

Getting and tweaking TIGER files in QGIS

The purpose of this exercise is to get the geographic data for your study area from the U.S. census. NOTE: for some jurisdictions, like the City of San Francisco, you can get a lot of ready-made GIS data. However we are going to learn how to acquire “raw” data from the US census, in a method that can be applied for any jurisdiction across the U.S.: from Winemucca to Kalamazoo.

1. Get the TIGER shapefile for your geographic area.

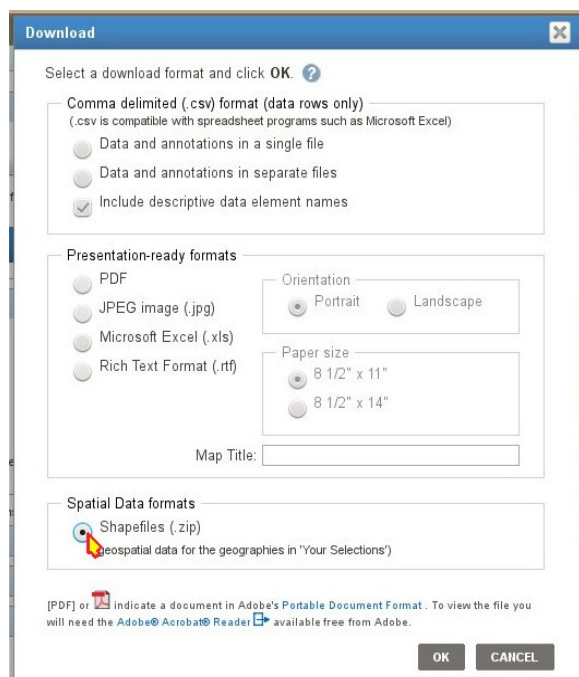
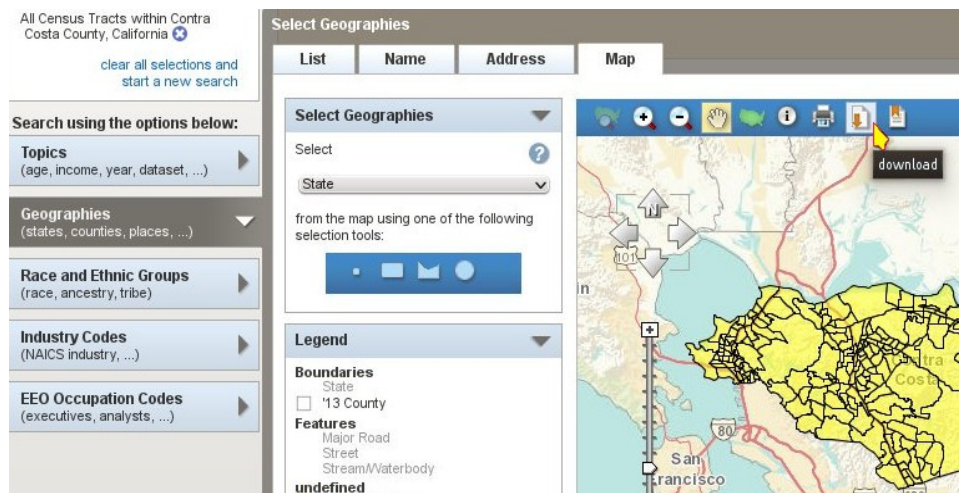
Rather than do a separate search down through the generic TIGER database, I get the TIGER map straight from American Factfinder2. The advantage of this method is that you have already set the FactFinder to seek data from a specific geographic level. For this course I have been using Census Tracts as the geographic unit, but I checked and this also works for Oakland Unified School district.

In the screenshot below, I show how to get the Census-Tract TIGER shapefile for the County of San Francisco. Use the same method for any county: from **Geographies** stay on the **List** tab; from **Geographic Type** select Census Tract – 140 (or whichever geographic unit you are using; could be ZIP codes, could be school districts); then choose **State** and **County** and set this as your the filter under **Your Selections** on the left-hand pane.

Then, click the **Map** tab in the “Select Geographies” window:

The screenshot displays the American Factfinder2 interface for selecting geographic areas. On the left, the 'Your Selections' pane shows 'Census Tract' selected under 'Search using...'. Below this, there are sections for 'Search using the options below:' including 'Topics', 'Geographies', 'Race and Ethnic Groups', 'Industry Codes', and 'EEO Occupation Codes'. The 'Geographies' section is expanded, showing 'Census Tract - 140' selected. The 'Select Geographies' pane on the right has the 'Map' tab selected. It shows filters for 'most requested geographic types', 'Census Tract - 140', 'California', and 'Contra Costa'. A list of results is shown at the bottom, including 'All Census Tracts within Contra Costa County, California'.

This will generate a preview which should look like the geography you are working with. And it should look projected. It is projected! Rather than a “Geographic Coordinate System” (GCS), shapefiles accessed through FactFinder are projected in the “Web Mercator” projection. This is the same projection used by Google Earth and smart-phone mapping applications. The datum is also different: older TIGER files are NAD83; this one is WGS84. Later in this tutorial we will deal with projections and datums, because your shapefiles need to be in the same datum/projection to do operations like Erase and Clip.



If it looks like the data you seek, then click the download button (notice my obnoxiously-recolored cursor). That invokes the **Download** dialog.

When you download it, remember to save the file with a name, and at a path that you can remember. Unzip the file, and it will generate a folder called “reference_map_shape” with a shapefile called “140_00”. This shapefile name is the geographic code-number for census tract data.

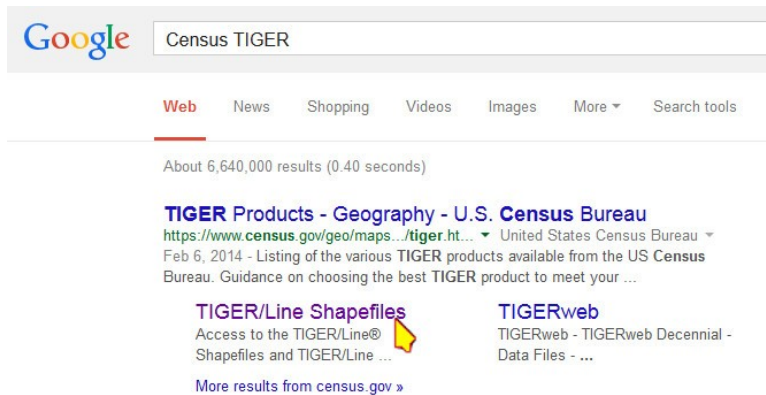
2. Get the “areawater” file for trimming the Census Tracts.

At first I thought this was an optional step. No! It is mandatory. Here is why: many types of geospatial analysis depend upon the area of the units being analyzed, such as population per square mile, dwelling units per acre, and crime or pollution within a given area. A generalized way of thinking about all these kinds of data is that they are *densities*. In data-analysis terms, this is a way of *normalizing your data*. A huge census-tract with a large raw number of units may not actually be very dense. And density or sparseness is what we actually experience on the ground, not the boundaries of census-tracts.

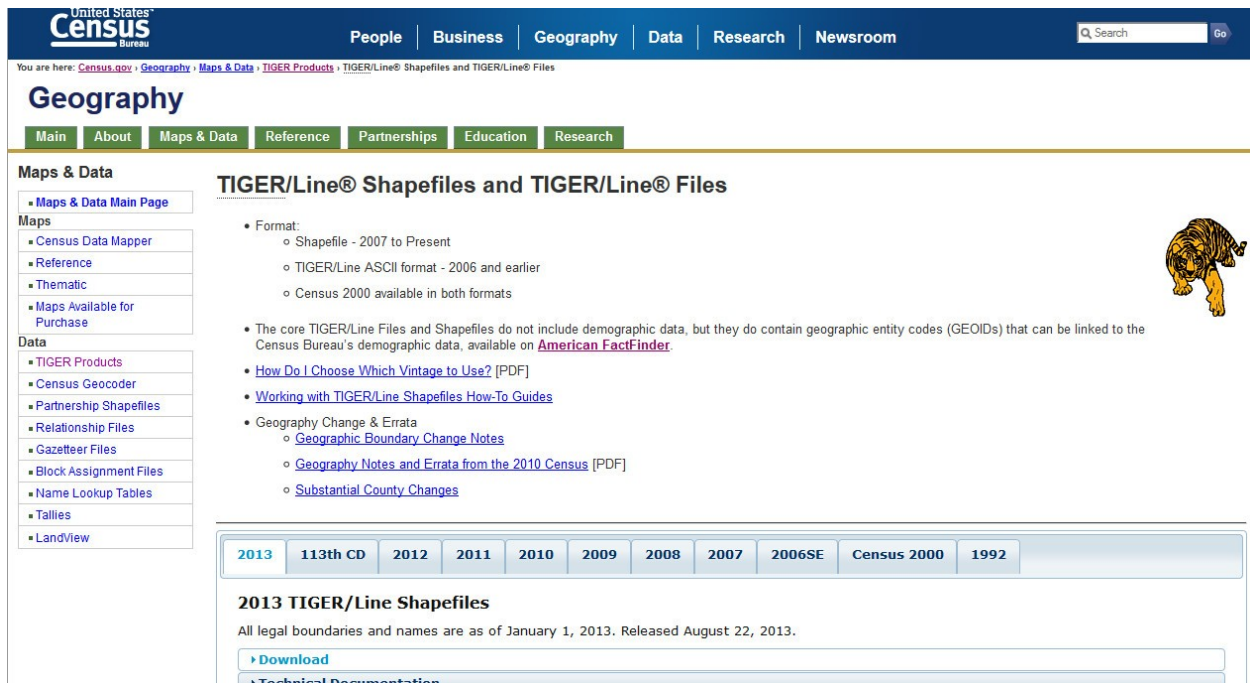
TIGER files include area-calculations in their attribute data. One reason for this is that the shapefile format is not ‘smart’, and it does not automatically track area. In fact you have to project your shapefile in order to get accurate area calculations. The new GIS format that is superseding the shapefile format is the *geodatabase* format. It is a ‘smarter’ format in which polygon-areas are tracked dynamically. However, that does not get

around the fact that a census tract which includes large expanses of open water will throw off the density-ratios of your analysis. So let us trim (Erase) the water and reproject our TIGER file right away.

Here is the method for Contra Costa County:



Yep, back to Google, thence directly into the TIGER database. You may also be able to do this through FactFinder2, but I have not researched that option.



Here you can pick the most recent data, or time-appropriate data if you know that the water-area has changed.



You can get Tract and Block-group polygons through this portal, but I am going after water.

2013 TIGER/Line® Shapefiles: Water

Return to: [Main Download Page](#) | [2013 TIGER/Line Shapefiles Main](#)

Linear Hydrography

--Select a state--

Area Hydrography

California

Linear hydrography (streams and rivers) would be good for buffer-analysis of environmentally sensitive areas, but I am going after area-polygons (the lower option) to trim back my tract-polygons.

2012 TIGER/Line® Shapefiles: Water

Return to: [Main Download Page](#) | [2012 TIGER/Line Shapefiles Main](#)

Contra Costa County

And there we have it: download, and unzip it. We will use this file to trim back the Census-Tract file.

However, in order to use the trimming function in QGIS (or ArcMAP), first the two shapefiles must have the same datum/geoid and be in the same projection. So next, I will show you how to reproject shapefiles.

3. Reproject the TIGER shapefiles.

During class I have been waffling back and forth between various datums(/geoids) and projections. Now I have made up my mind, based on changes in data available worldwide and from the U.S. Census:

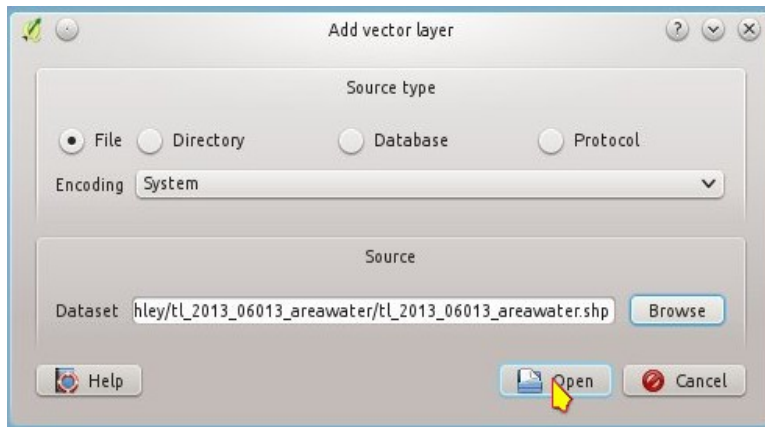
1) When possible, shift the **datum** to **WGS84**. You may get older files in NAD83 or even NAD27. Shift'em all to WGS84. [Reminder: the datum is the mathematical formula for the potato-shape of the earth. Several different formulas have been used over the past century, from North American Datum to the World Geodetic Survey.]

2) For urban files, **reproject** them to Universal Transverse Mercator (**UTM**). [Web Mercator/Pseudo Mercator is also OK; that is the projection used by Google Earth and smart-phone apps. But I think it may be time to retire the State Plane projection.]

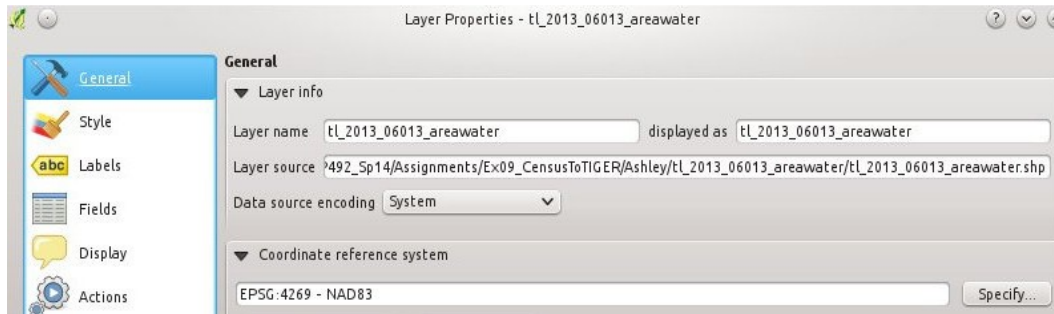
It is best to work with a series of shapefiles that are all in the same projection. Not just so that area-calculations are accurate, but also because some geoprocessing functions will not work if they are in different datums/projections.

Start QGIS. By default it opens a blank data set.

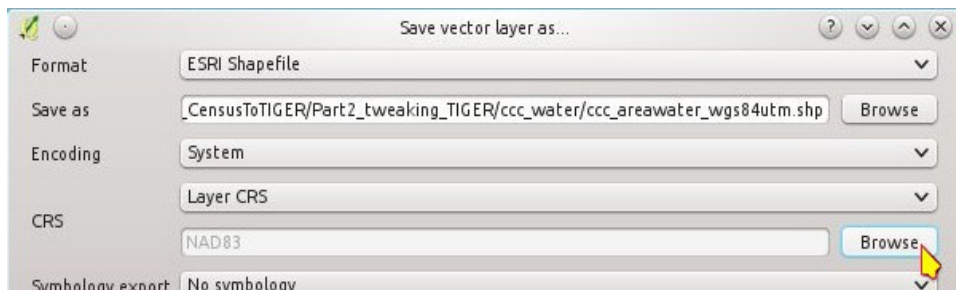
Click LAYER > ADD VECTOR LAYER... and then Browse.. the Path to your TIGER file, and open it.



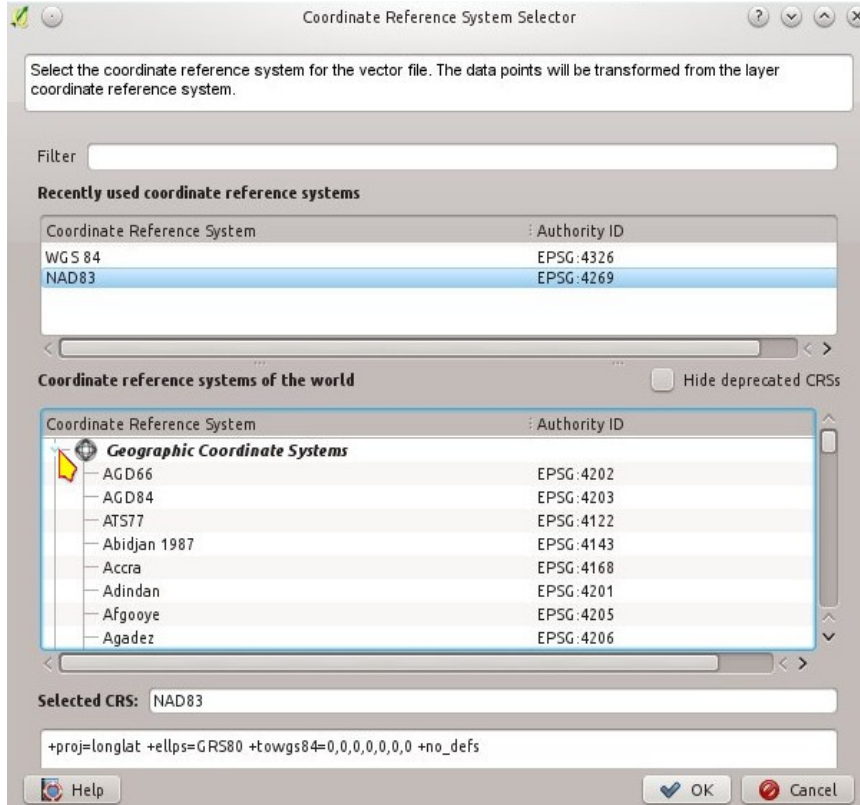
Double-click on the layer to examine its properties. In this case I see that the CRS is NAD83, and no projection is listed; so it is “Longitude X Latitude” or “Geographic”.



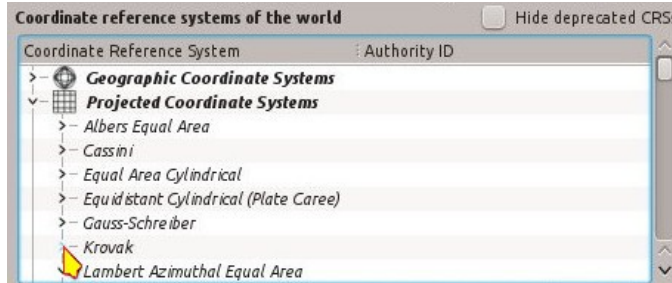
To reproject in QGIS, right-click the layer to reveal the drop-down menu of options, and choose “Save As...”



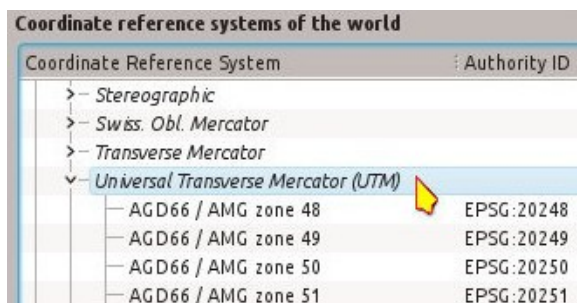
In the **Save as** field I have already specified a name ending in `_wgs84utm.shp`, which will remind me later that this is the reprojected shapefile. Next, I will **Browse** for a new projection, a.k.a. a new Coordinate Reference System (CRS):



In QGIS the projection-selector is a bit funky: all the folders are open and dropped down into a list that is hundreds of names long. Begin your search for the WGS84/UTM (Zone 10 North) projection by closing the “Geographic Coordinate Systems” folder. Later, when you know the EPSG code for your projection, you can simply type it in the search field and it will pop up.

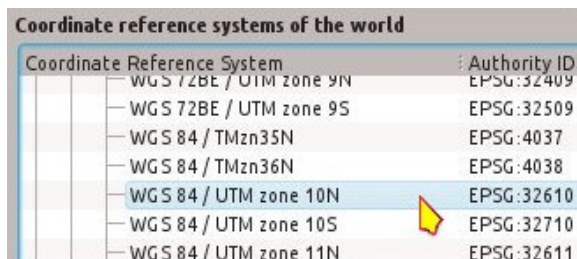


Keep closing projection sub-folders so you will not have to scroll down past all of their options.



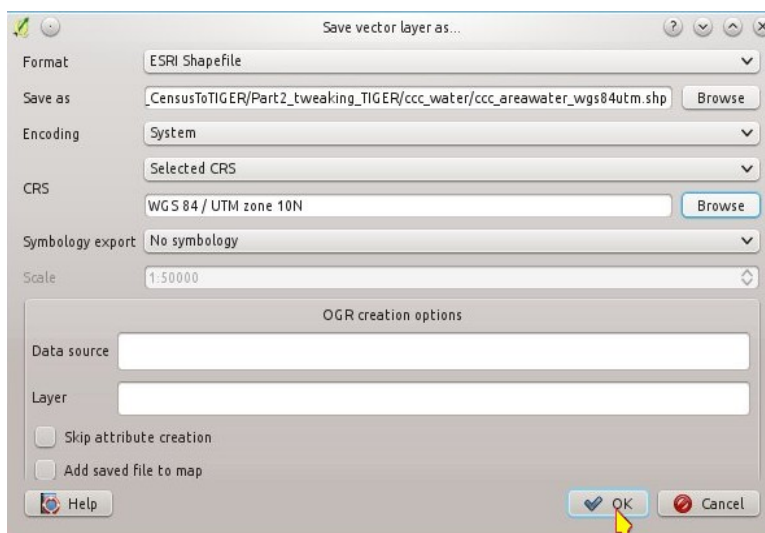
UTM will be near the bottom, because the list is sorted in alphabetical order.

Now we need to find the correct datum (WGS84) and the correct UTM zone, which is 10-North for northern/western California. Los Angeles and San Diego are in Zone 11N; New York is Zone 15N; and Kabul is Zone 42N.



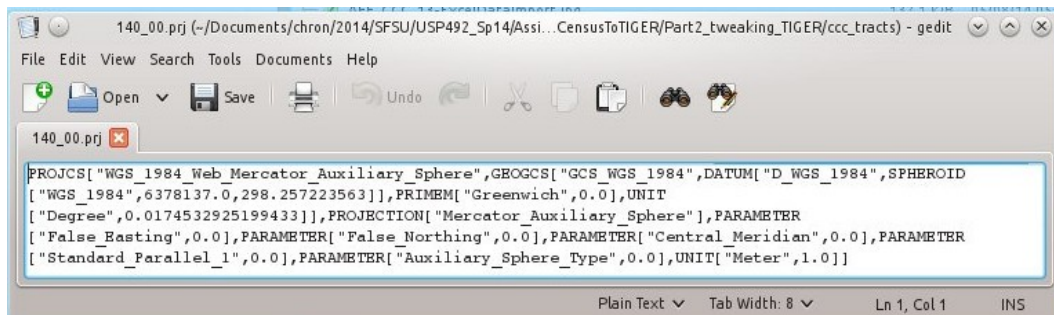
Furthermore, WGS84 will be near the bottom of the UTM projections as well, again because the list is alphabetical.

Now you can see that its EPSG code is 32610, so you can look it up much more quickly in the future. (EPSG stands for European Petroleum Survey Group)



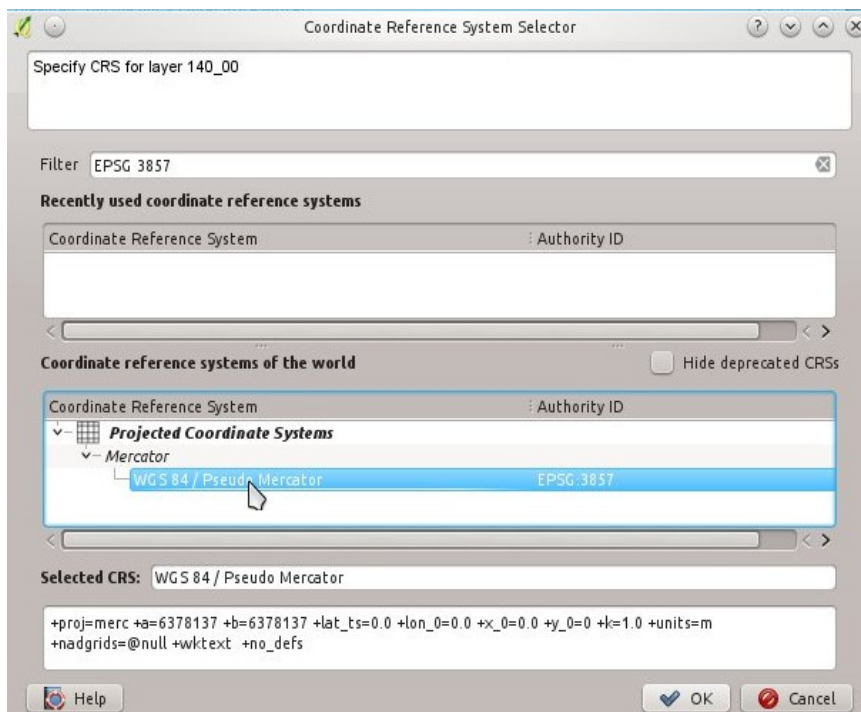
When you set the specs and click OK, QGIS will reproject the spatial data of your shapefile but keep the relationship between the .shp polygons and the .dbf attributes. It will offer to load the reprojected file into the working project. If it does, you may be surprised that it looks just like the original file. That is because QGIS will show the file in the current workspace coordinate reference system. If you quit and restart QGIS as a blank document, and then load the reprojected file *first*, you will see it displayed with its own projection.

QGIS also gives you some output as it generates the reprojected file. Here you should look for warning messages; it may explain what went wrong. It also tells you that the reprojection’s scale units will be meters.



```
PROJCS["WGS 1984 Web Mercator Auxiliary Sphere",GEOGCS["GCS_WGS_1984",DATUM["D_WGS_1984",SPHEROID["WGS_1984",6378137.0,298.257223563]],PRIMEM["Greenwich",0.0],UNIT["Degree",0.0174532925199433]],PROJECTION["Mercator_Auxiliary_Sphere"],PARAMETER["False_Easting",0.0],PARAMETER["False_Northing",0.0],PARAMETER["Central_Meridian",0.0],PARAMETER["Standard_Parallel_1",0.0],PARAMETER["Auxiliary_Sphere_Type",0.0],UNIT["Meter",1.0]]
```

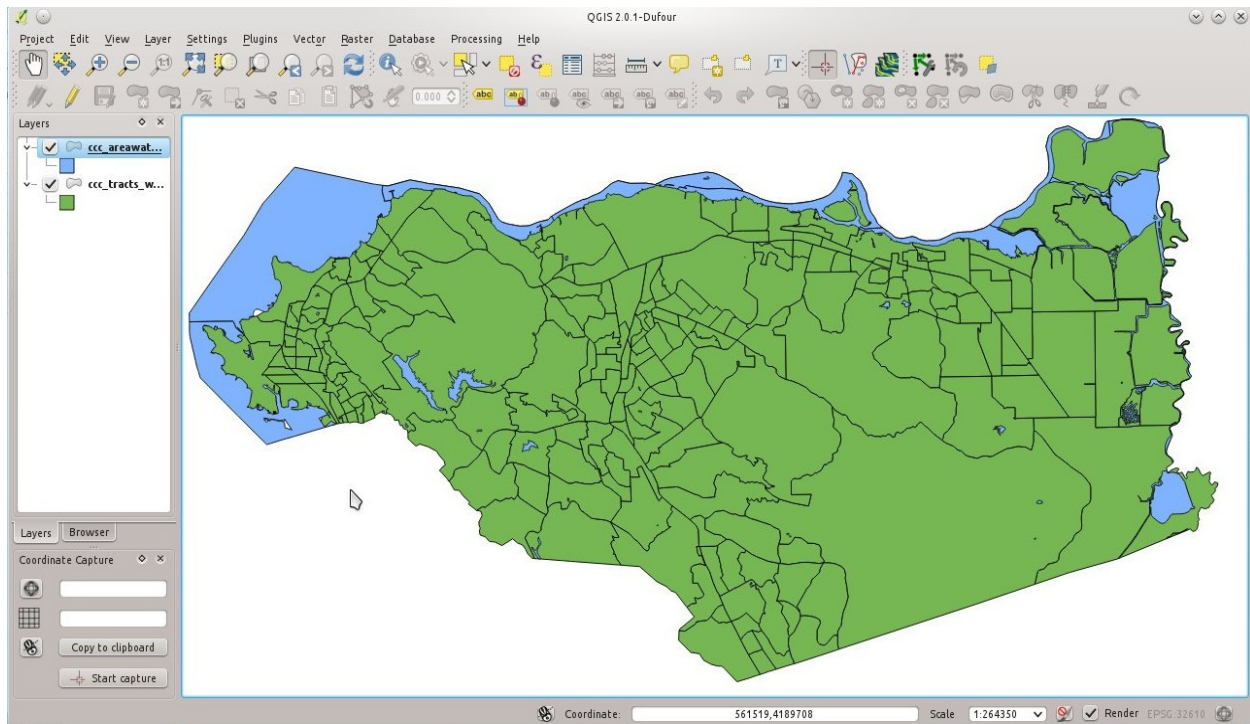
Rather than load this reprojected water file back into QGIS, proceed directly to reprojecting the Census Tracts shapefile to the same projection:



Unlike with the “areawater” file, QGIS will ask you to specify the Coordinate Reference System for this shapefile. Why? Because (at least for the moment) QGIS does not understand “web mercator” as a projection system. So I looked it up for you. It is EPSG 3857, which you can type up in the **Filter** field. QGIS calls this projection “WGS84/Pseudo Mercator”. I don’t know why. What I do know is that this is the correct projection, because the EPSG code matches.

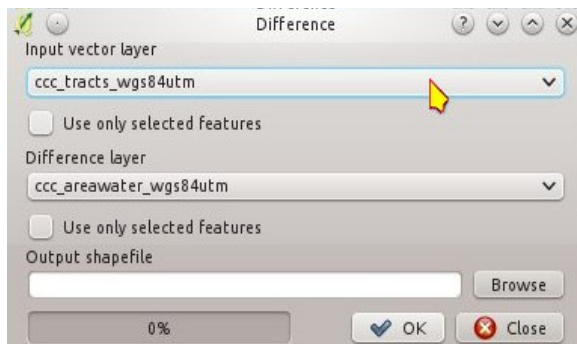
I go through exactly the same procedure to reproject this shapefile to WGS84/UTM Zone 10N.

Then I load both files.



4. Trim the Census-Tract shapefile using the “areawater” shapefile.

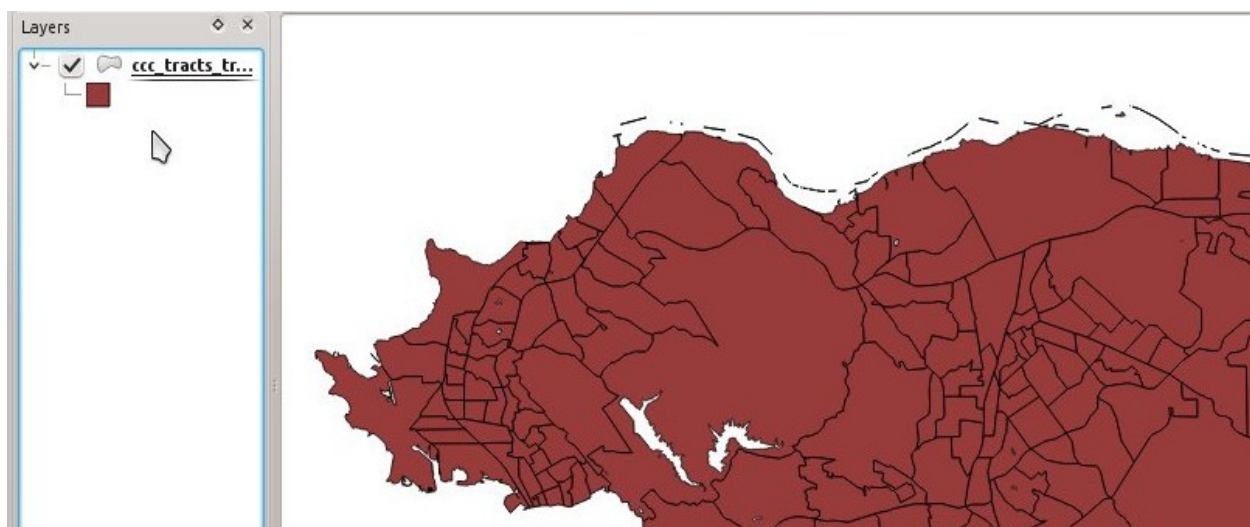
In QGIS, the way to trim one shapefile with another is to go to **Vector / Geoprocessing Tools / Difference...**



I called the output file “ccc_tracts_wgs84utm.shp”.

It generated smoothly, and then I loaded it into the workspace to review it (below).

In most ways it looks good, However there are some artifacts out in the water. Not sure what they are, but I noticed that ArcMAP did a better job of a clean erasure of polygons. In this case it will not matter because I am aiming to clip away most of Contra Costa County anyway, leaving only Richmond.



5. Create a GEOid join-field in the Attribute Table of the Census Tract File.

Attribute table - ccc_tracts_trimmed :: Features total: 207, filtered: 207, selected: 0

	GEO_ID	STATE	COUNTY	TRACT	NAME	LSAD	SHAPE_AREA	SHAPE_LEN
0	1400000US06013302006	06	013	302006	3020.06	Tract	23207378.64550000057	21465.66800430000
1	1400000US06013303102	06	013	303102	3031.02	Tract	10518794.38719999976	15089.79689170000
2	1400000US06013313102	06	013	313102	3131.02	Tract	3236227.45238000015	7154.77929570000
3	1400000US06013313206	06	013	313206	3132.06	Tract	2774919.23469999991	6876.07746981000
4	1400000US06013326000	06	013	326000	3260	Tract	3902590.59161000000	11676.41825440000
5	1400000US06013336101	06	013	336101	3361.01	Tract	1125502.85098000010	5092.13904159000
6	1400000US06013338301	06	013	338301	3383.01	Tract	3157875.58630000008	7693.36114581000
7	1400000US06013352201	06	013	352201	3522.01	Tract	11176254.78109999932	14429.95013070000
8	1400000US06013355109	06	013	355109	3551.09	Tract	4876677.62676000036	10605.63606110000
9	1400000US06013355115	06	013	355115	3551.15	Tract	3748104.69963999977	9358.15113468000
10	1400000US06013366002	06	013	366002	3660.02	Tract	1650978.05789999990	6290.56392389000
11	1400000US06013371000	06	013	371000	3710	Tract	2112953.85225000000	7044.69905641000

Show All Features

Okay, so here is some bad news about grabbing the TIGER file straight out of FactFinder2: it has a lot of the same attributes, but it does not have GEO.id2. What to do??? One option, if we backtrack a bit, is to use the GEO_ID attribute as the correlation-field. Another, which I will show here, is to use the **Field Calculator** function to concatenate together the STATE, COUNTY, and TRACT field values. Together, they constitute the GEO.id2 values. So! We will use this as an opportunity to generate a new attribute field!

Attribute table - ccc_tracts_trimmed :: Features total: 207, filtered: 207, selected: 0

	GEO_ID	STATE	COUNTY	TRACT	NAME	LSAD	SHAPE_AREA	SHAPE_LEN
0	1400000US06013302006	06	013	302006	3020.06	Tract	23207378.64550000057	21465.66800430000
1	1400000US06013303102	06	013	303102	3031.02	Tract	10518794.38719999976	15089.79689170000
2	1400000US06013313102	06	013	313102	3131.02	Tract	3236227.45238000015	7154.77929570000
3	1400000US06013313206	06	013	313206	3132.06	Tract	2774919.23469999991	6876.07746981000
4	1400000US06013326000	06	013	326000	3260	Tract	3902590.59161000000	11676.41825440000
5	1400000US06013336101	06	013	336101	3361.01	Tract	1125502.85098000010	5092.13904159000
6	1400000US06013338301	06	013	338301	3383.01	Tract	3157875.58630000008	7693.36114581000
7	1400000US06013352201	06	013	352201	3522.01	Tract	11176254.78109999932	14429.95013070000

Click the “Toggle editing mode” button in the upper left corner of the **Attribute table** window.

From here we could use the **Field Calculator** to simultaneously create a new field and fill it with calculated values; but I am going to create a field directly.

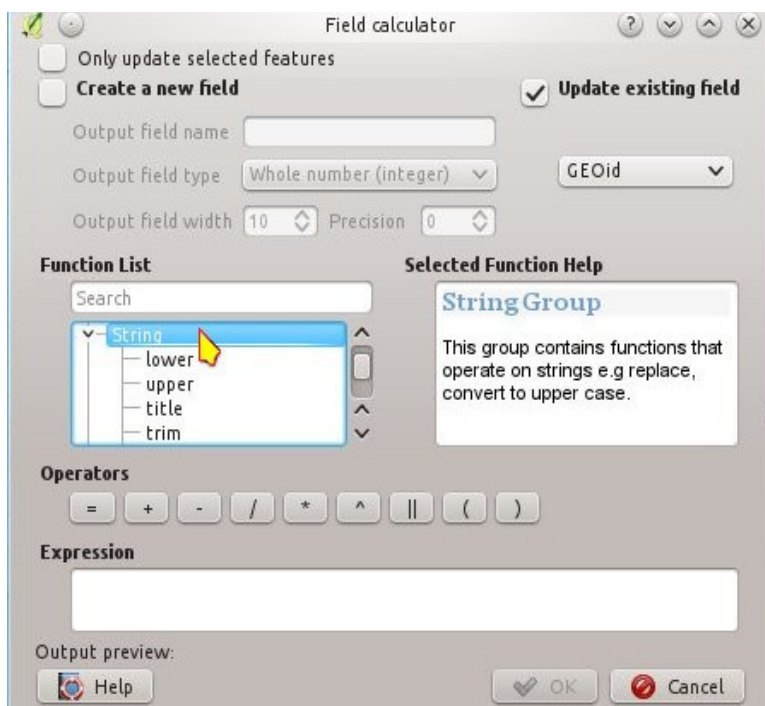
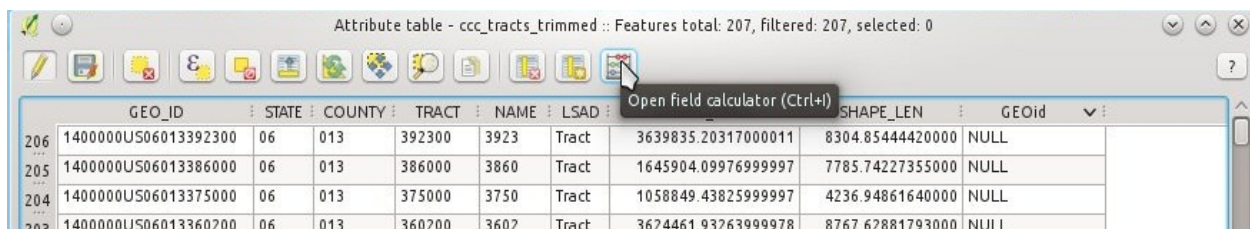
Attribute table - ccc_tracts_trimmed :: Features total: 207, filtered: 207, selected: 0

	GEO_ID	STATE	COUNTY	TRACT	NAME	LSAD	New column (Ctrl+W)	SHAPE_LEN
0	1400000US06013302006	06	013	302006	3020.06	Tract	1400000US06013302006	21465.66800430000
1	1400000US06013303102	06	013	303102	3031.02	Tract	1400000US06013303102	15089.79689170000
2	1400000US06013313102	06	013	313102	3131.02	Tract	1400000US06013313102	7154.77929570000
3	1400000US06013313206	06	013	313206	3132.06	Tract	1400000US06013313206	6876.07746981000

Below I have specified a new “GEOid” field, in which numbers will be treated as a text string, with space for eleven characters.



Now click the right-most button, which looks like an abacus. This starts the **Field Calculator**.



The Field Calculators of QGIS and ArcMAP look different, but they seem to do the same functions. Here you can build expressions, like the functions in Excel, but designed to manage “spatially-aware databases”.

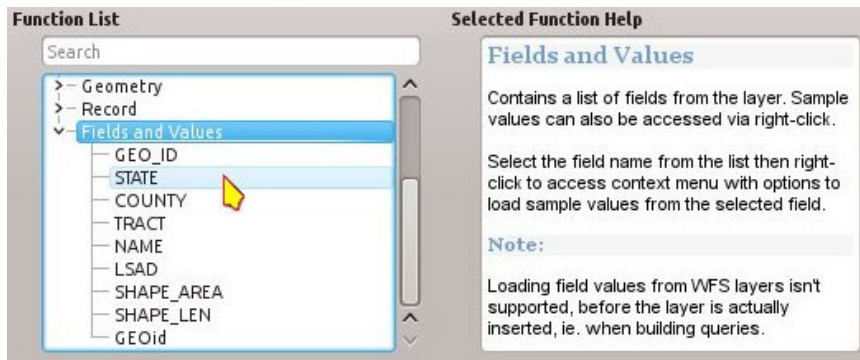
For now we will do a practical, simple function: concatenate the values in three fields to create a fourth field.

Concatenation is more of a ‘text-assembly’ function than a mathematical operation, so it is grouped under the **String** functions.

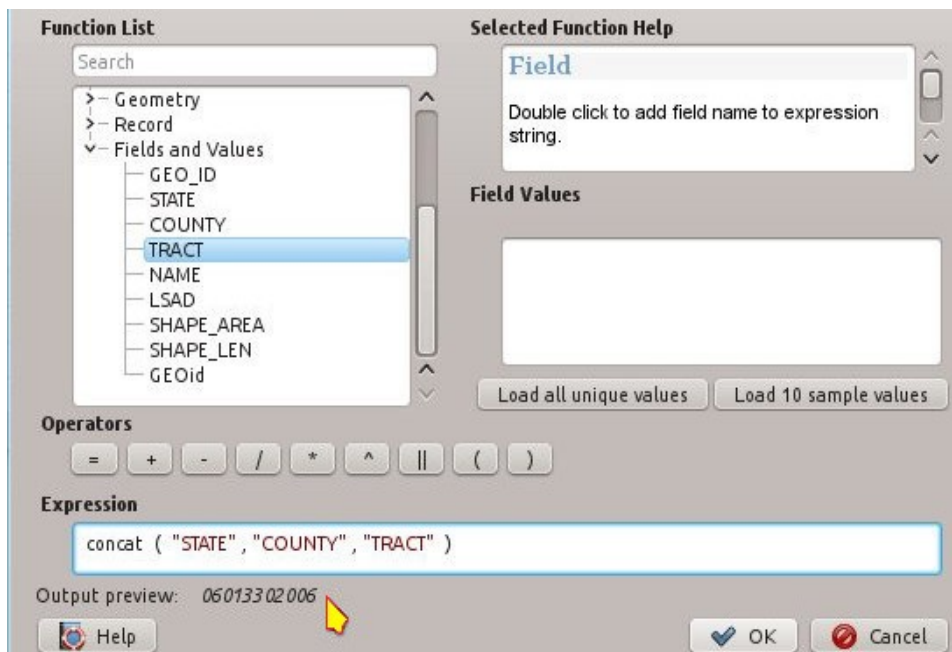


Double-click it and it will appear in the **Expression** field below. The **Selected Function Help** pane on the right includes descriptions and syntax-guidance.

Next, double-click on the **Field Names** and they will be inserted in the **Expression** area:



The full **concatenate** expression looks like this:



At the very bottom of the panel, notice the **Output preview**. Looks very much like the GEO.id2 values we worked with in the first Part of this project. Since it looks correct, I hit OK. Now the Attribute Table has one more column on the right-hand end:

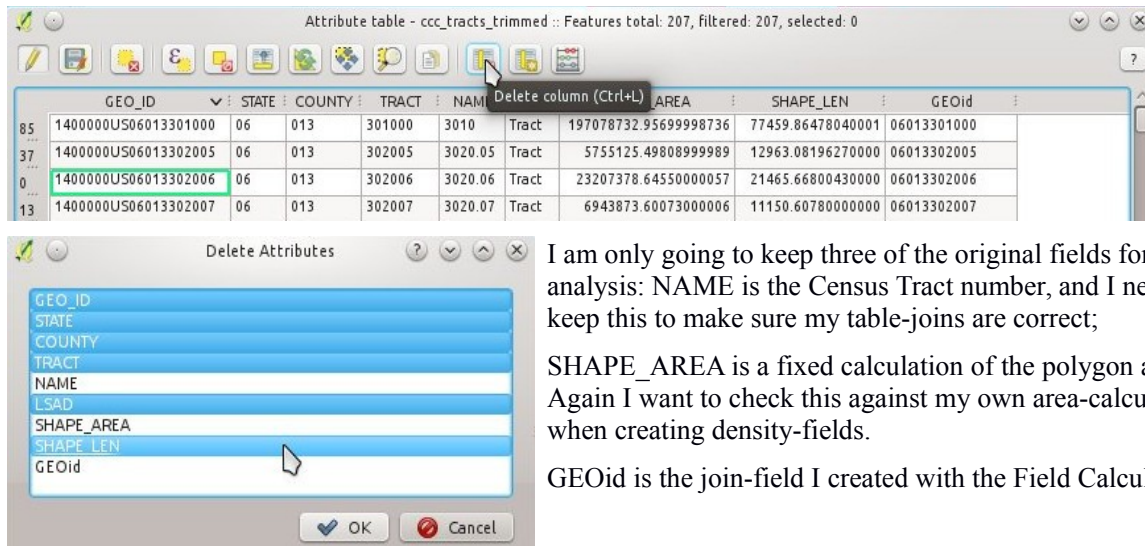
Attribute table - ccc_tracts_trimmed :: Features total: 207, filtered: 207, selected: 0

	GEO_ID	STATE	COUNTY	TRACT	NAME	LSAD	SHAPE_AREA	SHAPE_LEN	GEOid
85	1400000US06013301000	06	013	301000	3010	Tract	197078732.95699998736	77459.86478040001	06013301000
37	1400000US06013302005	06	013	302005	3020.05	Tract	5755125.49808999989	12963.08196270000	06013302005
0	1400000US06013302006	06	013	302006	3020.06	Tract	23207378.64550000057	21465.66800430000	06013302006
13	1400000US06013302007	06	013	302007	3020.07	Tract	6943873.60073000006	11150.60780000000	06013302007

This file now has a GEOid join-field for receiving FactFinder data. At this point you can click the “Save Edits” button, second from the left, to save your modified Attribute Table.

6. Delete unnecessary fields in the Attribute Table

Since we are in editing mode, this would be a good time to remove attribute-fields we will not use. The third button from the right is the “Delete Column” button.:

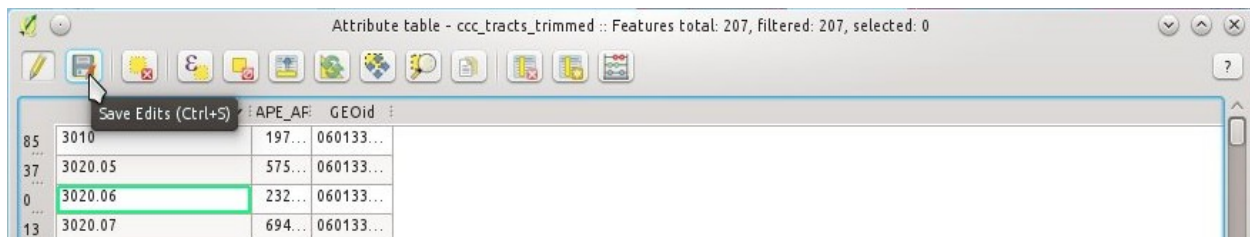


I am only going to keep three of the original fields for this analysis: NAME is the Census Tract number, and I need to keep this to make sure my table-joins are correct;

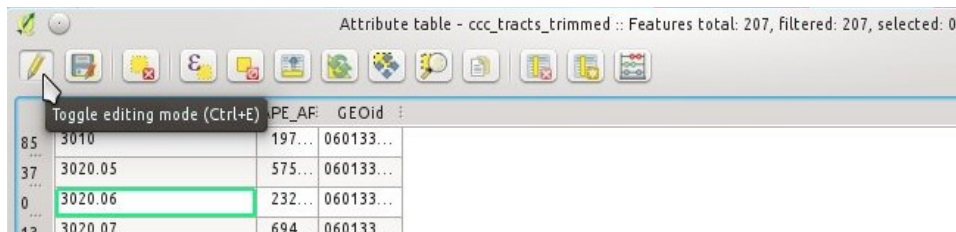
SHAPE_AREA is a fixed calculation of the polygon areas. Again I want to check this against my own area-calculations when creating density-fields.

GEOID is the join-field I created with the Field Calculator.

Again, I “Save Edits”:



And then click the “Toggle Editing Mode” button to shift out of editing mode.

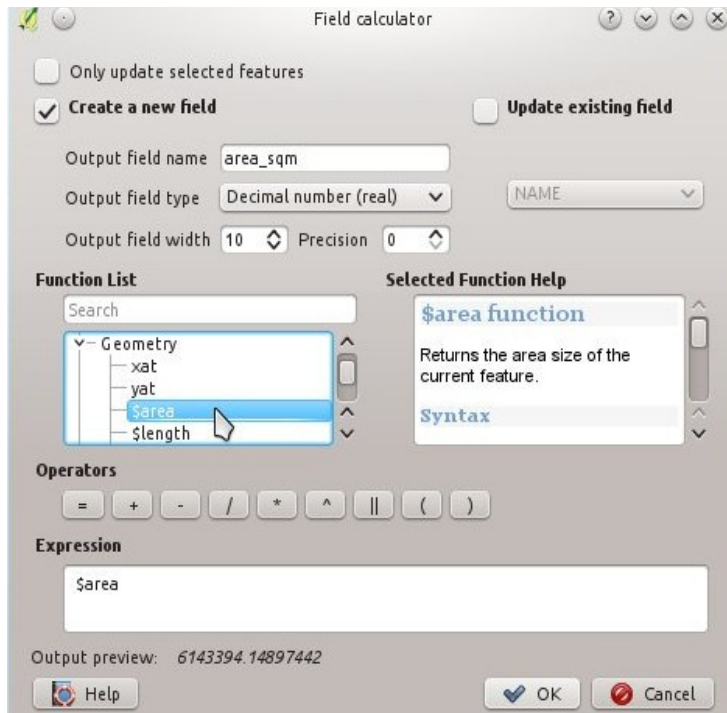


7. One last thing: a reality-check

So we have reprojected, cleaned, buffed, and polished this Census-Tract file. Now I want to derive my own area-calculations as a reality-check, and because my next step will be to create an area-acre field for the purpose of calculating data densities. That way, when I clip away parts of a Census-tract polygon and reduce its area, I will still know how many or how much of a given thing there might be in that area. Here is an example, to illustrate:

- Census tract #12345 is ten acres. It has 100 houses in it. So I calculate a density of 10 houses per acre.
- Then I clip away the parts of Census-tract #12345 that lie outside of Richmond, CA. What remains is 4 acres.
- However, in the Attribute Table of my clipped file, it still says 10 acres and 100 units. Both are wrong. Why didn't the software recalculate these? Because they are just numbers in the attribute table, and

GIS software does not know that it should change along with a change in polygon-area unless that number is deliberately linked to the polygon-area. Instead, we will need to recalculate it by hand.



In this case I use the Field Calculator to create a new field named “area”.

I set the **Output field type** to “Decimal number (real)”, and leave the **Field width** and **Precision** at their defaults.

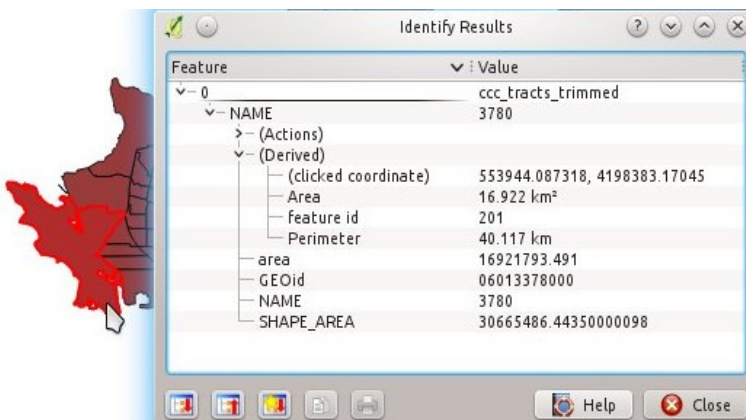
The function I select is **\$area** within the **Geometry** group of functions.

At the bottom you can see an **Output preview** of the calculation, and it looks good, so I click **OK**.

Unfortunately, as you can see below, the resulting area and the original SHAPE_AREA data we got from the Census are completely different. Maybe the original data is in square feet; maybe it was calculated before the polygon was reprojected. But it makes me wonder: is my reprojection wrong?

	NAME	SHAPE_AREA	GEOid	area
0	3020.06	23207378.64550...	06013302006	6143394.149
1	3031.02	10518794.38719...	06013303102	6525422.824
2	3131.02	3236227.452380...	06013313102	2004941.013
3	3132.06	2774919.234699...	06013313206	1718676.881
4	3260	3902590.591610...	06013326000	2421958.411
5	3361.01	1125502.850980...	06013336101	697941.576

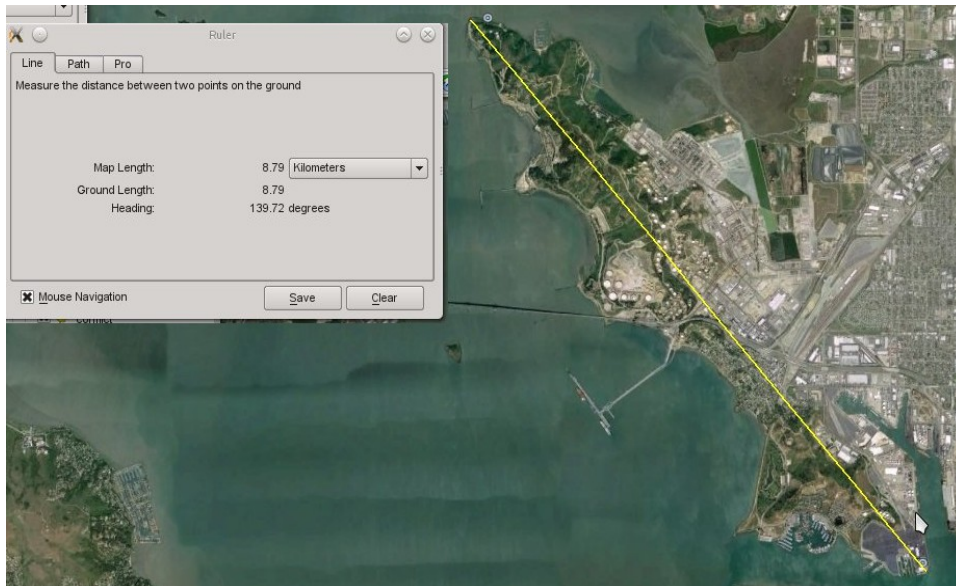
I use the **Identify Features** button to derive data from the reprojected shapefile:



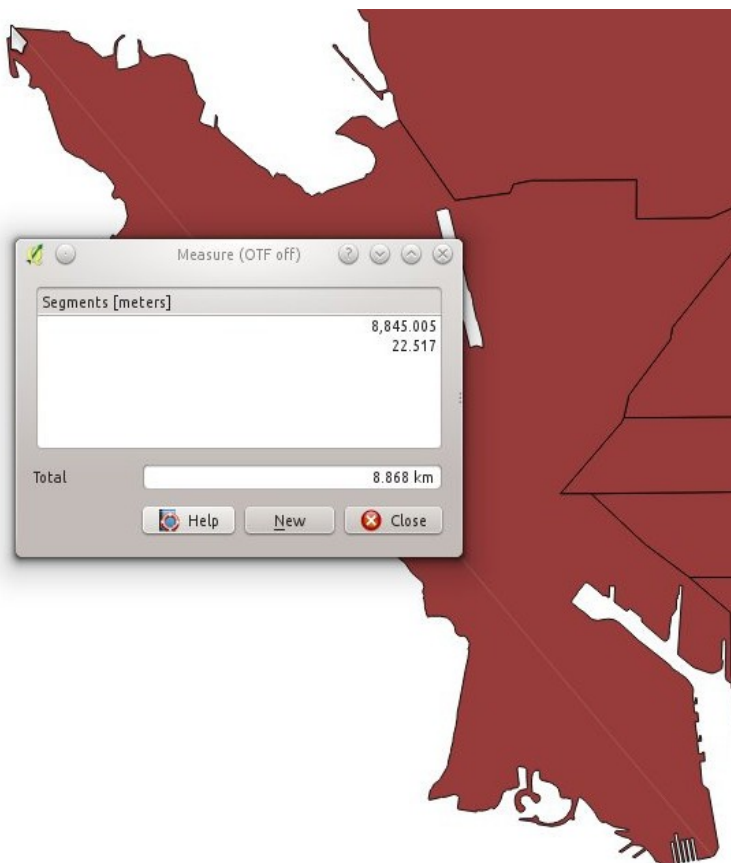
Hmm. The “Derived area” is 16.922 square kilometers. That matches my Field-Calculated area for that same Tract, in square meters.

But the original SHAPE_AREA number is totally discrepant. It cannot be square feet, either, because the ratio of square meters to square feet is about 1:10. So is the scale of my reprojected shapefile totally wrong? It indicates that Point Richmond is 16 sq km. Is that correct?

To check this, I fire up Google Earth and start comparing it to geographic features in my shapefile.



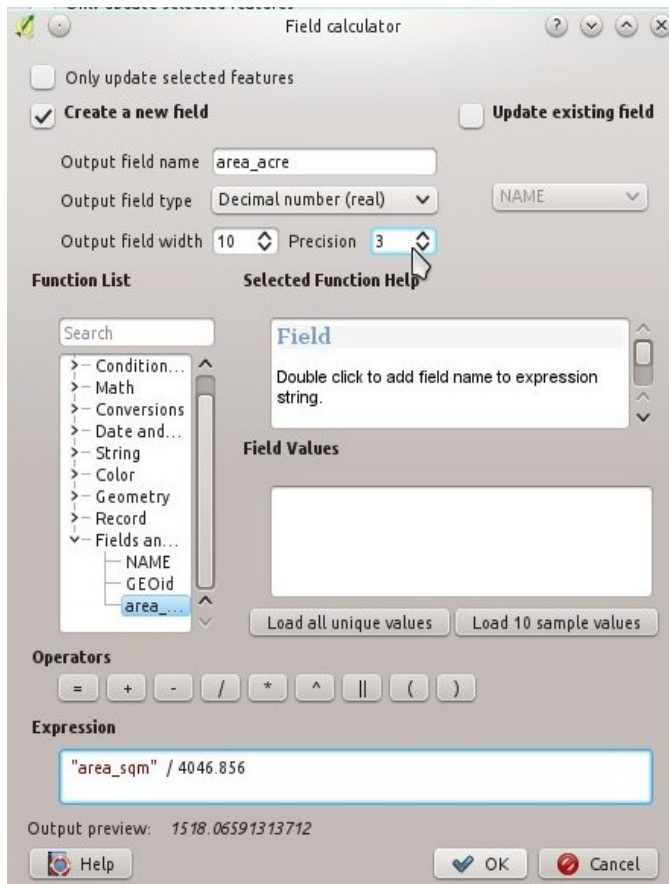
Here is Point Richmond in Google Earth. I set the units to Metric and used the measuring tool. It shows Point Richmond as 8.79 kilometers long. In which case an area-calculation of 16 square kilometers is plausible.



Here is the same feature in my shapefile. I use the Measure tool to measure its length. It measures as 8,845 meters; in other words, it matches the “real-world” measurement from Google Earth.

For me, this is a sufficient reality-check that my shapefile is the correct scale, and that the areas I am deriving are accurate.

I don't know what the SHAPE_AREA number represents; at some point I will research that. But I know that my own derived areas are accurate, and I am now going to calculate another field for acreage. That is the number I will use for my data-density calculations: people per acre, units per acre, etc. The reason I am taking this extra step is that local planners will be familiar with density-per-acre calculations. Density-per-sq meter may be accurate, but it is an abstract number that you cannot easily cross-check with other data.



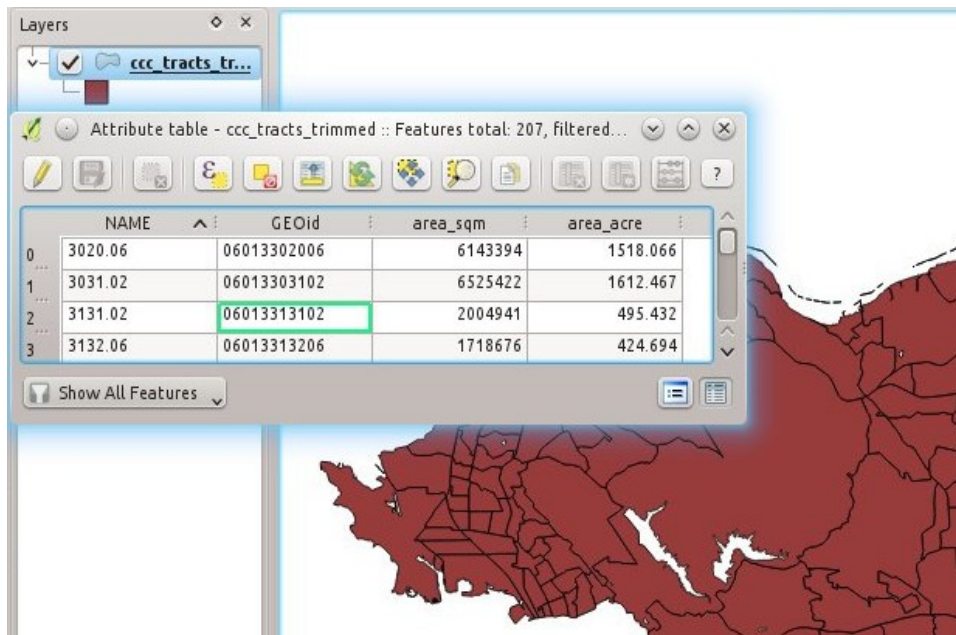
Using the **Field Calculator**, I will now create an “area_acre” field, derived from the “area_sqm” field [yes, I renamed the “area” field to “area_sqm” in the process of creating this tutorial. Sorry]

In this case, the function is to divide “area_sqm” by 4046.856. That is the ratio of square meters to acres; I looked it up again on the internet.

One more thing I needed to do in this case: I had to set the **Precision** to three decimal places, because I was doing an arithmetic operation in which I manually entered the divisor. And 4046.856 only has three decimal places, so I am settling for that degree of accuracy.

In this case the **Output preview** is misleading, because it shows a calculated result with 11 figures after the decimal. If I left **Precision** at zero, this operation would actually round to whole numbers. If I set it to eleven, I suppose it would calculate to the precision shown in the **Preview**. But realistically, I don’t need that level of precision.

Here is the result. Our shapefile is now ready for the next series of operations: Join, Normalize, Clip, and Analysis. Those will be in a separate tutorial.



8. Summary

Remember that the objective of this project is to analyze Richmond, California. This is not so easy, because Census-Tract data does not line up with Richmond. Therefore we will need to do the following:

1. join in the data we are interested in, but for all of Contra Costa;
2. ***normalize the data*** we will analyze for Richmond;
3. clip away all the rest of Contra Costa County;
4. Recalculate the areas of the partial Census-Tracts for Richmond only;
5. And then recalculate the raw figures (populations, housing, etc) for the partial Census-Tracts of Richmond itself.

We are getting there. In this tutorial I showed you how to produce a shapefile of Census-Tracts of Contra Costa County that is reprojected, with water-areas trimmed away, and with the Attributes we need and nothing more.