02.221 – Lab 3: Spatial queries, joins and geocoding

The data needed for this lab can be found on the course website or downloaded from Dropbox: https://dl.dropboxusercontent.com/u/4837647/02221-Lab3.zip. Extract the data from the zip file to an appropriate location within your documents.

Goals

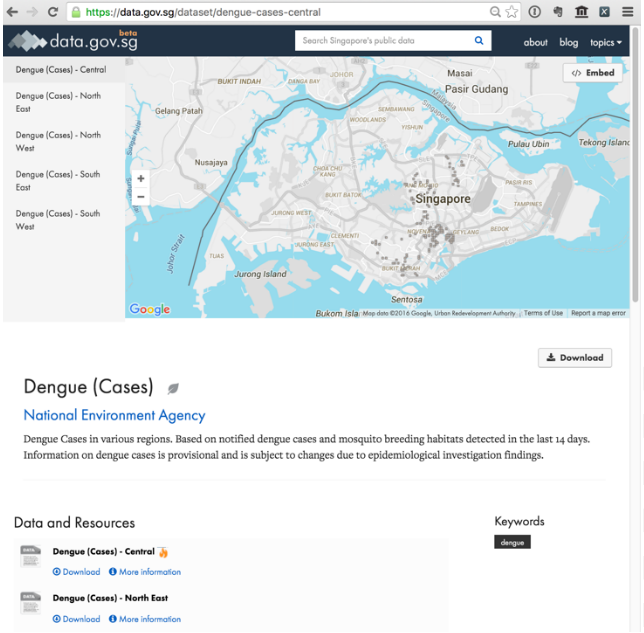
The primary goal of this lab to learn about the ways in which different spatial layers can interact and be combined. We will do so through a case study of recent cases of dengue fever in Singapore.

Spatial Querying

Imagine you are a data journalism intern at the Straits Times. After last year's El Nino year, with higher than usual dengue fever cases and the rise of Zika later in 2016, NEA is reporting that 2017 might see a spike in dengue cases as well. This has you concerned a little bit and you decide to dig deeper into the situation. You wonder whether dengue cases are evenly spread throughout the city and whether, for example, SUTD's campus might be close to any recent cases. You know the government is closely monitoring the situation and even provides some data and maps on sites like dengue.gov.sg and data.gov.sg. However, those are not as easy to read or insightful for the average newsreader. Perhaps you can improve on those and create a visualization to hand to your editor!

To get started, you think it's a good idea to look at some familiar surroundings: how many cases have been reported around the SUTD campus? Start a new project in QGIS and add the school-buffer vector layer. It might help if you add a basemap (Web | Openlayers) for some context. You will see that the school layer has one feature: a circle around the school's campus. Since the Aedes mosquito that spreads dengue does not travel very far (generally ~500 meter), you are only concerned with anything that happens within roughly 1 kilometer from campus. This is what the circle represents – you want to find out how many cases of dengue have been reported within 1 kilometer from campus.

To do so, you will need data on recent dengue cases. Go to data.gov.sg and search for 'dengue cases'. This will lead you to a spatial file that represents recent (14 days prior to reporting) dengue cases, which the government has kindly made available to the public. The file looks split over several regions, but you can download any one of them: they actually contain all the regions within a single zip file. Download the file and extract the zip archive.



https://data.gov.sg/dataset/dengue-cases-central

- Dengue (Cases) - Central
- Dengue (Cases) - North East
- Dengue (Cases) - North West
- Dengue (Cases) - South East
- Dengue (Cases) - South West

Map data ©2016 Google, Urban Redevelopment Authority | Terms of Use | Report a map error

Dengue (Cases)

National Environment Agency

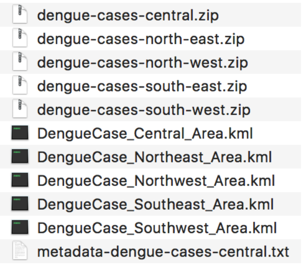
Dengue Cases in various regions. Based on notified dengue cases and mosquito breeding habitats detected in the last 14 days. Information on dengue cases is provisional and is subject to changes due to epidemiological investigation findings.

Data and Resources

Dengue (Cases) - Central
Download · More information

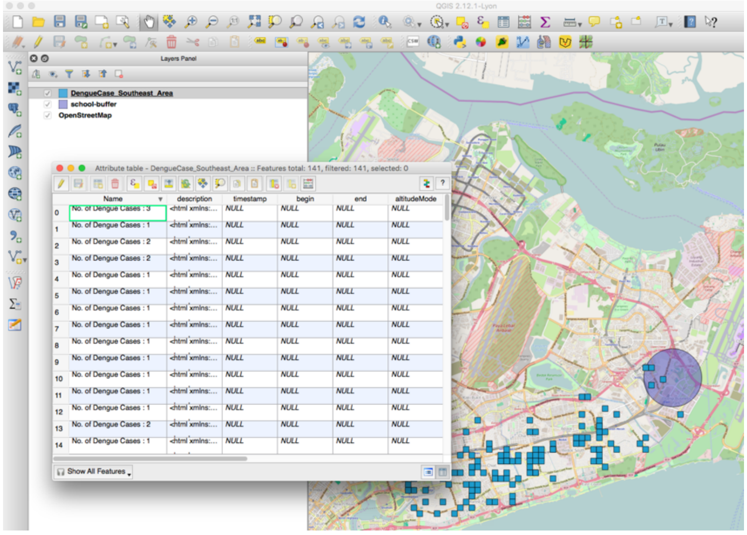
Dengue (Cases) - North East
Download · More information

Keywords: dengue



dengue-cases-central.zip
dengue-cases-north-east.zip
dengue-cases-north-west.zip
dengue-cases-south-east.zip
dengue-cases-south-west.zip
DengueCase_Central_Area.kml
DengueCase_Northeast_Area.kml
DengueCase_Northwest_Area.kml
DengueCase_Southeast_Area.kml
DengueCase_Southwest_Area.kml
metadata-dengue-cases-central.txt

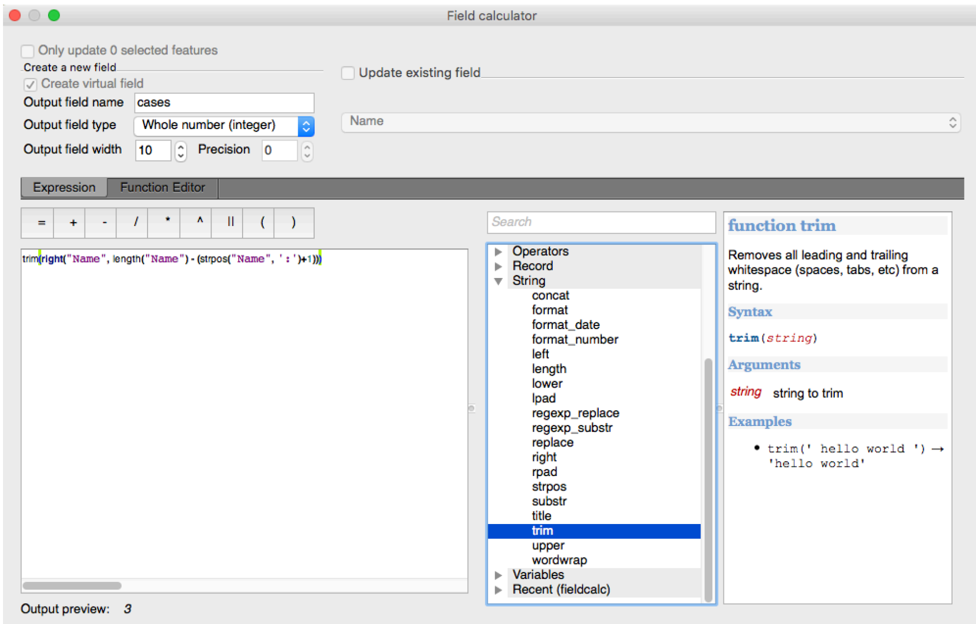
You will see each region is stored as a separate .kml file. For now, only add the South East file to your map in QGIS. Open the attribute table for the new layer to inspect what variables are available. Here we see an immediate downside of many public data sets, especially KML files. The variable of interest (# of cases) is hidden inside the 'Name'. Before we can count the number of cases within our circle, we need to extract and store the data in an appropriate format (it is a ratio variable and should be stored as integer).



Attribute table - DengueCase_Southeast_Area :: Features total: 141, filtered: 141, selected: 0

	Name	description	timestamp	begin	end	altitudeMode
0	No. of Dengue Cases : 3	<html xmlns:...	NULL	NULL	NULL	NULL
1	No. of Dengue Cases : 1	<html xmlns:...	NULL	NULL	NULL	NULL
2	No. of Dengue Cases : 2	<html xmlns:...	NULL	NULL	NULL	NULL
3	No. of Dengue Cases : 2	<html xmlns:...	NULL	NULL	NULL	NULL
4	No. of Dengue Cases : 1	<html xmlns:...	NULL	NULL	NULL	NULL
5	No. of Dengue Cases : 1	<html xmlns:...	NULL	NULL	NULL	NULL
6	No. of Dengue Cases : 1	<html xmlns:...	NULL	NULL	NULL	NULL
7	No. of Dengue Cases : 1	<html xmlns:...	NULL	NULL	NULL	NULL
8	No. of Dengue Cases : 1	<html xmlns:...	NULL	NULL	NULL	NULL
9	No. of Dengue Cases : 1	<html xmlns:...	NULL	NULL	NULL	NULL
10	No. of Dengue Cases : 1	<html xmlns:...	NULL	NULL	NULL	NULL
11	No. of Dengue Cases : 1	<html xmlns:...	NULL	NULL	NULL	NULL
12	No. of Dengue Cases : 1	<html xmlns:...	NULL	NULL	NULL	NULL
13	No. of Dengue Cases : 2	<html xmlns:...	NULL	NULL	NULL	NULL
14	No. of Dengue Cases : 1	<html xmlns:...	NULL	NULL	NULL	NULL

Use the Field Calculator to create a new field. We now need to extract only the actual number of cases from the 'Name' field. You might be tempted to just use the last character of string: ‘right("Name", 1)’. However, that wouldn’t work if the number of cases was more than 9. Therefore, it would be better to split the string at the colon (‘:’) and then use the part that comes after it. There are multiple ways to do so. This is one: ‘trim(right("Name", length("Name") - (strpos("Name", ':')+1)))’.



□ Only update 0 selected features

Create a new field

☑ Create virtual field

Output field name: cases

Output field type: Whole number (integer)

Output field width: 10 Precision: 0

□ Update existing field

Name

```
trim(right("Name", length("Name") - (strpos("Name", ':')+1)))
```

function trim

Removes all leading and trailing whitespace (spaces, tabs, etc) from a string.

Syntax

`trim(string)`

Arguments

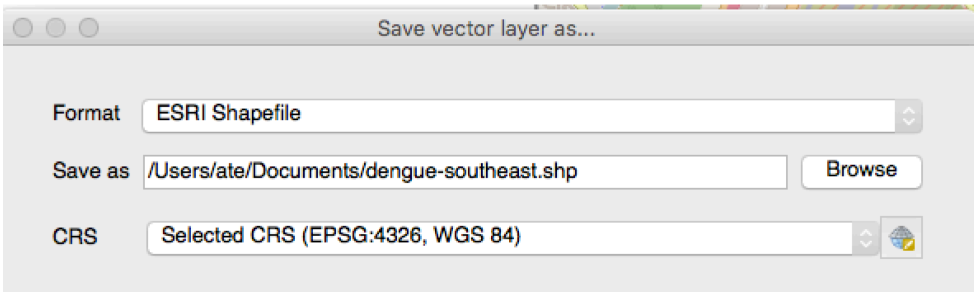
string string to trim

Examples

- `trim(' hello world ') → 'hello world'`

Output preview: 3

Use the help function available inside the field calculator to figure out what each function in the expression does (trim; right; length; strpos). *Write your answer down.* Give an appropriate name and type to your new field and hit ‘OK’. One final step: KML is not a great format for storing additional variables. This is why you need to save the Southeast layer as a shapefile. Do so (right-click on the layer) and add the resulting layer to your map. Remove the ‘old’ KML version.

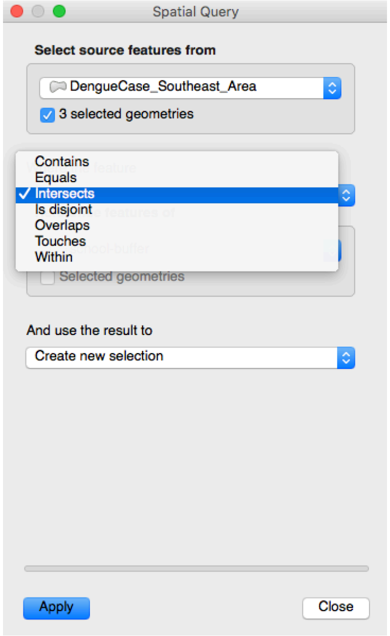


Format: ESRI Shapefile

Save as: /Users/ate/Documents/dengue-southeast.shp

CRS: Selected CRS (EPSG:4326, WGS 84)

Finally, all our data is in place. To figure out which cases are inside of our circle, we need to execute a ‘spatial query’. To do so, download the Spatial Query plugin and use it to select dengue case that *intersect* with the circle. What do you think happens when you choose *within* instead?



Select source features from
DengueCase_Southeast_Area
☑ 3 selected geometries
☐ Selected geometries
And use the result to
Create new selection

Go to Vector | Analysis Tools | Basic Statistics. You can use this tool to calculate the sum of the cases for the selected cells close to our campus. *Write down the answer, as well as the total number of cases in the Southeast file.*

N.B. Naturally, you could have counted the number of cases within the circle by hand. Think about how you can extend the procedure above to other applications where counting by hand is not so convenient. You can find all cases within 1km for nursing homes; from schools; or find water bodies close to dengue clusters to aid in finding sources of infection.

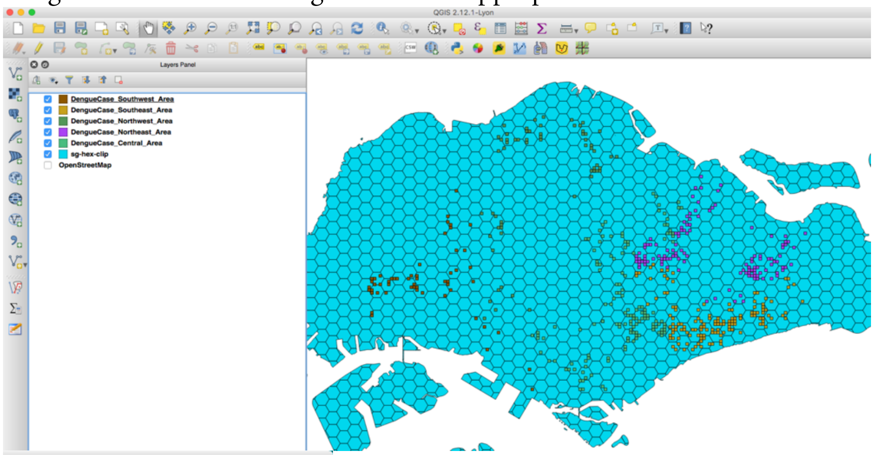
Spatial Joining

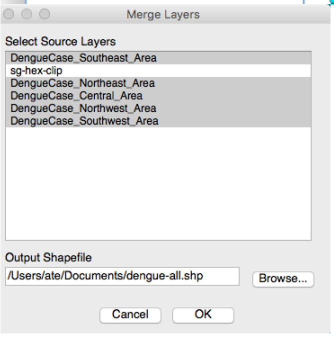
After you satisfied your initial concerns about dengue cases within SUTD's immediate vicinity, you want to explore whether specific clusters exist elsewhere in Singapore. Because the small cells in the original government data do not lend themselves to making thematic maps, your boss has provided you with a hexagonal grid that you can use to visualize data. These cells are bigger and thus much easier to read in a small, printed map of Singapore. Add the sg-hex-clip file to your map. Now you need to add the dengue cases in all the regions to your map and somehow count the number of cases in each hexagonal cells. To do so you need to do the following steps:

1. Merge all regional files together
2. Extract number of cases in separate integer field
3. Spatially join individual cases to hexagons and calculate the total number of cases per cell

[1] Before we start, we need to one more plugin for the next section. Install the MMQGIS plugin. To merge all the regional files together, first add all of the kml

files to the map. Now go to MMQGIS | Combine | Merge Layers, select the layers to merge and save the resulting file with an appropriate name.


QGIS 2.12.1-Lyon
Layers Panel
☑ DengueCase_Southwest_Area
☑ DengueCase_Southeast_Area
☑ DengueCase_Northwest_Area
☑ DengueCase_Northeast_Area
☑ DengueCase_Central_Area
☑ sg-hex-clip
□ OpenStreetMap


Merge Layers

Select Source Layers

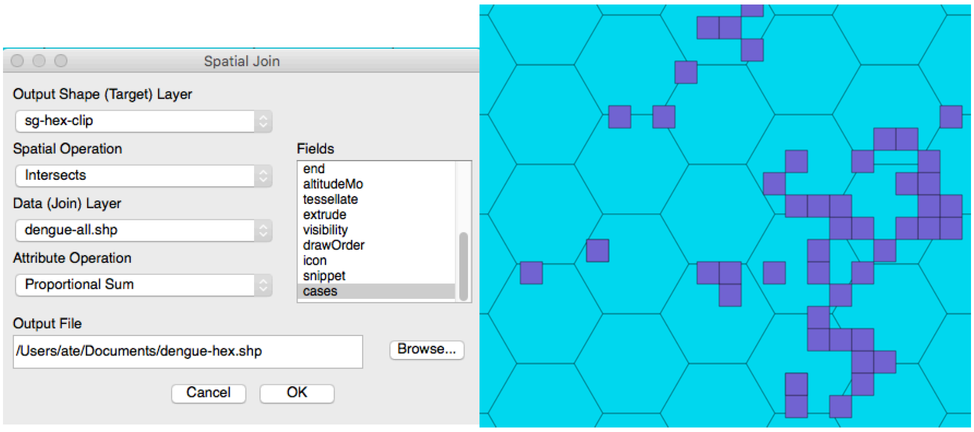
DengueCase_Southeast_Area
sg-hex-clip
DengueCase_Northeast_Area
DengueCase_Central_Area
DengueCase_Northwest_Area
DengueCase_Southwest_Area

Output Shapefile

/Users/ate/Documents/dengue-all.shp

[2] You already know how the extract the number of cases from the Name field and store the results in a new field. Repeat the procedure from the start of the lab to do so.

[3] The last step is the spatial join. In a spatial join, data from one spatial layer is joined to another spatial layer. This is not done based on a common variable or attribute, but based on a *common location*. Go to MMQGIS | Combine | Spatial Join to start the process. Since the dengue cases are represented as small squares instead of points, if we join based purely on an 'intersect', a single case might be counted in multiple hexagons. This would effectively create additional phantom cases and skew the results and we cannot allow that to happen. Luckily, MMQGIS allows us to calculate a 'proportional sum'. In this way, a case is added to a hexagon based on the relative overlap. This is not a 100% ideal as this will create fractional dengue cases but for now this is OK. Select the right join and target layers, the correct attribute calculation and the field to calculate this attribute for. Then hit 'OK'.



Spatial Join

Output Shape (Target) Layer: sg-hex-clip

Spatial Operation: Intersects

Data (Join) Layer: dengue-all.shp

Attribute Operation: Proportional Sum

Fields: end, altitudeMo, tessellate, extrude, visibility, drawOrder, icon, snippet, cases

Output File: /Users/ate/Documents/dengue-hex.shp

Almost there! Check the attribute table to make sure the join, including fractional cases, was successful. You can now use the Style properties of your new layer to create a graduated color map, just like you did in the previous map. Your editor is interested in including a reference to a few places in the final article. If you had to tell your editor what the main clusters of current dengue cases are, what would they be? Additionally, how many total dengue cases have there been in this period? *Write your answer down.*

You will quickly notice that there are many hexagons without any cases. They currently have a 'NULL' value and they may throw off the final visualization. Remember, we are dealing with a ratio variable. In this case, NULL means zero and thus means there are no dengue cases. There are several strategies available to solve this problem. Use Stackoverflow to figure out how to execute one of them. For example, you could replace all null values with zero ('qgis replace null zero').

Finish and polish your map. Remove the border around each hexagon, choose an appropriate color scheme (e.g. from Colorbrewer) and class breaks (choose 'jenks' for now). You might also want to include some geographic reference. You can do this by including a reference map similar to what you made in Lab 1, or for a more minimalist approach, use the planning area boundaries from Lab 2. Finally, use the print composer to add a legend and a map title and *save the map as png to include in your assignment.*

Save the entire project for later use. In the next lab, we will use it again and learn how to polish the output from QGIS in Adobe Illustrator to make your visualization more refined and ready for inclusion in, for example, a newspaper.

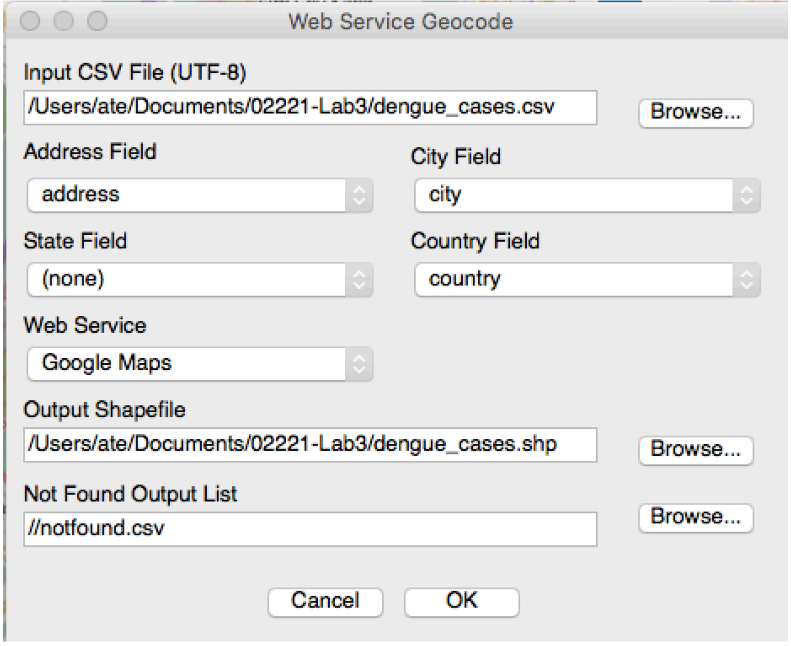
Geocoding

The government data source we used in the previous sections was last updated more than a week ago. Your editor has asked you to look into any more recent developments. A contact at the NEA has provided you with a list of cases that were reported in the last week but have not been included in the public statistics yet.

Unfortunately, the list is provided as a csv file without any explicit spatial attributes. Open the dengue-cases.csv file in Excel and check out the attributes.

The file does contain some kind of geographic information: the address of each case. You could manually look up each address in Google Maps and create a new spatial layer to enter each point by hand. Lucky for you, QGIS provides a way to do this in a more automated fashion but the principle remains the same. It is called *geocoding*: using implicit spatial information (like an address, or a city) to tie features to a more specific, explicit, spatial location.

In QGIS go to MMQGIS | Geocode | Geocode CSV. This tool allows you to read a csv and use the fields within it to geocode your dengue cases. It is pretty smart so should find the right fields by default.



Web Service Geocode

Input CSV File (UTF-8): /Users/ate/Documents/02221-Lab3/dengue_cases.csv

Address Field: address

City Field: city

State Field: (none)

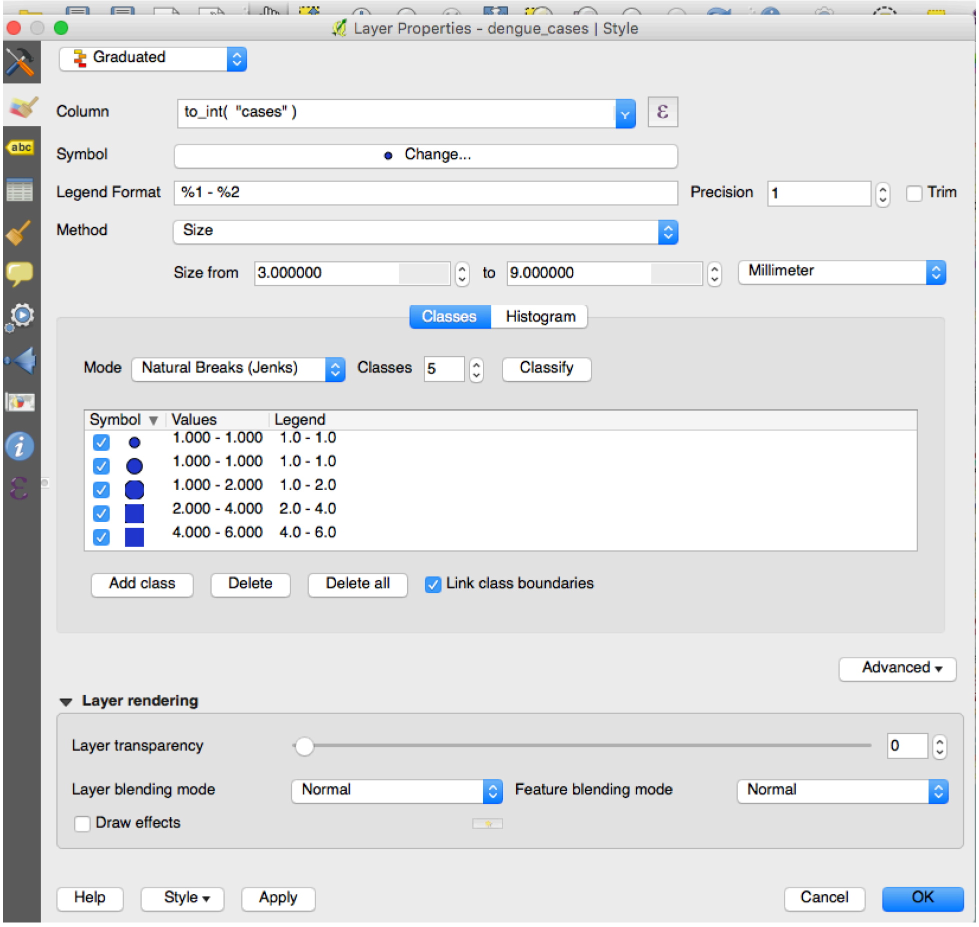
Country Field: country

Web Service: Google Maps

Output Shapefile: /Users/ate/Documents/02221-Lab3/dengue_cases.shp

Not Found Output List: //notfound.csv

Once geocoded, you will have a new point layer with the 15 addresses. You can, for example, create a proportional point map now using the number of cases in each location:



Layer Properties - dengue_cases | Style

Graduated

Column: to_int("cases")

Symbol: ● Change...

Legend Format: %1 - %2 Precision: 1 ☐ Trim

Method: Size

Size from 3.000000 to 9.000000 Millimeter

Mode: Natural Breaks (Jenks) Classes: 5

Symbol	Values	Legend
☑ ●	1.000 - 1.000	1.0 - 1.0
☑ ●	1.000 - 1.000	1.0 - 1.0
☑ ●	1.000 - 2.000	1.0 - 2.0
☑ ■	2.000 - 4.000	2.0 - 4.0
☑ ■	4.000 - 6.000	4.0 - 6.0

☑ Link class boundaries

Layer rendering

Layer transparency: 0

Layer blending mode: Normal Feature blending mode: Normal

☐ Draw effects

(N.B. "cases" is a string field so it needs to be converted to integer before visualizing).

Try zooming in and find the location of the apparent increase in recent dengue cases. There could be reasons in favor or against including a map like this in a newspaper article and sometimes you might use the map in your research but only publish certain insights gained from it and not the actual map. For example, think about how both the data source and the spatial precision of the data differ from the data we used to create the hexagon map. *Write down 1-2 paragraphs* on your decision whether to send the map on to your editor and, if you choose not to send it, how you can perhaps use information gained from it within a news article.

Assignment

On the class website, you will find the assignment for this lab. It consists of the questions above where you were asked to write down your answer and the thematic map of dengue cases.

The assignment needs to be submitted as Word or PDF file. Please make sure you submit the assignment by **February 15.**