Spatial Processing: Inside the Black Box

Introduction

There is little doubt that "GIS for everyone" has been a persistent theme in software development over the past several decades. ESRI, creator and purveyor of globally popular "GIS" software, evolved from command line processing (ArcInfo) to desktop GUI's for point and click (ArcVIew to ArcMap), to web based "black box" processing of data to perform analytics, manage data, and make maps and apps (ArcGIS Online, see "What is ArcGIS Online? https://doc.arcgis.com/en/arcgis-online/reference/what-is-agol.htm). This evolution toward "GIS for everyone" has raised flags among "GIS" professionals concerning unintended consequences of downplaying the importance of the profession. Stephen Keen (March 2014) in his article "GIS for Everyone? writes:

> But in our fervor to spread the GIS message, we are indeed turning to strangers and saying, "Come on, have a go, everyone can do it, and it is fun," when sometimes we should be saying, "Put your hands in the air and step away from the data." (https://www.directionsmag.com/article/1422, accessed 3/29/2019)

Concerning improved availability of and access to spatial data, Bearman et. al. (2016) write

> …[This] has reduced the level of spatial literacy among those who use spatial data, endangering these skills of critical spatial thinking. The increased availability of geodata has opened up the potential of applying GIS and geocomputation techniques to a wide variety of areas, but a lack of sufficient geospatial skills and low levels of spatial literacy have severely inhibited this in the UK (ESRC, 2013 International benchmarking review of UK human geography. Economic and Social Research Council. Retrieved from http://www.esrc.ac.uk/funding-and-guidance/tools-and-resources/impact-evaluation/UK-human-geography.aspx).

In addition, they propose that GIS education focuses on the "technical skills associated with using " GIS software "rather than developing the theoretical understanding of spatial problems, the science behind this (i.e. Geographic Information Science) and the usefulness of spatial data…" Bearman et. al. argue that the emphasis on technological skills is a result of the early GIS issues associated with computer processing power and data storage. Now, however, GIS education should refocus on "what do we want the technology to do?" rather than "how can we get the technology to do what we want?".

In the past few years, the tide has turned a bit primarily due to the rise in popularity of open source programming languages and environments like R (Rstudio) and Python (Jupyter Notebook). Both languages offer a wide variety of GIS related libraries/modules for geospatial and geostatistical analysis and visualizations (2D, 3D maps, and charts). Results may be offered as static or interactive graphics on the internet. Recently, ESRI has incorporated R and Python into their workflow discussions. The ArcGIS API for Python is now in version 1.3 and python notebooks have been appearing on ESRI's developer site.

The March 2019 ESRI Training newsletter promotes "Python for every GIS professional" mentality.

> Hands down, Python is the top scripting language favored by ArcGIS users, and it is very popular throughout the GIS community. But not all GIS professionals have felt the need

> to learn a scripting language.
>
> A new ArcGIS blog series aims to change that by introducing Python to nonprogramming GIS users. Authored by Esri's Olivia Iannone, herself a nonprogramming GIS user, the blog series is designed as a let's learn together approach to mastering Python basics.

Additionally, ESRI is sponsoring a new free Training Seminar scheduled to occur April 13, 2019: The seminar description is:

> **Integrated scripting, spatial analytics, and data science.**
>
> This seminar introduces basic concepts of data science, machine learning, and artificial intelligence (AI) in the context of ArcGIS Notebooks—a new Python scripting environment in ArcGIS Enterprise. The presenters share Python scripting tips and tricks using examples that include Jupyter Notebook, ArcPy, and ArcGIS API for Python. You'll see how to script an analysis, automate repetitive tasks, and discover how the fields of spatial data science and GeoAI can help GIS professionals, researchers, scientists, and data scientists more fully understand and solve complex problems.

What have these trends meant for GIS educators who debate the importance of teaching fundamental concepts, such as digital representation of spatial phenomena, standard spatial processing techniques like buffering, overlay: intersection, union, difference, as well as spatial autocorrelation and its impact on pattern analysis and interpolation? Most importantly, to teach GIScience effectively, should educators emphasize how to put these techniques and methods together to solve problems? Teaching point and click type GIS modules where students are provided data and "cookbook" type instructions to process the data overlooks the importance understanding the whole problem-solving process. And, as Bearman et. al. write:

> This approach develops the student's ability to use the software in question, but it is debatable whether this approach adds much to their knowledge about the types of question that a GIS can answer, or how to apply the tools available to other data-sets.

What follows is an example from teaching graduate level geospatial programming classes using python and cross-listed undergraduate/graduate course on web – based mapping that currently focuses on ArcGIS Online. In the exercise demonstrated here, the topic for both classes was spatial processing. Workflow and results are compared using

1. ESRI ArcGIS Online *Summarize Near* analysis tool (Black Box)
2. ArcGIS Online Dissolve, Buffer, Overlay, and Join tools (Under the Hood)
3. Python

The original intent of the exercise was to teach undergraduates how basic spatial processing methods can be strung together for data analysis. However, the algorithm (workflow) applies to the programming approach too. Data acquisition, prep, and analysis are discussed to compare the approaches described above.

Workflow

Understand the problem

Before undertaking GIS based analysis, understand the question. The question asked was:

- How many people live in the shadow of limited access highways in Buffalo, NY? Describe the characteristics of this population.

Students should consider the following with an eye toward understanding how these initial decisions may impact results.

1) What data are needed?

   The location of limited access highways and a polygon feature to summarize characteristics of the population.

2) What is meant by "shadow"?

   Typically, ask the sponsor/employer/GIS team, or research planning literature, or apply a range of distances for the analysis. For this exercise use ½ mile distance to characterize the highway shadow (impact zone).

3) What is meant by "characteristics"?

   Again, analysts should clarify this with the sponsor/employer/GIS team and/or research the literature. In this example, race/ethnicity and housing occupancy are the characteristics used since this information is part of the city's open dataset. Other data from the Census or ESRI data enrichment may be incorporated later.

4) For what period of time is the answer required?

   It seems that the City's data are from the 2010 Census although the metadata does not specify the time period. This may become a temporal analysis to investigate how these characteristics have changed over time.

5) How to summarize results?

   A variety of measures may be applied such as

   - Totals,
   - Area density (totals per area),
   - Linear density (totals per linear distance)

   Summary measures should be calculated for each highway but an additional consideration may be to investigate the number of people living in the shadow of two highways, three highways, etc.

Data Acquisition

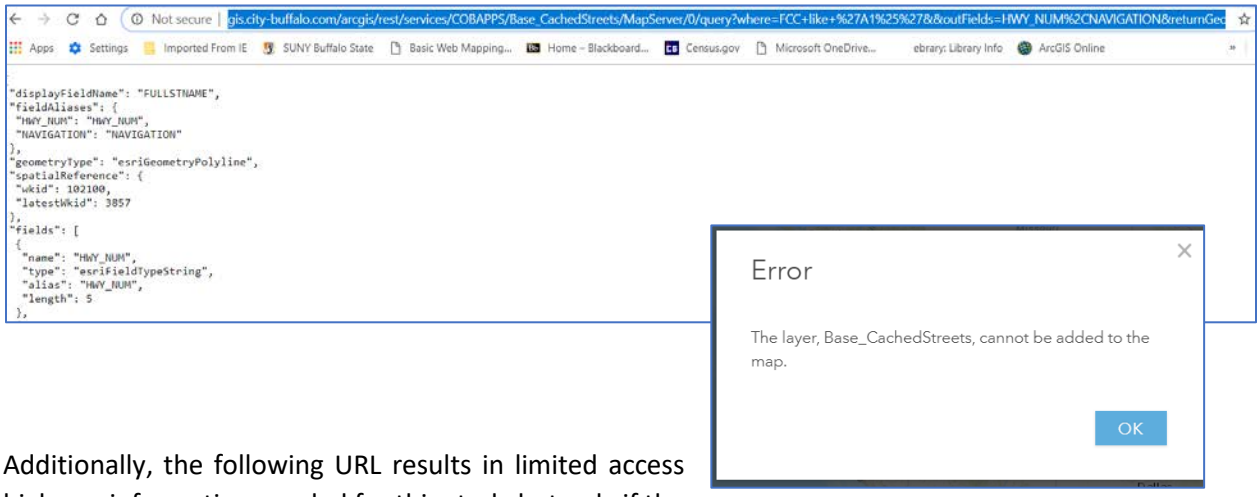
For all approaches, the next step is to get the data. The City provides open data via a Socrata platform (for the python class) and REST services (for the web-mapping class). The Socrata site names these geospatial datasets "Roads" and "Block Groups".

ArcGIS Online:

REST services for the required data are:

- Highways: http://gis.city-buffalo.com/arcgis/rest/services/COBAPPS/Base_CachedStreets/MapServer/0
- Block Groups: http://gis.city-buffalo.com/arcgis/rest/services/COBAPPS/Open_Data/MapServer/16

Since only limited access highways are desired an SQL statement using the highways REST service query function (?where=FCC like 'A1%') might be used. However, this query results in an error in ArcGIS Online.



Additionally, the following URL results in limited access highway information needed for this study but only if the format is KML. (every other query with valid SQL fails!)

- http://gis.city-buffalo.com/arcgis/rest/services/COBAPPS/Base_CachedStreets/MapServer/0/query?where=FCC+LIKE+%27A1%25%27&outFields=HWY_NUM%2C+NAVIGATION&f=KMZ

BUT users cannot perform analysis on a KML so the easiest thing to do is add the entire highways dataset then Filter it using FCC starts with A1. Finally, notice that the Spatial Reference of the data is found in the Details of the input layer.

Python (Jupyter Notebook):

In python the URL may contain SQL "where" and "select" conditions that results in a Geopandas dataframe containing polyline segments (geometry) and attributes. The same is true for importing the block group layer; select is used to reduce the number of columns to only those needed in the analysis.

The coordinate reference system (CRS) of the input data is identified and easily converted to another CRS if desired. The input data is in WGS84 (EPSG:4326) with decimal degree units. It is converted to WGS84 (EPSG: 3857) with units in meters.

Get highway data and plot

Use select = to limit the fields read in to the geodataFrame.

```
In [2]: url_rd ="https://data.buffalony.gov/resource/45ez-ykin.geojson?"+\
        "$select=the_geom,hwy_num,navigation"+\
        "&$where=fcc%20like%20%27A1%25%27"+\
        "&$limit=1000000"

In [3]: rd = gpd.read_file(url_rd)

In [4]: rd.head()
```

Out[4]:

	hwy_num	navigation	geometry
0	190	S	(LINESTRING (-78.880825967333 42.881363036534,...
1	5	S	(LINESTRING (-78.865562006088 42.854303043079,...
2	190	S	(LINESTRING (-78.841781969994 42.870071965687,...
3	190	S	(LINESTRING (-78.857530971801 42.873273963022,...
4	33	W	(LINESTRING (-78.84206404133801 42.92416998742...

```
In [6]: rd = rd.to_crs({'init': 'epsg:3857'})
```

Get Block Group data and overlay with buffers

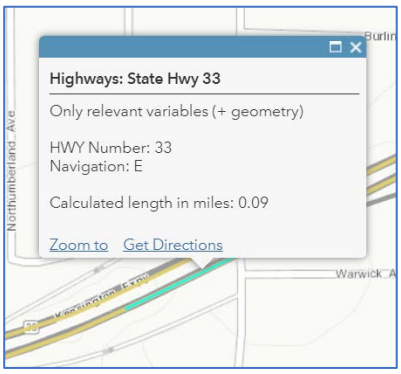
- Calculate area in sq. kilo of block groups.
- Calculate area in sq. kilo of intersected block groups.

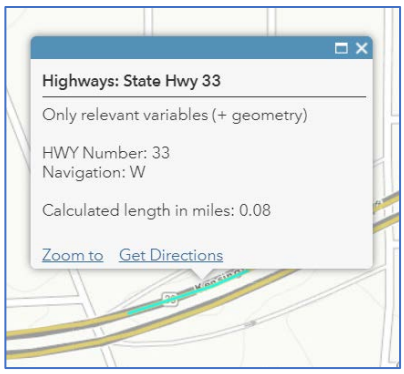
```
In [9]: url_bg = "https://data.buffalony.gov/resource/u7fm-4kwb.geojson?"+\
        "$select=the_geom,geoid10,total_pop,white,black,hispanic,hse_units,occup"+\
        "&$limit=1000000"

        bg=gpd.read_file(url_bg)
        bg = bg.to_crs({'init': 'epsg:3857'})
        bg['area_o']= bg['geometry'].area/10**6 #sq kilometers
        bg.shape[0]
```

Out[9]: 287

Data Prep

For this exercise, the conceptual representation of "limited access highway" is an entire stretch of roadway associated with a route within the city limits. Direction of traffic flow is not important. This concept does not correspond to the digital representation. In the digital representation each highway has two directions represented with different collections of polylines. The number and length of individual polyline segments varies.



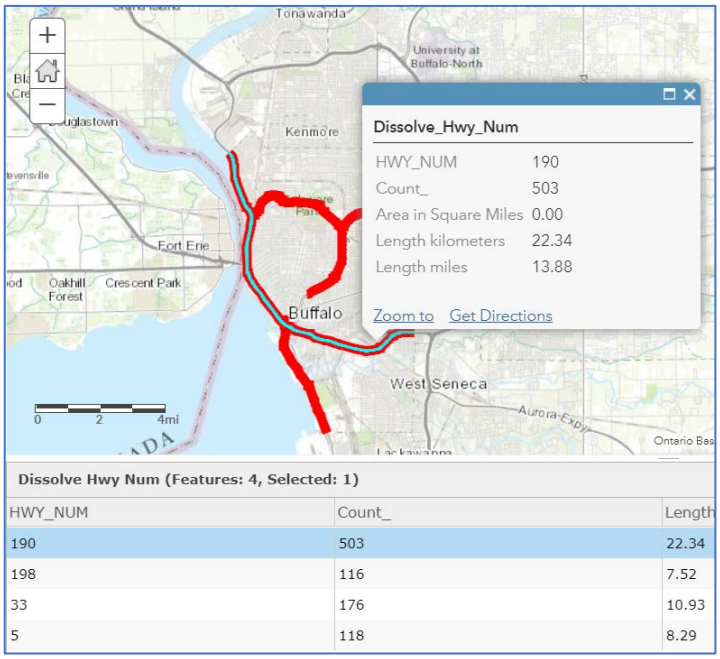


ArcGIS Online:

One approach is to buffer each segment and merge the buffers on the highway number. However, the ArcGIS Online Analysis Tool *Summarize Nearby* takes a very long time to run. There are 938 polyline segments representing the four limited access highways in Buffalo. So instead, the dissolve method may

be used to create one feature, that is a collection of many segments, for each of the four limited access polylines.

The dissolve method in ArcGIS Online Analysis/Manage Data seems to be designed for polygons. The tool name, "Dissolve Boundaries", and the instruction, "Choose area layer whose boundaries will be dissolved", along with the help information reinforce this notion. However, when applied to the filtered highway layer, setting the method to "Areas with the same field Value" = HWY_NUM and "Create multipart features" selected, the desired results are achieved.



Dissolve Hwy Num (Features: 4, Selected: 1)

HWY_NUM	Count_	Length
190	503	22.34
198	116	7.52
33	176	10.93
5	118	8.29

Python (Jupyter Notebook):

Geopandas' dissolve method is followed by a statement to calculate the length of each highway corridor.

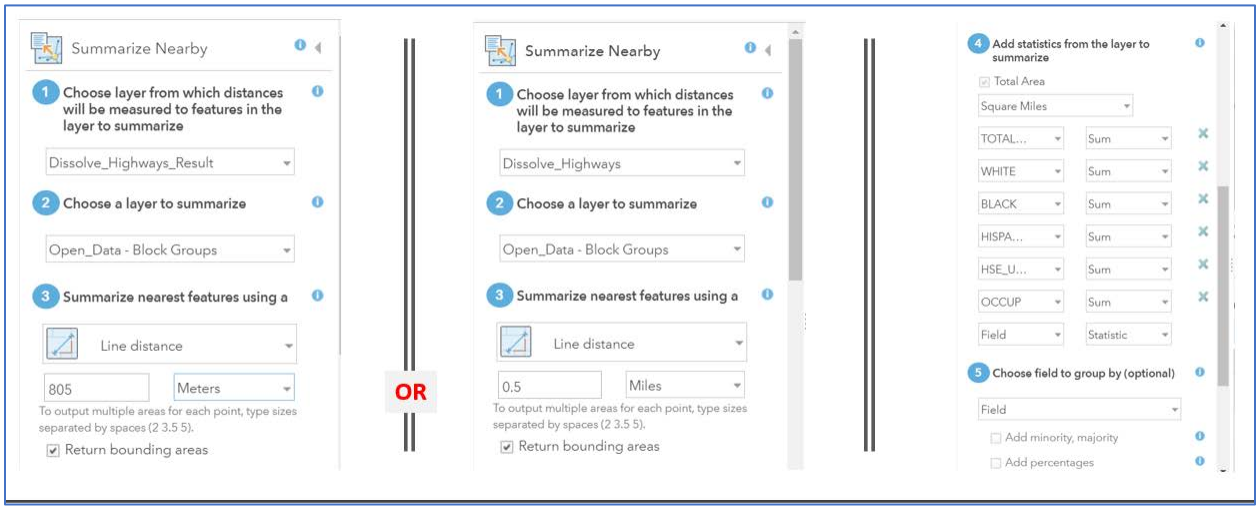
```
hwys = rd.dissolve(by='hwy_num').copy().reset_index()
hwys['lengthm']=(hwys.geometry.length*0.000621371)/2
hwys.head(10)
```

	hwy_num	geometry	navigation	lengthm
0	190	(LINESTRING (-8775153.716918258 5291882.042280...	S	13.880485
1	198	(LINESTRING (-8776819.278160788 5299771.023128...	W	4.675618
2	33	(LINESTRING (-8779138.063381437 5295734.920947...	W	6.788832
3	5	(LINESTRING (-8777898.62519433 5286538.6148021...	S	5.149639

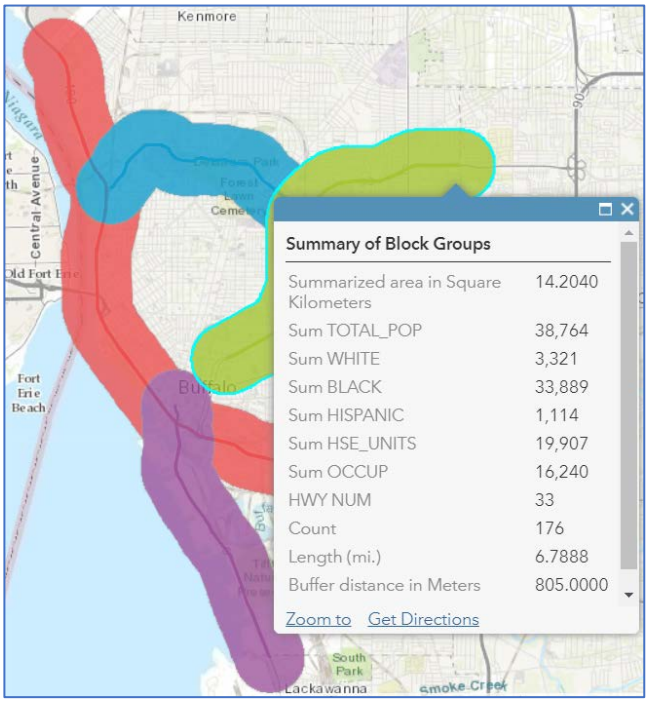
Data Analysis

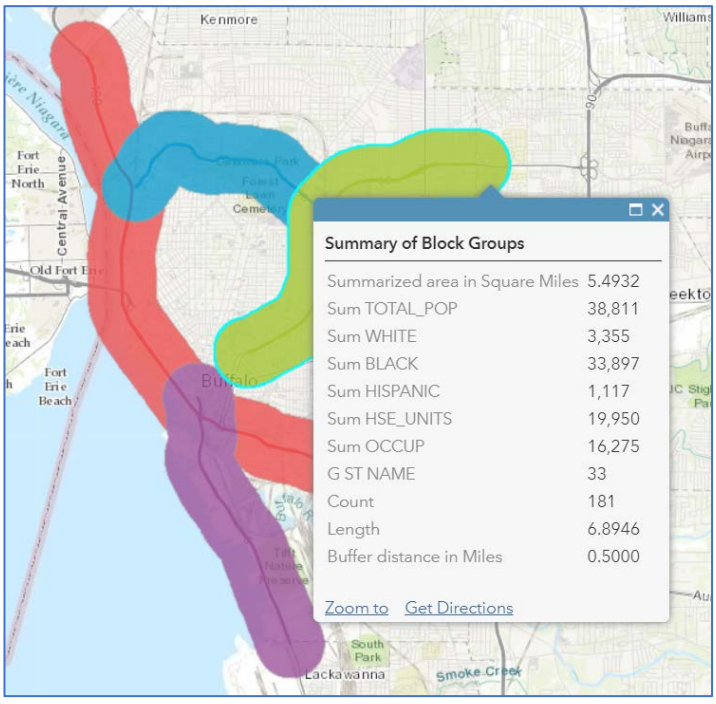
ArcGIS Online - Summarize Nearby Tool

The first approach is to use the "*Summarize Nearby*" analysis tool on the dissolved highways using either 805 meters or ½ mile as the distance. The results for each highway should contain the total population, white population, black population, Hispanic population, total housing units, and occupied housing units.



Results vary slightly depending on the buffer distance used.

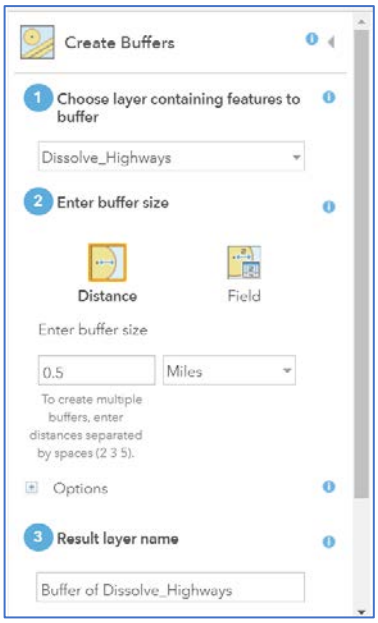


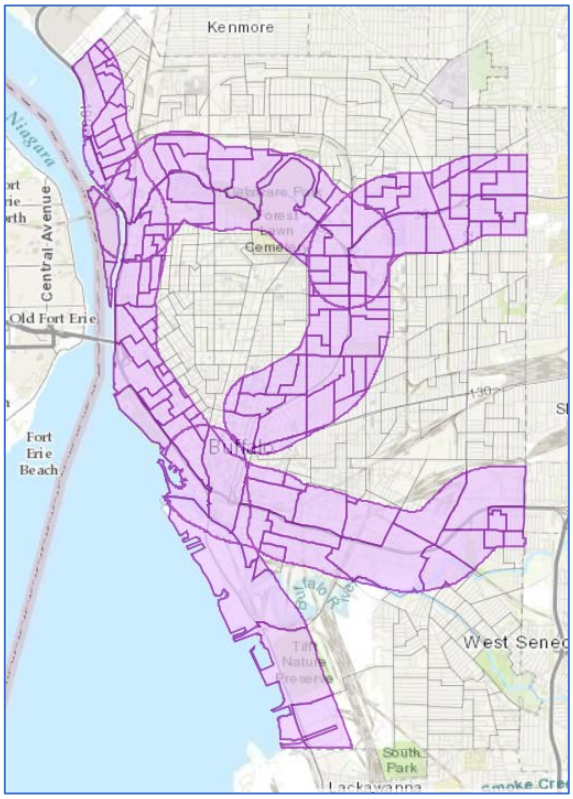


ArcGIS Online - Inside the Black Box

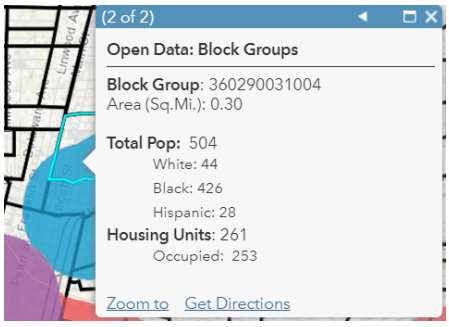
This first approach offers no insight into spatial processing; tools like buffer and overlay. So a second analysis approach, manually manipulating the data using ArcGIS Online, is undertaken to help students learn what tools/methods are used "under the hood" of the "*Summarize Nearby*" tool. Experience with the fundamentals reinforces an understanding of the data, especially how spatial representation impacts and is impacted by spatial processing methods.

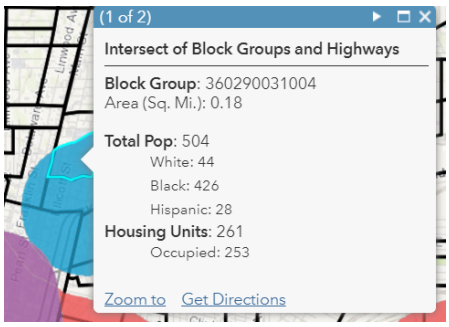
"Create Buffers" analysis tool defines the shadow area of each highway. Then the *Overlay* tool, intersect method breaks each buffered area into a collection of block group pieces.





Overlay/Intersect does NOT assign values to polygons based on the proportion of the intersection area to the original area. As shown below, the same values are assigned to different polygons: entire block group and partial block group near the highway.





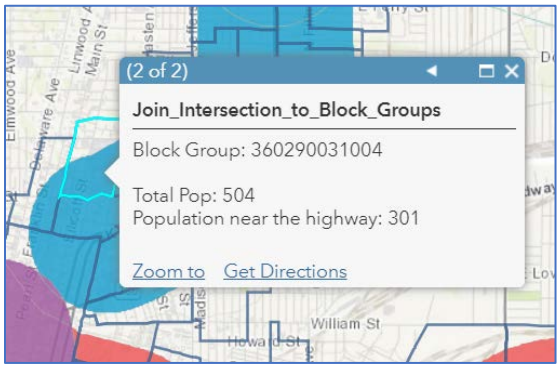
According to the "*Summarize Nearby*" tool documentation, values are allocated to the intersect polygons based on the ratio of the new polygon area to the original polygon area. The underlying assumption is that the characteristics of interest are evenly distributed throughout each polygon. This may not be valid

in areas with many high-rise housing units and/or areas dominated by parks, commercial, and retail development.

Before the intersect operation, Arcade is used on the polygon layer to calculate the area of each block group. Then the new layer from the intersect will contain this total area information. After the intersect operation, the intersect polygons' area are calculated using Arcade. A new field for each characteristic of interest is added to the intersect attribute table and the ratio of intersect area to original area is multiplied by the total block group value to determine the proportion of the original value that falls in the highway shadow.

An example of the Arcade expression for total population is:

round($feature["TOTAL_POP"]*($feature.AreaSqMi/Area($feature, 'square-miles')))



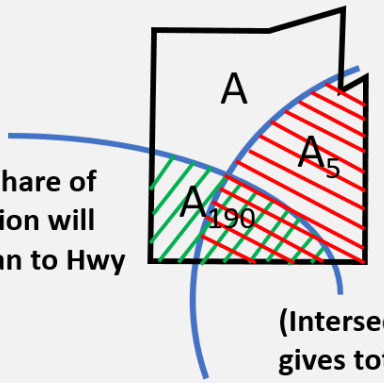
Finally, dissolve the intersection layer by the highway number to get the total population in the shadow of each highway. This step may be a bit confusing if you do not understand the structure of the spatial dataset resulting from the intersect process.

What happens when a block group falls in the shadow of more than one highway? The shadow areas may intersect each other (as shown below) or may result in separate pieces. In either case, the id for the original polygon will appear multiple times in the intersect layer. The issue involves how to identify the population in the shadow of two (or more) highways, the $A5 \cap A190$ piece in the diagram.

A = Area of Block Group

A_5 = Area of Hwy 5 buffer intersecting block group

A_{190} = Area of Hwy 190 buffer intersecting block group

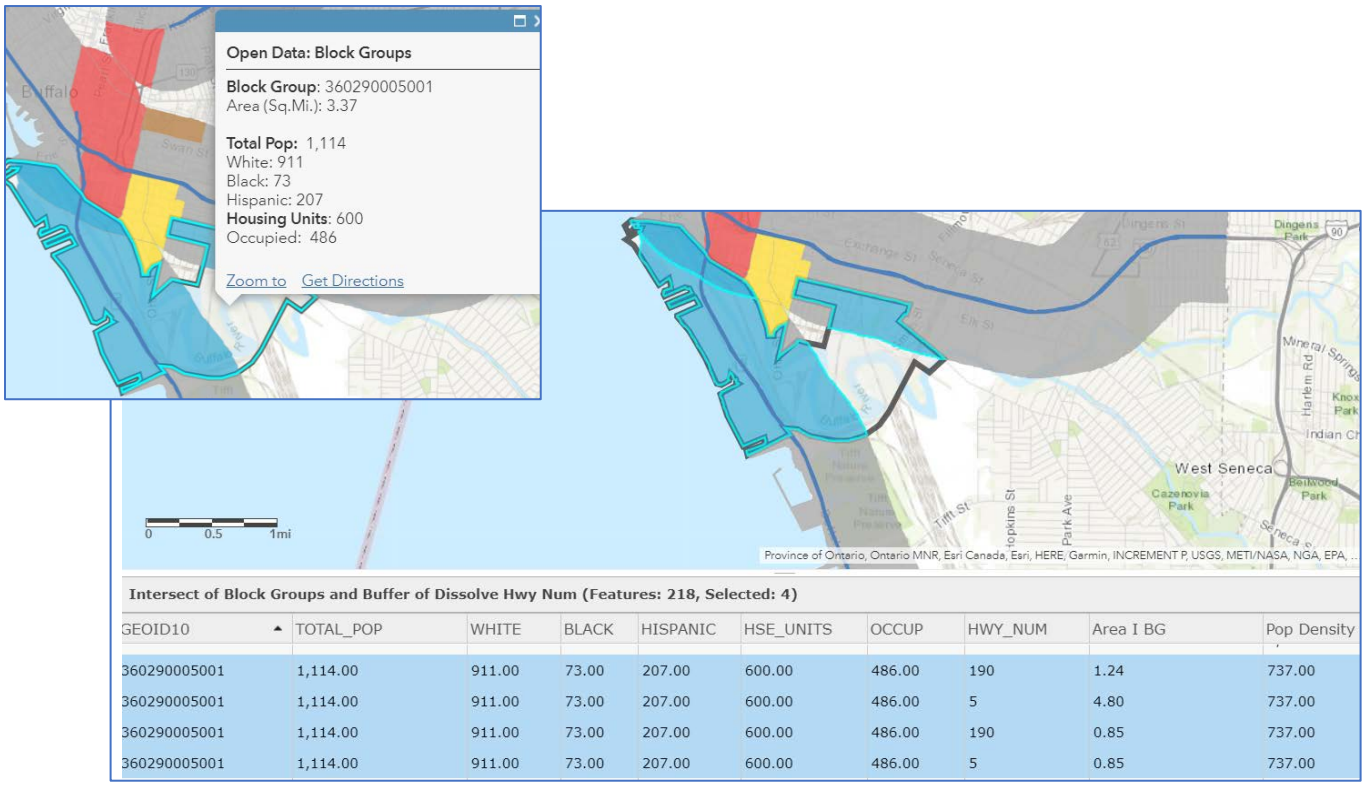


The GEOID of the block group will appear four times in the intersection layer: twice for hwy_num = 5 and twice for hwy_num =190.

A_5/A > A_{190}/A so a larger share of the block group's population will be allocated to HWY 5 than to Hwy 190

(Intersection of A_5 & A_{190})/A gives total number of persons near two highways

An example from the ArcGIS Online results shows that there is no direct way to estimate this population unless you assume that if two polygons have the same area and different highway numbers then they represent the same location within the original block group.



Intersect of Block Groups and Buffer of Dissolve Hwy Num (Features: 218, Selected: 4)

GEOID10	TOTAL_POP	WHITE	BLACK	HISPANIC	HSE_UNITS	OCCUP	HWY_NUM	Area I BG	Pop Density
360290005001	1,114.00	911.00	73.00	207.00	600.00	486.00	190	1.24	737.00
360290005001	1,114.00	911.00	73.00	207.00	600.00	486.00	5	4.80	737.00
360290005001	1,114.00	911.00	73.00	207.00	600.00	486.00	190	0.85	737.00
360290005001	1,114.00	911.00	73.00	207.00	600.00	486.00	5	0.85	737.00

Total Pop	BG area	Area of Intersection	HWY	Intersection Pop
1114	3.37	1.24	190	410
		4.8	5	1587
		0.85	190	281
		0.85	5	281

Population Near		
	Hwy 5	1868
	Hwy 190	691
	two hwys	281

This topic will be discussed further using the Python method.

To calculate summary statistics, it is necessary to add a field for each statistic then write an Arcade expression to calculate each new field. In this application, nine statistics were calculated. The analyst adds nine fields and calculates each one individually. This gets a bit tedious!

Python (Jupyter Notebook):

Using python, the Geopandas buffer method is applied to the dissolved highways.

```
hwys['bufgeo']=hwys.buffer(805)# buffer returns a geoseries
hwys=hwys.set_geometry('bufgeo')
```

Then the overlay module is applied to get the intersection between the buffered highways and the block groups. The area of the intersected polygons is computed.

```
from geopandas.tools import overlay
newdf = overlay(hwys, bg, how="intersection")#returns only those geometries that are contained by both GeoDataFrames
newdf['area_n']=newdf['geometry'].area/10**6
```

New fields containing the desired statistics are calculated. Results are aggregated by highway.

```
newdf['area_ratio']=newdf['area_n']/newdf['area_o']
```

```
fields = ['total_pop','black','white','hispanic','hse_units','occup']
newfields = ['tPop_n','black_n','white_n','hisp_n','hu_n', 'ohu_n']
newdf[fields] = newdf[fields].apply(pd.to_numeric, errors='coerce')
```

```
for f in zip(newfields,fields):
    newdf[f[0]]=newdf['area_ratio']*newdf[f[1]]
    newdf[f[0]]=newdf[f[0]].round()
```

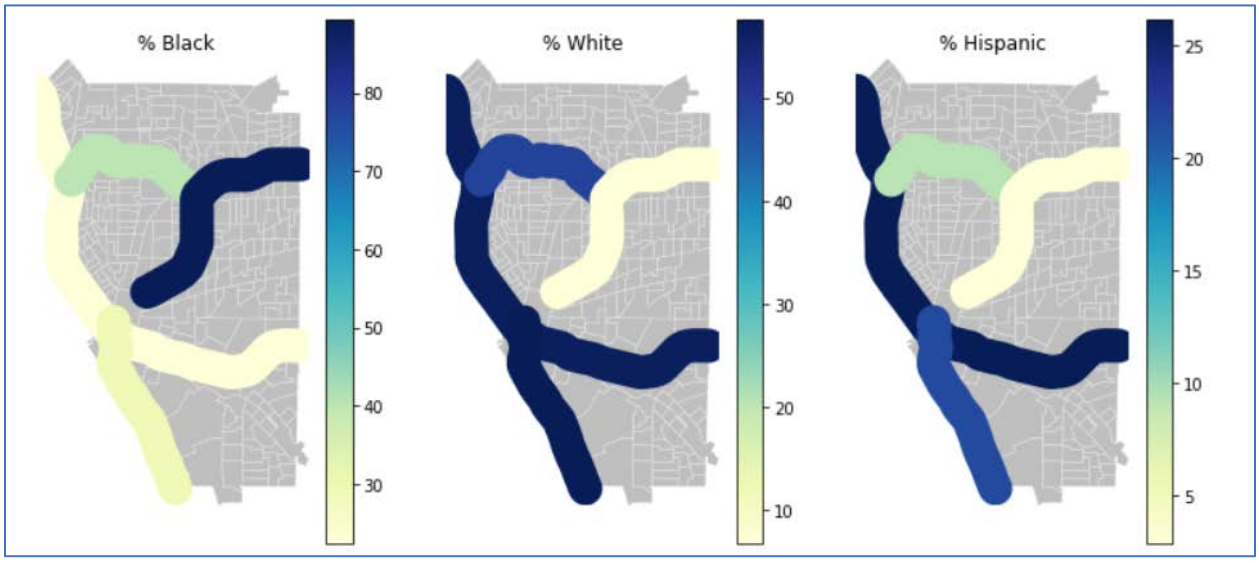
This table is merged with the highway buffer layer and summary statistics calculated for each highway shadow. These statistics include the proportion of those in the shadow by race and ethnicity as well as the proportion of housing units that are occupied. Linear density, such as the number of white persons per mile, is determined as is the ratio of Black to White persons.

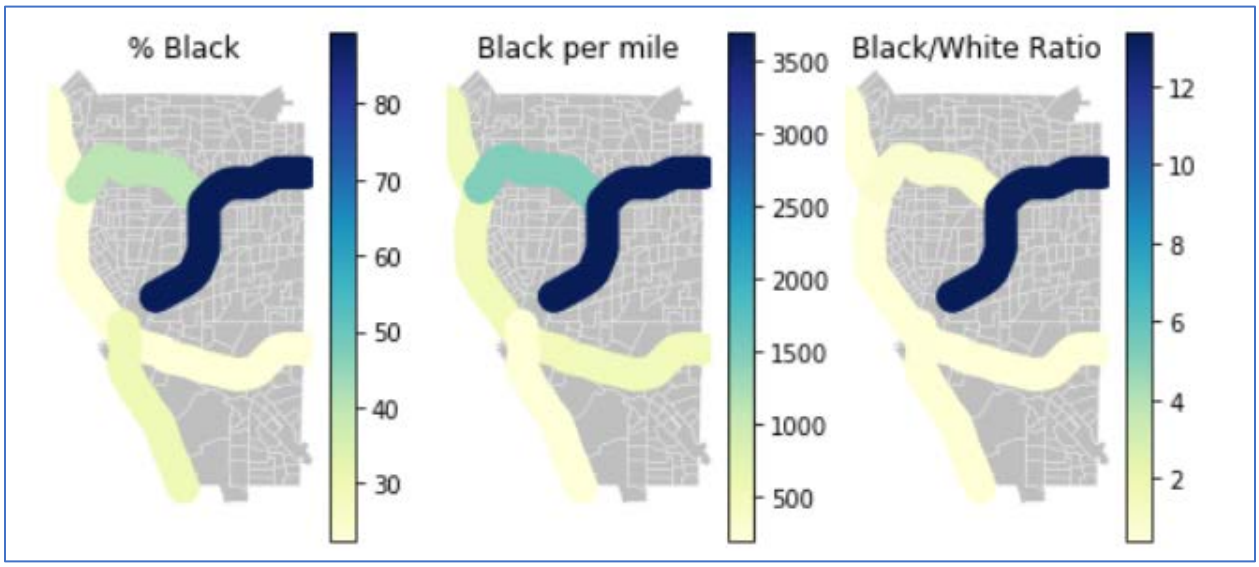
Aggregate Population and housing data by road name

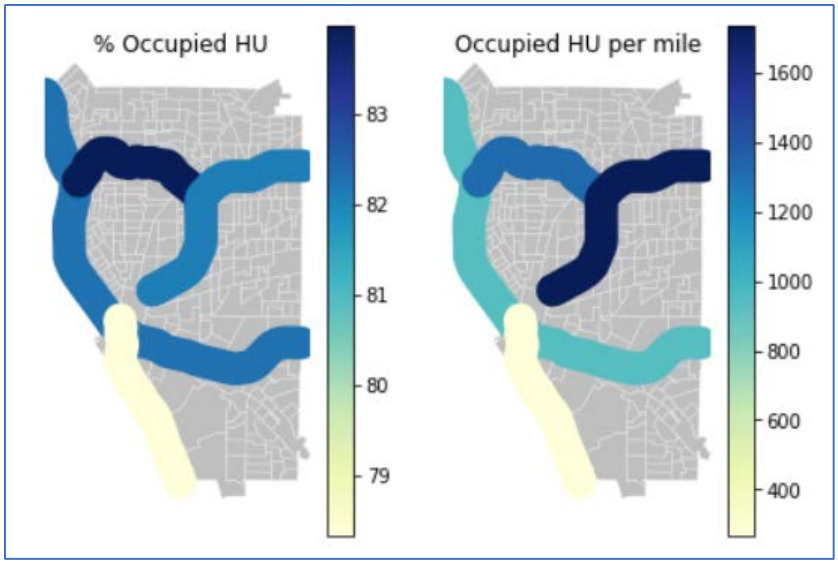
```
fields = ['hwy_num','total_pop','tPopn','hispanic','hisp_n',
          'black','black_n','white', 'white_n','hse_units','hu_n',
          'occup', 'ohu_n' ]
newdf1 = newdf.loc[:, fields]
groupdf = newdf1.groupby(['hwy_num']).sum().reset_index()
groupdf
```

	hwy_num	total_pop	tPopn	hispanic	hisp_n	black	black_n	white	white_n	hse_units	hu_n	occup	ohu_n
0	190	52082	31206.0	12798	8166.0	12226	6966.0	29522	17772.0	26151	15802.0	21602	13002.0
1	198	29878	17195.0	2865	1599.0	13727	6904.0	12910	8413.0	13670	7359.0	11378	6180.0
2	33	48233	28127.0	1528	801.0	40812	25136.0	5393	1879.0	24850	14382.0	20205	11809.0
3	5	8986	3216.0	2042	697.0	2493	946.0	5257	1853.0	5160	1739.0	4117	1362.0

Another advantage of python is the ability to plot several maps at a time for comparison. This is not possible in ArcGIS Online.



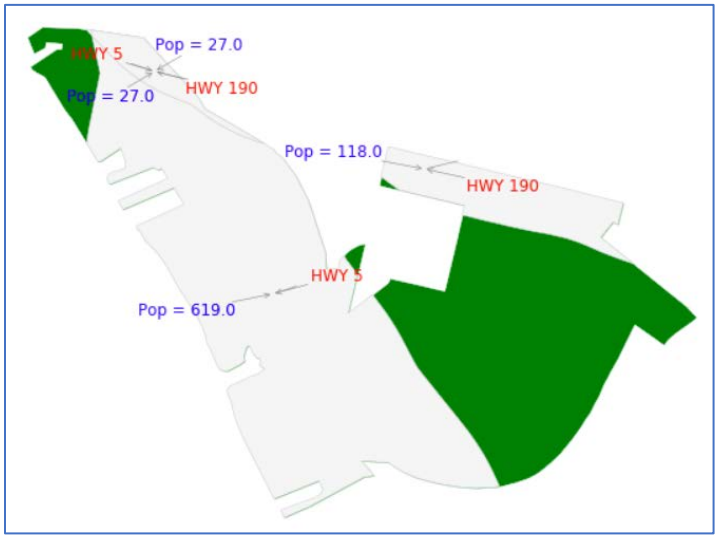




To answer questions concerning the number of people living in the shadow of two or more highways intersect the highway buffers with themselves, first. The resulting table assigns both highway numbers for intersected areas. Polygon ids identify block groups with multiple pieces. If the highway numbers in the intersection with the block groups are equal that population falls in the shadow of one highway. If they are different the population falls in the shadow of two highways.

```
overlapdf = ifdf[ifdf['geoid10']== '360290005001'].copy()
dups = ifdf[(ifdf['geoid10']== '360290005001') & (ifdf['hwy_num_1']!=ifdf['hwy_num_2'])].copy()
origbg = bg[bg['geoid10']=='360290005001'].copy()
overlapdf.shape[0],dups.shape[0]
```

```
(4, 2)
```



Results and Conclusions

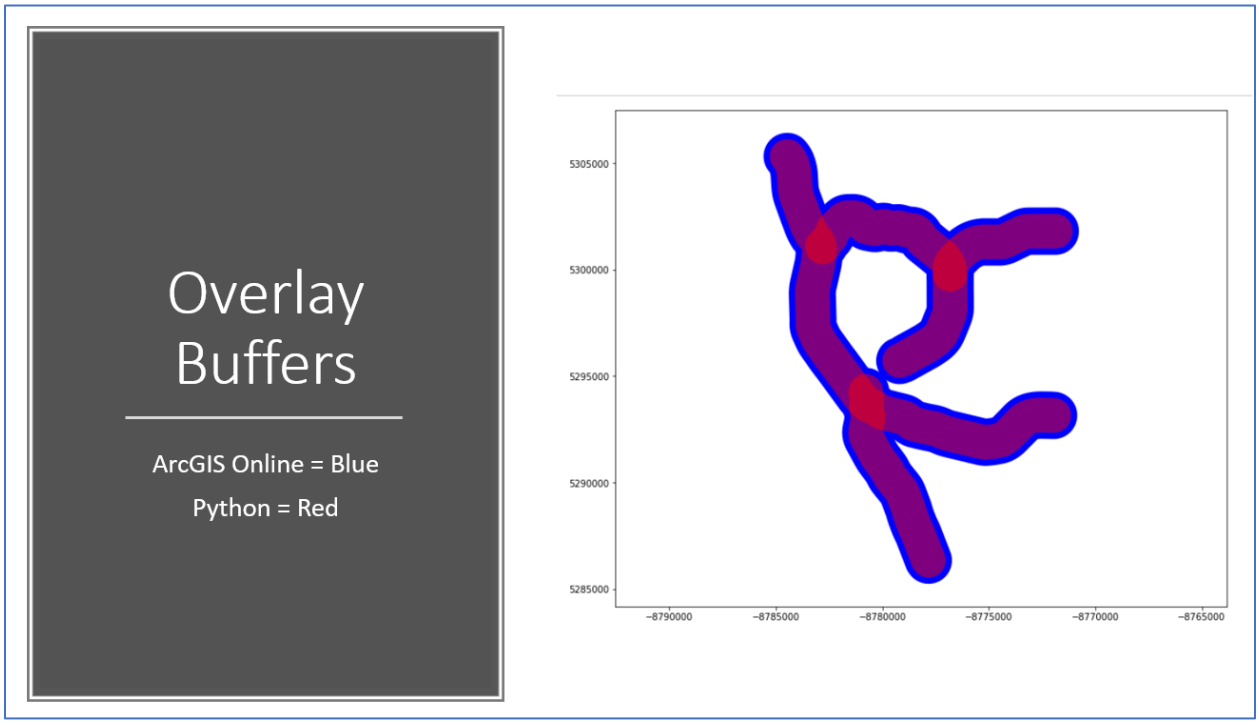
The results from Python are not the same as results from ArcGIS Online. Buffers associated with both ArcGIS online methods are larger than buffers from the python method. This results in higher estimates of the population within the shadow of each highway.

Total Population

HWY	ArcGIS Online	Python	% *Difference*
190	44,076	31,206	41%
198	25,152	17,195	46%
33	38,764	28,127	38%
5	4,884	3,216	52%

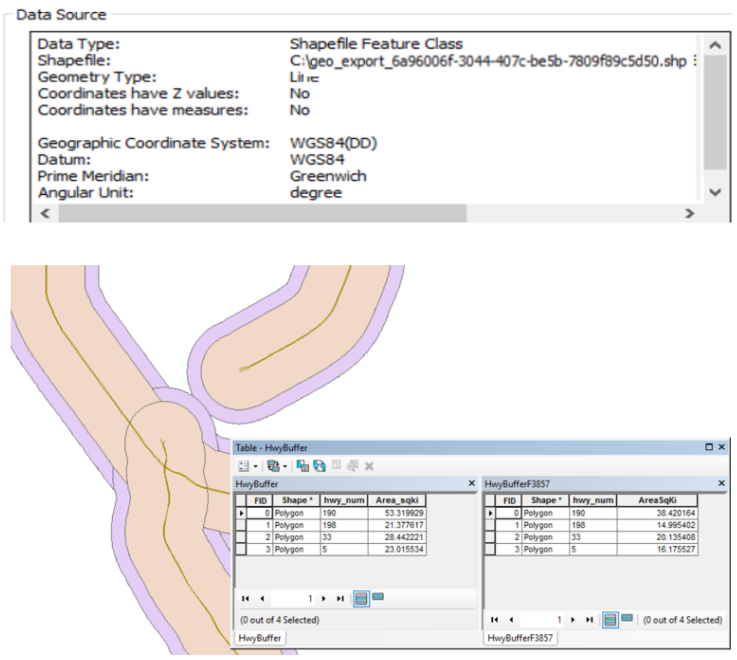
To complicate matters, both ArcGIS Online analysis tools (*Summarize Nearby* and *Buffer*) return area in the units requested. But, if you use Arcade functions, Area and Area_ geodetic, you end up with three different estimates of area for each highway buffer.

This analysis was repeated using ArcGIS Desktop – ArcMap and data extracted from the ArcGIS Online web maps. Results from ArcMap using the extracted feature layers projected to EPSG:3587 are the same as the results obtained using Python.



Using ArcMap

- Export/Extract line features and block groups from ArcGIS Online
- Highways in EPSG 4326 (WGS84, decimal degrees)
- Buffer highways – get ArcGIS Online result .
- Project highways to 3857 (WGS84, meters)
- Buffer projected highways - get same result as with Python

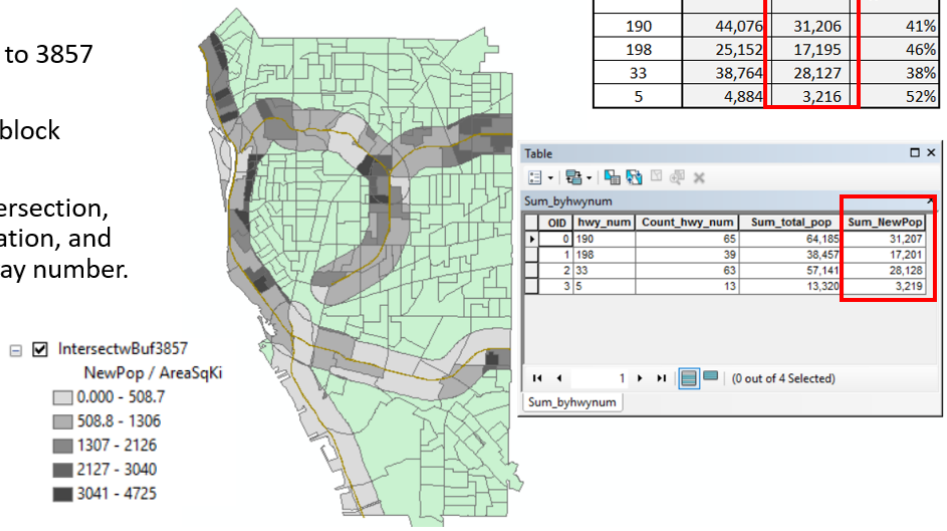


ArcMap Results

- Project block groups to 3857 and calculate area
- Intersect buffer and block groups
- Calculate area of intersection, compute new population, and summarize by highway number.

Total Population

HWY	ArcGIS Online	Python	% Difference
190	44,076	31,206	41%
198	25,152	17,195	46%
33	38,764	28,127	38%
5	4,884	3,216	52%



The "*Summarize Nearby*" tool provides GIS analysis for everyone. However, GIS analysts should know what is inside the Black Box, if merely to understand how the numbers were derived and what they mean. Python provides a very flexible hands-on approach to implementing the workflow – especially in Jupyter Notebook. Programming students with little or no GIS background understand the relationships between spatial and attribute data and spatial processing methods.

This is not a “cookbook” exercise. The entire workflow is addressed. Students consider

- spatial representation and how it effects the approach to generating the desired results
- coordinate systems and units of measurement
- data availability and access
- which methods/tools to use and when to use them

Comparison of the different technical approaches to solving the problem was implemented by the instructor. Consequently, each class accepted their results as accurate. Additionally, they did not consider the pros and cons or limitations of any approach.

Additional References

Nick Bearman, Nick Jones, Isabel André, Herculano Alberto Cachinho & Michael DeMers (2016) The future role of GIS education in creating critical spatial thinkers, Journal of Geography in Higher Education, 40:3, 394-408, DOI: 10.1080/03098265.2016.1144729