
-76.958504	38.8931572	11	0	0	0	0	0	0	0	0	0	0	0	0	111	1881	1992	
-76.974367	38.8484766	11	0	0	0	0	0	0	0	0	0	0	0	0	66	2880	2967	
-77.08712	38.9112783	11	34	0	10	0	0	0	60	26	0	0	16	60	238	2073	2792	1.
-77.091734	38.9237666	11	5	177	0	13	0	18	142	0	0	82	35	260	527	4756	6410	0.
-76.989558	38.8555543	11	0	0	0	0	0	0	0	0	0	0	0	0	28	2828	2856	
-77.105945	38.9115205	51	12	30	13	0	0	0	6	0	0	5	16	19	134	3248	3572	0.
-77.113429	38.8327451	51	1486	571	0	69	0	0	0	0	94	27	0	31	210	1206	3885	3
-77.138189	38.8729084	51	27	0	13	19	94	3	52	59	0	0	0	29	584	2633	3766	0.

Some characteristics to note:

- Some languages are grouped together. These are generally languages which are geographically related rather than related in other ways, e.g. the groups “African languages” and “Other Asian”. This is an inherent limit in the ACS data - small languages are grouped together.
- There are no neighborhood names. Census tracts are identified by complex numbers, e.g. “Census Tract 4523.01”, which I am opting not to use. Instead, I will assign “Neighborhood ID” numbers to each census tract. In a future project, the lat-long values could be aligned with neighborhood names in DC, though the mapping will not be simple.
- Dataset provides raw number of speakers of each languages within the census tract.
- Additional pre-processing steps will be undertaken, including limiting the dataset to only those census tracts within the District of Columbia proper, which is bounded on the west side by the Potomac River and extends across the Anacostia River in the east.

Foursquare API venue data

The linguistic and geospatial data will be combined with data about venues obtained via the Foursquare API. For each latitude-longitude in the geospatial data (i.e. for each census tract in DC), a list of venues close to its center will be obtained (grocery stores, restaurants, coffee shops, etc.)

The venue categories (and the diversity of categories present in each neighborhood) will be the most important feature of these data for the purposes of this project.

The data obtained from Foursquare will be used to develop a picture of the local character of each neighborhood in terms of its venues: does it have a variety of places to eat, entertainment, and other amenities?