Introduction to Vector Analysis
(A Hands-On Tutorial with ArcGIS Pro)

Introduction:

This tutorial is designed to introduce you to vector-based analysis (using points, lines, and polygons) in ArcGIS Pro. For information about how to access ArcGIS Pro software at Dartmouth, check out the following page: https://sites.dartmouth.edu/gis-geography/software/

Before beginning the exercise, familiarize yourself with foundational GIS concepts, including vector vs. raster data models and coordinate systems. We recommend the following learning pathway: https://sites.dartmouth.edu/gis-geography/basics-of-gis-training/

Important reminders for working with GIS software:

- **Use specific naming conventions for GIS**. Make sure that your computer folders, file, and ArcGIS project do **NOT** have spaces or special characters in their names, apart from an underscore. Computers interpret spaces and special characters as a command and will run into mysterious errors during analysis.

- **Your data are not stored in a GIS project**. Geospatial data is stored on your computer's hard drive in file folders. The GIS project only points the computer to where these are located, runs analyses on them, and displays them. Once you start a project, if you change anything about your existing file configuration (move or rename any folders or files), the software will not be able to find them. To fix this, open the dataset's properties and use "Set Data Source" to link ArcGIS back to the correct location.

- **Know where your output files are going and what they are named**. You will produce many, many files while running analysis in a GIS. If you do not specify where these files are saved when you run tools, they will be sent to a default location that you will **NOT** easily find. Similarly, make sure to specify a clear name for output files that you will remember, as default names are **NOT** intuitive.

- **Each GIS dataset is made up of multiple files**. A single GIS data layer (e.g. a shapefile) is not just a single file on your computer, but a collection of 4-8 files that computers interpret together. We recommend only moving, deleting, or renaming these files through a GIS software rather than a file explorer, and zipping up all files together to share.

- **Save and back up your project frequently**. GIS software, while useful, is known for crashing unexpectedly, often at the worst possible time. Similarly, computer issues always strike when least expected. We recommend copying your entire working folder with all associated files onto a USB flashdrive or cloud drive (Google Drive, Dropbox, Onedrive, etc.).

Outline:

In this tutorial, you will explore a case study of a (fictional!) mosquito-borne disease outbreak in Denver, Colorado. The disease, South Platte Virus (SPV), is a hypothetical future variant of West Nile Virus.

On November 1, 2030, the first new cases of South Platte Virus (SPV), a novel variant of West Nile Virus, are reported in Denver, Colorado. Over the following weeks, several hundred SPV cases emerge across the city of Denver, spread by mosquito bites. By the time doctors realize the virus can also be transmitted by blood transfusions and personal contact, the number of cases has exploded, surpassing the capacity of local hospitals.

In this simulated scenario, the City of Denver has hired you to identify locations where temporary field hospitals can be built for the SPV patients. They are looking for sites that meet the following criteria:

- Within walking distance (0.25 miles) of patients who have poor hospital access (more than 5 km away from hospitals)
- In areas with high SPV prevalence (high proportion of population infected)
- Within walking distance (0.25 miles) of major roads

Along the way, you will be introduced to the following essential vector-based methods:

- Geocode addresses
- Calculate Euclidean distance
- Query data by attributes (Select by Attribute)
- Query data by spatial proximity (Select by Location)
- Calculate field attributes
- Spatial buffer
- Spatial overlay (Clip and Intersect)

Prepare your working space

a. Create a new working folder on your computer to hold the files from this tutorial. Remember to name it without spaces or special characters (e.g. *GIS_Tutorial_Vector* instead of *GIS Tutorial – Vector*).

b. Download the zip file labeled “Vector_Analysis_Tutorial_Data.zip”. Unzip the file (right-click its name and select Extract All) into your working folder.

c. Open ArcGIS Pro and create a new project within your working folder. See the “Introduction to GIS Software” tutorial to remember project basics.

d. Add a new folder connection (in the Catalog pane, right-click on Folders → Add Folder Connection). Navigate to your working folder and choose OK. When you expand the arrow next to Folders, you should see the working folder appear (right-click and refresh if not).

e. Within the Vector_Analysis_Tutorial_Data folder, you should see the following datasets provided by the City of Denver's fictional SPV task force:

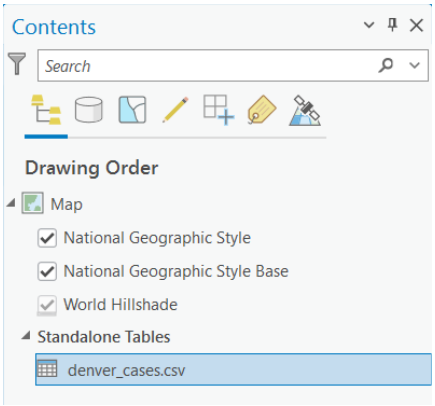
- **denver_cases.csv**: a table with addresses of new patients diagnosed with the fictional South Platte Virus disease at this point in time
- **hospitals_proj.shp**: a shapefile of hospital locations in Denver, projected into the Colorado Central State Plane coordinate system
- **colorado_blockgroups_2020_proj.shp**: a shapefile of Census block groups and population information for the state of Colorado, projected into the Colorado Central State Plane coordinate system
- **denver_city_boundary_proj.shp**: a shapefile of the Denver city boundary, projected into the Colorado Central State Plane coordinate system
- **colorado_major_roads_proj**: a shapefile of major roads in the surrounding Denver area, projected into the Colorado Central State Plane coordinate system

f. Add all five datasets to your ArcGIS project by clicking and dragging them into the Map, or by right-clicking their names and choosing Add to Map.

PART A:

1. Geocoding Addresses

Your disease case data are stored in a non-spatial table with street addresses instead of coordinates, like many other social, economic, and health related datasets at the individual level. Desktop GIS software packages like ArcGIS Pro position objects based on map coordinates – geographic (latitude/longitude) or projected (XY). The process of converting street addresses to map coordinates is referred to as "**geocoding**".

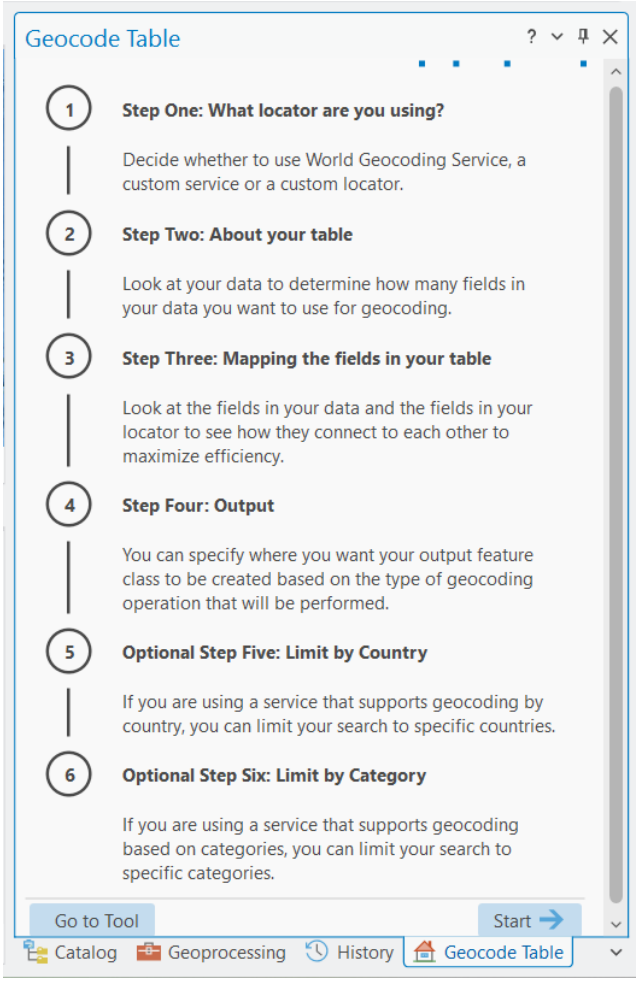


There are many different ways to geocode addresses, depending on the number of records. For more guidance, visit https://sites.dartmouth.edu/gis-geography/geocoding

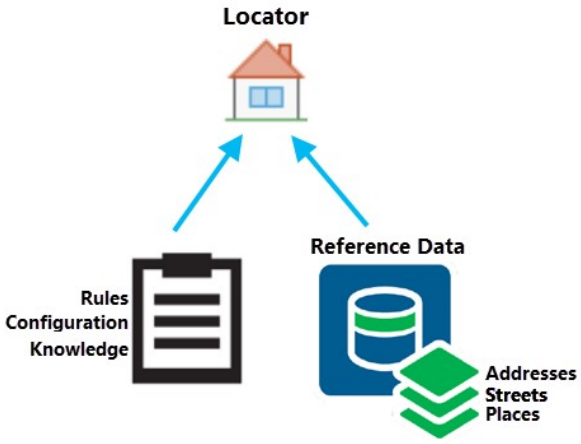
a. At the bottom of your Contents pane, right-click *denver_cases.csv* and select Open. You should see street addresses for 527 (fictitious) patients who presented with symptoms of SPV at a local hospital or clinic.

*** **Very important note** *** Because this is a simulation of a fictitious disease, the addresses for "SPV cases" are entirely arbitrary. There is no implication of risk to particular neighborhoods/homes. If these were real data used in a real public health study, you would need to follow data security practices to protect individuals' private information; e.g., password-protecting data files.

b. Right-click on *denver_cases.csv* and select **Geocode Table** to open the geocoding wizard (screenshot at right). Click on Start.

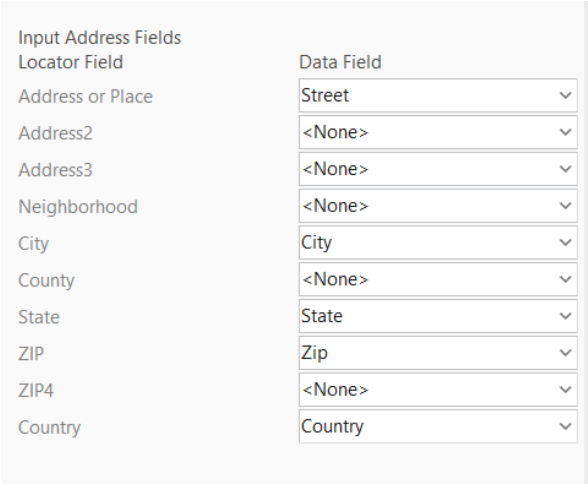


c. For input locator, select ArcGIS World Geocoding Service. This is a tool that compares addresses to reference data in order to generate coordinates. You can use different types of locators or create your own, but we will stick with Esri's cloud-based service for now.



d. On the following page, make sure the input table is *denver_cases.csv*. Examine the table to see how the addresses are structured: is it contained in one field, or spread across multiple fields? Select the dropdown option accordingly.

e. On the next page, check that the correct fields are being used to interpret the address (screenshot at right).



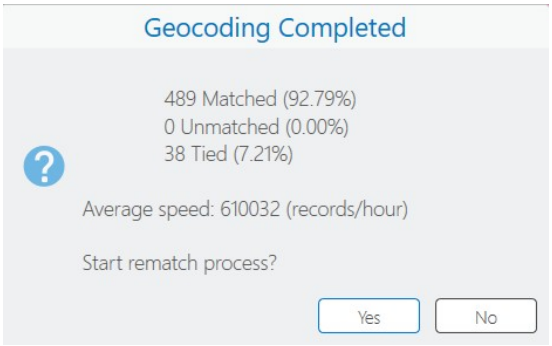
f. Select next, then set the output to a location that you will remember. Give it a name like *denver_case_points* and click on the gray globe next to output to set the coordinate system. In the search bar, type in the EPSG code '6427', then select the projected NAD 1983 (2011) StatePlane Colorado Central system. Leave the location type as Address Location and output fields as All.

g. Skip the following page about country locations, as we have already specified a country field.

h. Similarly, skip the option to limit addresses by category. Select Finish to see all the parameters.

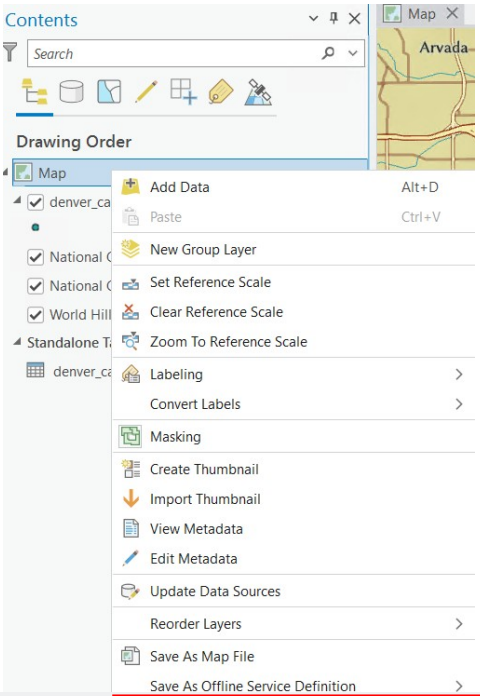
i. At the top of the tool, select 'Click to estimate credits'. Credits are a currency used by ArcGIS Online to run cloud-based tools and data processing. It should be a reasonably small number compared to your total credits available (2000), as we only have 527 addresses. If you were geocoding many more addresses, you may want to consider a different option.

j. Run the tool. It may take several minutes before being added to your map. When finished, you should see a notice of how many addresses were matched. Select No for this, as we are using fictional addresses where the exact locations are not required.



k. You should now see points within the city of Denver, representing individual locations of each patient with SPV.

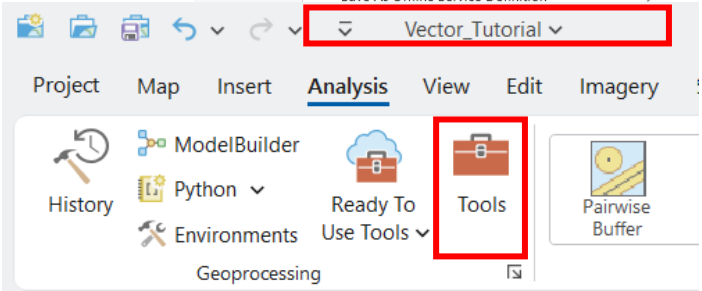
l. In the Contents pane, right-click on Map and select Properties. Under Coordinate Systems, you should see the new points layer showing up in the Colorado Central State Plane system. Make sure your map matches this by clicking once on the system's name and selecting OK.



2. Calculate distances from each SPV case to the nearest hospital or clinic

To assess patient access to hospitals, you will first calculate the average distance from each SPV case to the nearest hospital or clinic. Ideally, these "distances" would actually be *travel times via car or public transportation*. However, that requires methods known as "network analysis" that are beyond the scope of this case study, so for now you will compute the straight-line distance rather than travel time. (See https://sites.dartmouth.edu/gis-geography/network-analysis for guidance on this more advanced analysis).

a. In the upper ribbon, open the Analysis tab and select Tools to open the Geoprocessing pane



at right. In its search bar, search for Near. This will calculate the nearest hospital to each patient.

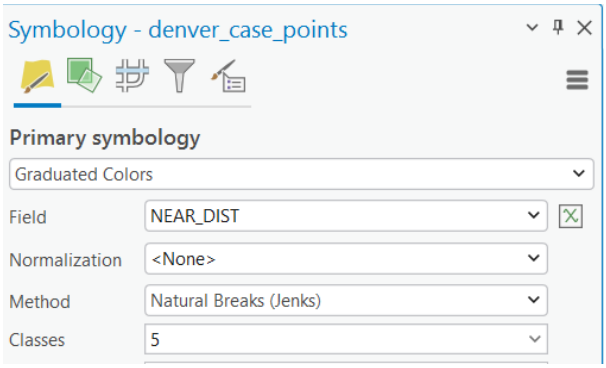
b. Set the input features as your case points and the near features as the hospital points. For the method, select Geodesic, then run the tool.

c. When the process finishes, right-click on the case points layer and select Attribute Table. Scroll all the way to the right to see two new fields:

- NEAR_FID: an ID field that indicates which hospital was identified as closest
- NEAR_DIST: the calculated distance between each patient's address and the hospital

Question: What is the unit for distance here? Why did ArcGIS use that unit?[1]

d. To visualize the points based on distance to the nearest hospital, right-click on the case points name, go to Symbology and change the dropdown from Single Symbol to Graduated Colors. Change the field to be the calculated distance, and select an appropriate color scheme.

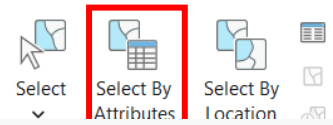


3. Identify low access cases (Select by Attribute)

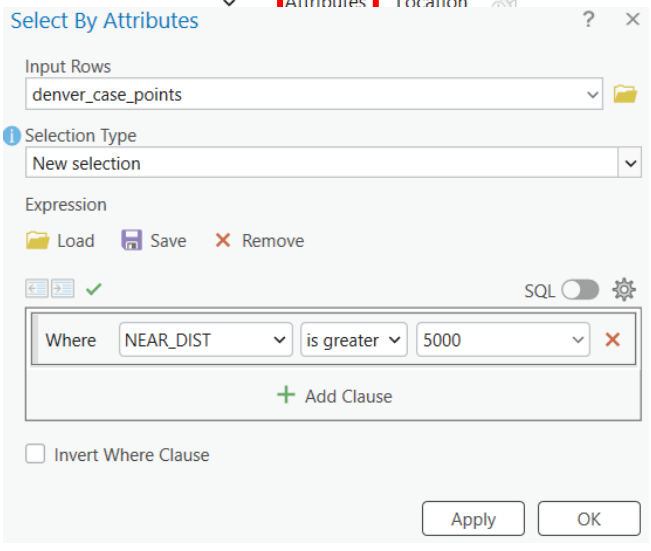
Public health researchers, healthcare workers, and urban planners are often interested in studying access to medical care, including identifying under-served areas to optimally site a new facility. 5 km is generally considered to be the maximum acceptable distance for good hospital access. You will identify patient locations outside of this distance to determine where the city might implement new field hospital sites.

To do this, you will use the Select by Attributes tool, which allows you to query (question) the data's attributes using Structured Query Language (SQL).

a. In the upper ribbon, go to the Map tab and open the Select by Attributes tool. Leave the input rows as the case points.



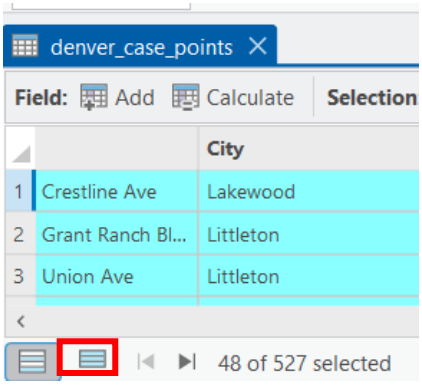
b. Construct a query to find all the patient locations that are greater than or equal to 5000 meters, or 5 km (screenshot at right).



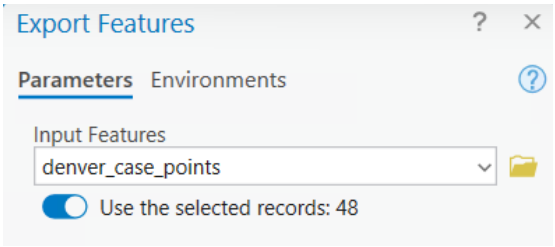
c. Select OK, then look at the attribute table to see how many records were selected. They will be highlighted in teal in both the table and the map,

[1] Hint: check the coordinate system of the case points layer

with a number like 48 out of 527 selected. If you want to view just the selected records, you can click on the teal record button at the bottom.



d. To create a new data layer with just these low access locations, right-click on the case points layer in Contents and go to Data → Export Features. Make sure that "use selected records" is toggled on so that only those that fit our query will be exported. Save it in a reasonable location with a name like *denver_cases_outside5km*.



e. When the new layer appears, check the box next to the total *denver_case_points* layer to hide it cases and make sure that you see only the cases beyond 5 km to hospitals on the map.

4. Identify areas within walking distance of low access patients (Buffer)

To ensure the field hospital sites are accessible for these patients, you will first need to identify areas within their walking distance. Good walkability is generally considered to be 0.25 miles, or 400 meters. You will use the tool Buffer to do this, which generates a zone of access around each feature at some specified distance.

a. In the Geoprocessing pane, search for the tool **Buffer**. Set the input features as your cases outside of 5 km, and save the output features into a reasonable location with a name like *low_access_400mbuffer*. Set the distance as 400 and the linear unit as meters. Leave the method as planar, and set the dissolve type to dissolve all output features into a single feature. Run the tool.

b. When finished, you should see circular buffer zones around each low access case. When buffers overlap, they have been dissolved into a larger polygon.

c. Using the same method, create a 0.25-mile (400 m) buffer around the streets layer.

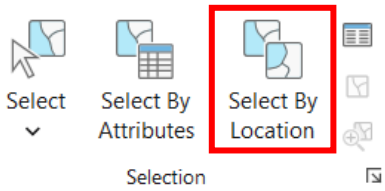
5. **Calculate areas with high disease prevalence**

So far, you've been looking at the raw number of cases. However, variation in the distribution of these cases may be due to variation in population density (where there are more people, we would expect more cases). Instead, you can calculate the prevalence of SPV, or the proportion of the population that is infected (e.g. total number of cases / total population) by first summing up the number of cases within each block group, then dividing by total population in each block group.

You will use Census block groups for your area of analysis, as population and demographic data are available at this geography level from the US Census Bureau. First, you will need to go through the following steps:

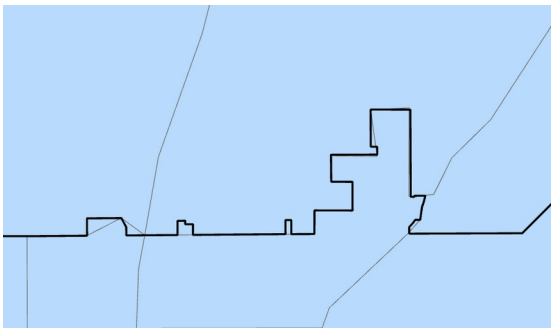
I. *Limit block groups to the city of Denver*

a. Rather than doing analysis with the entire state's block groups, you should narrow it down to Denver. Open the Select by Location tool from the upper ribbon (Map tab). Use the following parameters to run the tool:
   - **Input features**: Colorado block groups
   - **Relationship**: Have their center in
   - **Selecting features**: Denver city boundary



b. In the same way as you did after the Select by Attributes tool (step 3d), export the selected features into a new dataset of only block groups that intersect with Denver's city boundary. Give it a name like *denver_block_groups_2020* and save it to a reasonable location.

c. Remove the Colorado block groups and check your results. Note: we could have done a similar process with the tool Clip, which would cut out the block groups using the city boundary as a cookie cutter. However, the two datasets were digitized with different levels of detail, and clipping would create small sliver polygon errors around the edges.



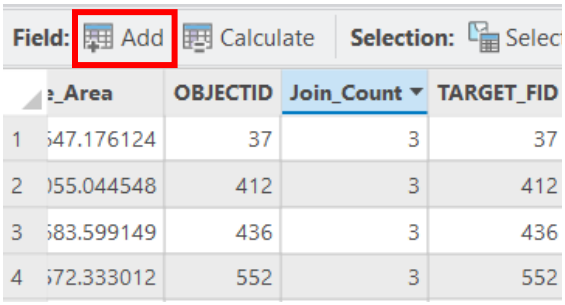
II. *Calculate the number of cases within each block group (Spatial join)*

To calculate disease prevalence, you will first need to join the attribute table for the block groups and case points together. Since there is no equivalent field in the attribute tables, you can use spatial proximity as the linking key instead with a **spatial join.**

a. Right-click on the Denver block groups and select Joins and Relates → Add Spatial Join. Leave the target features as the block groups, which is the host table to which you will join the points. Set the join features as the low access cases, and change the Match Option to Contains. Click OK.

b. When finished, open the attribute table for the Denver block groups and scroll through the table until you find the new fields, which should include "Join_Count" (might be in the middle of the table, or at the far right). Join_Count should tell you how many cases are in each block group. Check the map to make sure this seems reasonable.

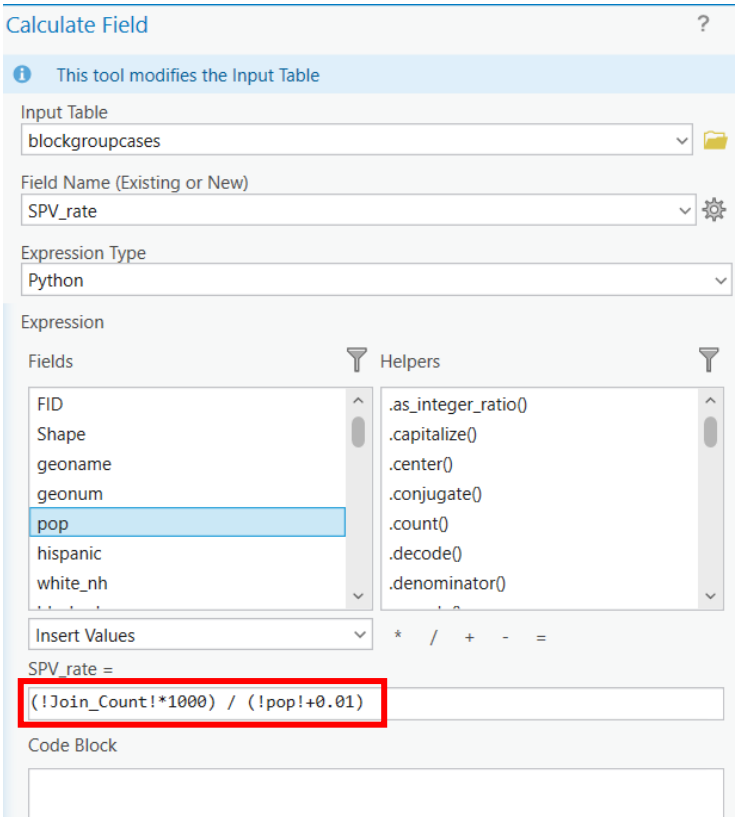
c. Similar to attribute joins, spatial joins are only temporary and will disappear once the program is closed. Export the Denver block group layer into a new dataset (e.g. *block_group_cases*) to make it permanent.



III. *Calculate the disease prevalence*

a. To calculate prevalence for each block group at once, add a new field to the attribute table of *block_group_cases* called something like SPV_rate with the Data Type "Double". Save your edits.

b. Back in the attribute table, scroll all the way to the right to see the new SPV_rate field. Right-click on its name and select Calculate Field to open the Field Calculator.



c. In the formula box, type in the following, making sure to exactly match the spelling and cases of your field names. The "+0.01" is to avoid divide by zero errors, and the "*1000) is to inflate the rate per 1,000 people.

*(!Join_Count!*1000) / (!pop!+0.01)*

Note: you can double click on the layer names in the Fields list to use them in the expression, rather than typing by hand.

d. Click OK to apply the equation, and check the attribute table to make sure the results are reasonable.

IV. *Identify areas with high disease prevalence*

a. Using the Select by Attribute tool, find block groups with a SPV rate higher than 0. Export these into a new dataset called something like *high_prevalence_block_groups*.

6. Overlay and identify potential sites

Now that you have successfully identified areas that meet each specification, you will combine them using a spatial overlay tool to find locations that meet all criteria for a new field hospital site. You will use the Intersect tool for this, which will only preserve overlapping areas between all the different layers.

a. Run the tool Intersect with the following parameters:
   - **Input features**: high prevalence block groups, 400 m buffer around low access patients, 400 m buffer around streets
   - **Output features**: something like *potential_sites*

b. Check the results; you may need to turn off extra layers or change your basemap to satellite imagery (Map tab → Basemap). How many sites did you identify? Do they seem like reasonable locations based on satellite imagery?

In this exercise, you tested various vector-based methods to analyze spatial data of infectious disease cases. Needless to say, this only begins to scratch the surface of how GIS methods can be used in public health. If you want to learn more, consider taking Geography 79, *GIS in Public Health*.

For an extra challenge:

1. What are some other factors the City of Denver might consider for picking new field hospital locations? Try adding at least one additional factor to your analysis.

2. Are there any hotspots of disease cases? Run the Anselin Local Moran's I tool on the disease case data to find significant clusters and outliers. Use the ArcGIS Pro tool page for Local Moran's I to understand the parameters and results.

3. Create a map layout showing the final sites you recommend for new field hospitals, including existing hospital locations, streets, and patient locations. Use the following page for tips on creating map layouts in ArcGIS Pro. https://sites.dartmouth.edu/gis-geography/cartography-and-making-a-simple-map/