Random selection, QGIS

With the random selection -tool you can easily sample only parts of your data set. This is very handy when a data set contains tens, hundreds and thousands of objects, and a quality control for individual objects should be conducted. All objects are impossible to go through one-by-one, and manually choosing, for example, the first ten objects from the attribute table might cause skewness in the results.

How many samples should be examined, then? This depends on the size of the data set. Usually, a dataset that contains maximum of 50 objects can be examined thoroughly and no sampling is needed. After that, the object number is beginning to be too high and all objects would be too laborious to examine one-by-one. A core rule is that 1 to 10 percent of the objects should be examined in order to get an impression of the positional quality of the objects. However, with a limited amount of time and resources in manual quality checking, the percentage can be even smaller.

Thresholds could be, for example, as following:

- 1 to 50 objects → no sampling needed
- 50 to 500 objects → sample size is 10 %
- 500 to 5000 objects → sample size is 1 %
- 5000 to 50 000 objects → sample size is 0,1 %
- 50 000 to 500 000 objects → sample size is 0,01 %

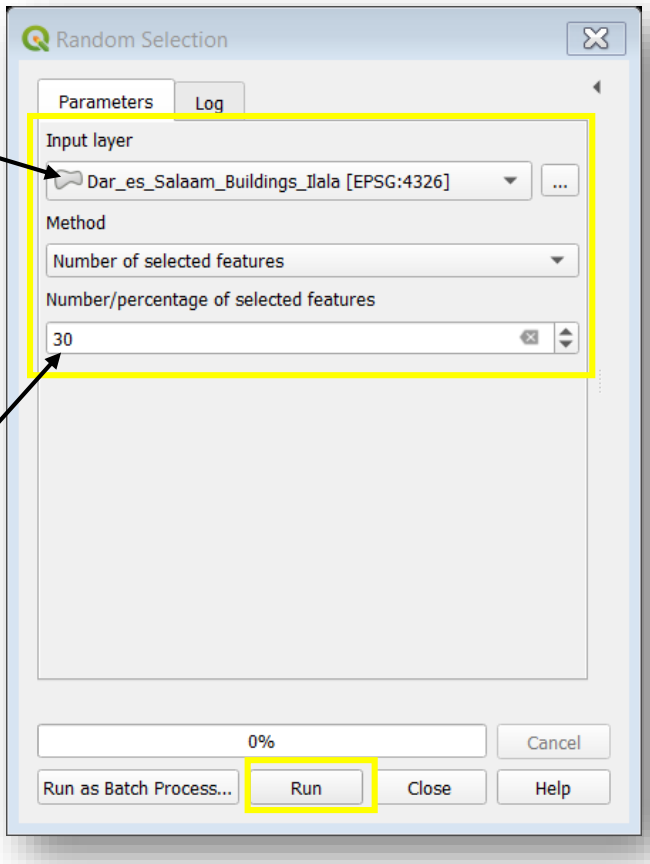
Random Selection

1. Open your vector layer in QGIS.
2. Check the number of objects by opening attribute table. Number of objects is shown in the top bar of the table.
3. Open Random Selection -tool. Go to top-navigation bar of QGIS and navigate to *Vector → Research Tools → Random Selection…*
4. The tool window pops up. Select your layer from the *Input layer* drop-down menu.
5. Set the selection method from the second drop-down list (number or percentage). NOTE: the tool doesn't handle decimal numbers, so if your data set is large and the sample should be smaller than 1 percent, calculate the number of sampled object beforehand and use method=number.
6. Set the percentage or number of selected objects.
7. *Run* the process.

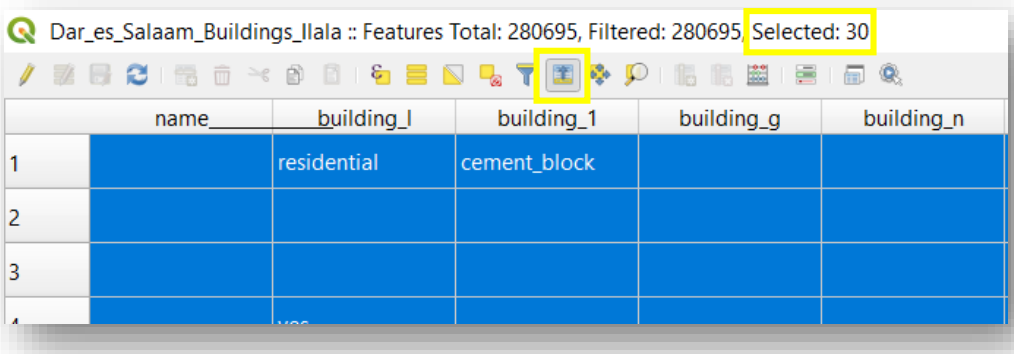
Here, a layer of buildings in Ilala district of Dar es Salaam is examined.

The layer has over 280 000 objects. 10 % sample of those would be 2 800 objects, which is overwhelmingly too many objects to be examined manually.

Thus, a much more smaller percentage is used (0,01 %), which leaves the sample number to 28. In this example, a nice round number of 30 is used.



8. After running the process, open the attribute table of your layer. Now, the number of selected objects is shown in the top-bar of the attribute table.
9. Optional: If you want to save the selection for later examination, export the selected objects and save them as separate layer. This can be useful if you don't have the time to finish the quality control at once, or if there are some problems in the precision and accuracy of the selected objects, and you would like to present them to the data provider as an example.
10. Click the *Move selection to top* icon.



Positional evaluation

1. Examine the positional precision and accuracy of each sampled object by right-clicking the first object in the attribute table and choosing *Zoom to Feature*.
2. The view of the QGIS map window is now zoomed to the first selected object. The object is highlighted as yellow. Evaluate visually whether the object matches with your reference satellite imagery: is the position correct (object is where it should be) and is the position precise enough for the purpose of the data (shapes of the objects are detailed enough)?

> Reference satellite imagery
> For reference aerial data, use for example satellite images provided by Google, ESRI or Bing which can be opened directly in QGIS. To do this, add the satellite maps as *XYZ tiles* via URL links, or by installing *QuickMapServices* plugin. Here are links to useful blog posts which provides instructions for adding base maps to QGIS:
> Instructions for connecting XYZ tiles
> Instructions for using QuickMapServices

3. To ease the comparison of the objects and the reference image, you may change the visualization of your layer to partly transparent: *Right click the layer → Properties → Symbology → Opacity.*
4. The selection opacity can be modified from *Project → Properties → General → Selection color.*
5. Examine all the randomly selected objects one-by-one. Mark down how many of those objects had some positional problems and identify what kind of problems they were. You may do this by using an Excel-table where you write the FID of the object and reason for an error, or by creating a new temporary field to the attribute table. In the quality report, provide a comment of the error rate (number of errors divided by the number of samples) and description of the most common problems.

> Error rate
> Error rates are usually conducted by calculating omission errors, commission errors, and user's and producer's accuracy metrics (read more from this Humboldt University blogpost). However, error rates can be calculated also in more simpler cases wherever a produced data set is compared to a reference data. If the data set contains 100 observations (or objects) and the sample size is 10, and of those ten two were positionally inaccurate, the error rate would be 2/10 = 20 %.
>
> A skilled would then evaluate what the reasons behind the inaccuracies might be: are they due to GPS errors, digitizing errors, errors in reference data (either in data production phase or in this quality evaluation phase) or something else? Can these errors be easily fixed or not?

In the case of Ilala district building data, multiple positional errors can be found. In the first screenshot the selected object (yellow) is actually in the correct position when compared to the ESRI satellite image, but the buildings around are clearly out of place. Also, some objects cannot be found from the satellite image, but that is just an error of the reference data (outdated satellite image).

In the respect of calculating the error rate, this object would be marked as OK in terms of positional accuracy and precision. The corners of the object matches with the corners of the real-world building footprint. However, when moving on to other randomly selected objects, most of them are not in correct locations.

The reason for positional errors in this data set is most probably the choice of reference data (i.e. background image) in the digitizing process. This data is OpenStreetMap data and mostly produced via HOT OSM id editor. There, the coordinates of background images sometimes varies between each other. This is due to the fact that satellite images always stretches towards the edges of that image strip, and images taken from different angle, distance and width always has some skewness in them.

This problem is avoided when the same background satellite image is used thorough the digitizing process, or they are aligned with a simple step before starting digitizing. However, the OSM data is produced by crowdsourcing and multiple mappers might not always notice to choose the same image others have been using. Thus, errors between the final objects can be found from all around the building data set.

The errors are only couple of meters between the objects, but such errors might affect analysis where very detailed information is needed, which is why these kind of errors are important to be mentioned in the metadata of the data set, but also in the quality report filled by the Data Managers of the Climate Risk Database.