Tutorial for delineating natural cities using population count

Haofeng Tan

Urban Governance and Design Thrust, Society Hub
Hong Kong University of Science and Technology (Guangzhou), China
Email: haofeng_tan@foxmail.com

I. Introduction

Natural cities refer to human settlements or human activities in general on the Earth's surface that are objectively or naturally defined and delineated from massive geographic information of various kinds, and based on the head/tail breaks method. Traditionally, a city is characterized as a relatively large and permanent human settlement. But how large a settlement must be to qualify as a city remains ambiguous. Additionally. many cities have a particular administrative, legal, and historical status according to their local laws, which are very subjective. This subjectivity is also demonstrated in the physical boundaries of cities, which are legally and administratively determined. Deriving natural cities based on head/tail breaks method brings a new idea: given a variable X, if its values x follow a heavy tailed distribution, then the mean (m) of the values can divide all the values into two parts: a high percentage in the tail, and a low percentage in the head. The heavy tailed distribution refers to the statistical distributions that are right-skewed, for example, power law, lognormal, and exponential. Obviously, the density of population count, street nodes, the size of street blocks, and the nighttime imagery pixel values all exhibit a heavy tailed distribution, which implies that there are far more small things than large ones.

In this tutorial, natural cities are identified by the head tail breaks method proposed by Prof. Bin Jiang using population count data. Given that cities tend to be highly populated, cities and rural areas can be distinguished with the help of population count data. Therefore, natural cities are delineated by selecting locations where population count exceeds the regional average value. Then, the hierarchy of natural cities is interpreted. The rest of this tutorial has three sections. Section 2 provides information of data resources and software used. Section 3 is a detailed guidance for delineating natural cities. The hierarchy of natural cities is interpreted in the Section 4.

II.Introduction Data sources and software support

This tutorial is based on the Chinese district of Nansha, Guangzhou. The administration division shapefile can be downloaded from GitHub (https://github.com/GaryBikini/ChinaAdminDivisonSHP). The input data used in this tutorial is population count data obtained from WorldPop Hub (https://hub.worldpop.org/project/categories?id=3), specifically Constrained Individual countries 2020 UN adjusted 100m resolution China (*chn_ppp_2020_UNadj_constrained.tif*).

The geoprocessing algorithms in this tutorial were implemented using QGIS-LTR 3.28.10-Firenze (https://qgis.org/en/site/). Other GIS (Geographic Information Systems), such as ArcGIS (https://desktop.arcgis.com/en/arcmap/), also can be utilized.

The head/tail breaks calculation (https://github.com/OneDay48Hours/Head-tail-Breaks) was implemented in a Python file (https://www.python.org) using NumPy (https://numpy.org), pandas (https://pandas.pydata.org) and GDAL (https://gdal.org/index.html). Image break values can be calculated by *head_tail_breaks_image.py*. The head/tail break metric, such as area, was exported to a csv file in one column prior to calculation. Then it was imported into *head_tail_breaks_csv.py* to calculate break values.

III. Delineating natural cities from population count data

In this tutorial, population count data was used to identify the boundaries of natural cities. Any

location where the population count exceeds the average level in a specific region will be regarded as a city. Conversely, locations with a population count below the average level will be regarded as rural areas.

1. Download and open the population count chn_ppp_2020_UNadj_constrained.tif dataset in QGIS (Layer -> Add Layer -> Add Raster Layer).

2. Download and open the district division dataset district.shp in QGIS ((Layer -> Add Layer -> Add Vector Layer).

3. Open the Attribute Table of the district.shp by right-clicking and select the Nansha district using Field Filter, c.f. Figure 1.

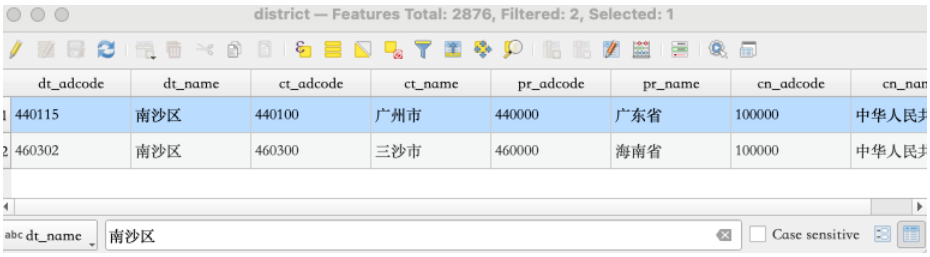


Figure 1: (Color online) Select Nansha district through attribute table

4. Clip the population count raster by the selected Nansha district feature (Toolbox -> GDAL -> Clip Raster by Mask Layer and tick selected feature only), save the clipped raster as nansha_ppp.tif.

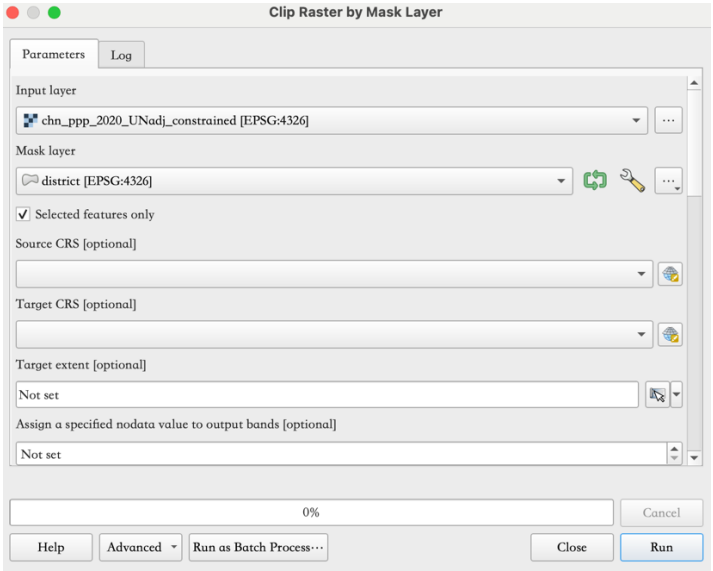


Figure 2: (Color online) Clip population count data

5. Calculate break mean values for clipped population count data running *head_tail_breaks_image.py* via command line.

6. Due to the first mean of the head/tail breaks method is not sufficient, and the second mean instead will be used. Delineate natural cities using the second mean population count value (Toolbox -> Raster Analysis -> Reclassify by Table), run it with the clipped population count dataset (nansha_ppp.tif), and set grid value to 1 (below average value) or 2 (above average value) respectively.

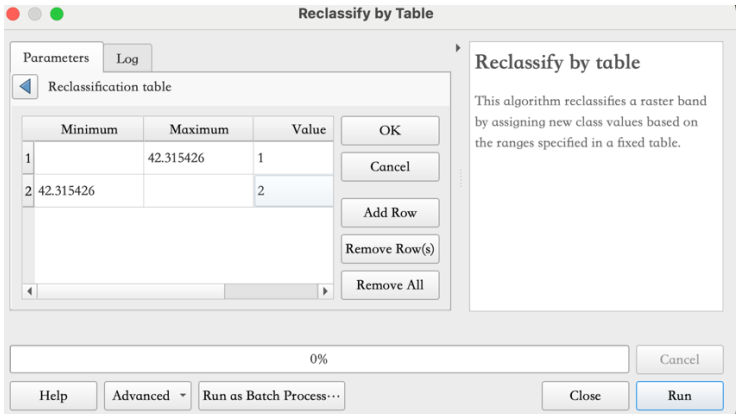


Figure 3: (Color online) Delineating natural cities using second mean value

7. Convert the classified data into polygon features (Toolbox -> GDAL -> Raster Conversion -> Polygonize), create new field DN, and project the polygons from *WGS84*, Geographic Coordinate System, to *CGCS2000/3-degree Gauss-Kruger CM 114E*, Projected Coordinate System (Toolbox -> Vector General -> Reproject Layer).

8. Highlight the reprojected polygons, open *Attribute Table*, select polygons with DN = 2, and export them to *natural_cities_first.shp* and *natural_cities_second.shp* by right-clicking the layer (*Export -> Save Selected Features as..*).

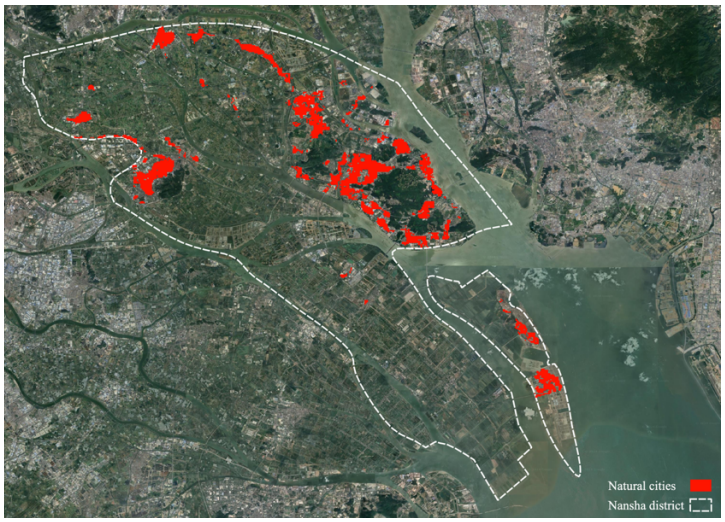


Figure 4: (Color online) Natural cities from second average value

IV. Interpreting the hierarchy of natural cities

The natural cities derived from population count data can be further analyzed by head/tail breaks method to explore their hierarchy structure based on the idea that far more small substructures than larger ones. In this tutorial, the criterion for classifying large natural cities is based on the metric - area of each natural cities.

1. Open the Attribute Table of the natural_cities.shp generated in the last section.

2. Click the Field Calculator  , tick Create a new field, name the new field Area, set it to Decimal type, and use $area function expression to calculate its value.

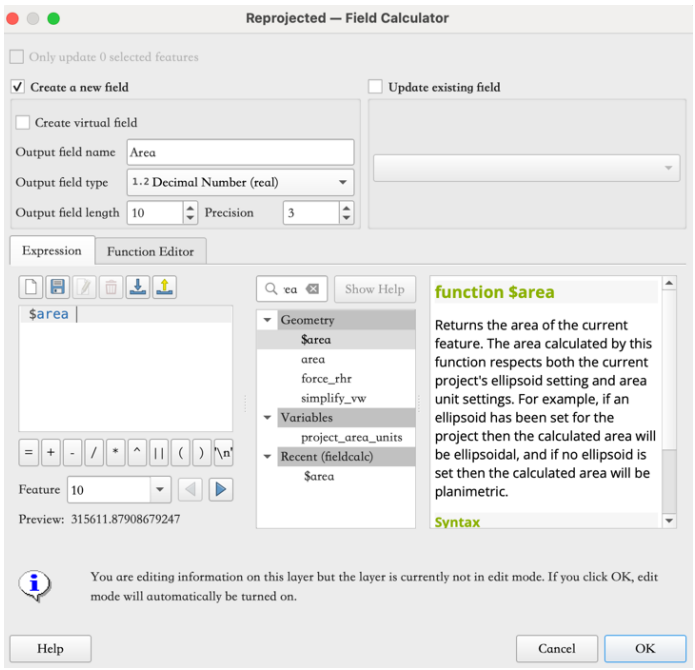


Figure 5: (Color online) Calculate area of each natural city

3. Highlight the natural cities layer, and export only Area field to naturalcities_area.csv (Export -> Save Selected Features as..).

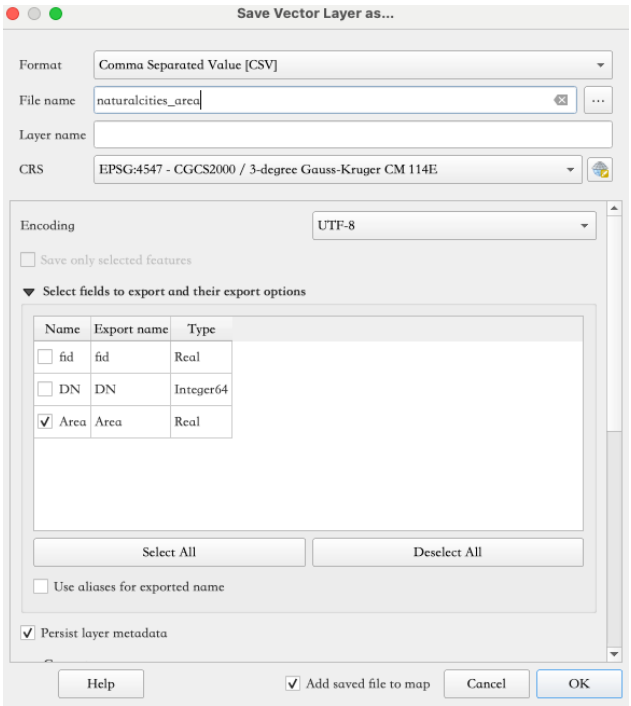


Figure 6: (Color online) Export natural cities area as csv

4. Run ht_break_excel.py by command line to calculate the cut values. (-f represents the excel file path, -c represents the column index, -v represents the head/tail breaks version)

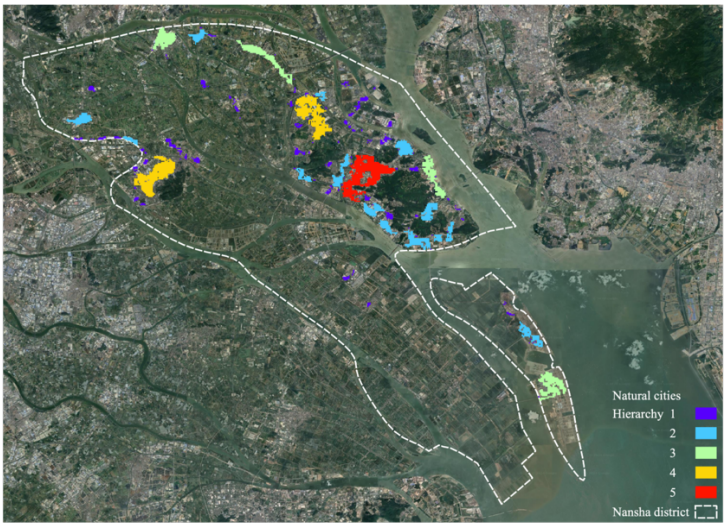
```
=====================================================
Head/tail version1 Breaks Detail:
   rows          mean  head  tail  head percentage  tail percentage
0   139  2.669860e+05    22   117         0.158273         0.841727
1    22  1.453238e+06     7    15         0.318182         0.681818
=====================================================
Break Points:
[266986.04022302164, 1453237.9651818182]
```

Figure 7: (Color online) Results of break values calculated by head/tail breaks version1
(Note: (Command Line) python ht_break_excel.py -f 'your_excel_file_path' -c 0 -v 1)

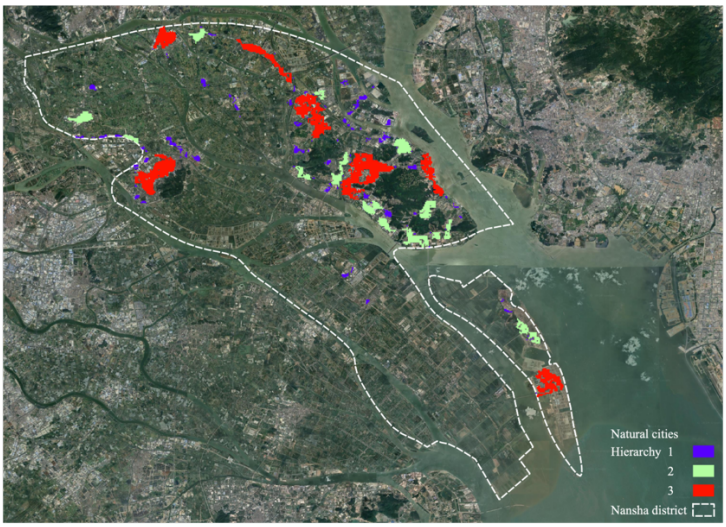
```
=====================================================
Head/tail version2 Breaks Detail:
   rows          mean  head  tail  head percentage  tail percentage  avg head percentage
0   139  2.669860e+05    22   117         0.158273         0.841727             0.158273
1    22  1.453238e+06     7    15         0.318182         0.681818             0.238228
2     7  3.065793e+06     3     4         0.428571         0.571429             0.301676
3     3  4.526692e+06     1     2         0.333333         0.666667             0.309590
=====================================================
Break Points:
[266986.04022302164, 1453237.9651818182, 3065792.913857143, 4526691.6296666665]
```

Figure 8: (Color online) Results of break values calculated by head/tail breaks version2
(Note: (Command Line) python ht_break_excel.py -f 'your_excel_file_path' -c 0 -v 2)

5. Using the cut values to classify the natural cities. Go to the properties -> symbology of the natural cities and select Rule-based, with Area as value. Then click the classify button and manually select the correct number of classes and break values.



(a)



(b)

Figure 9: (Color online) Natural cities interpreted by head/tail breaks (a) version1 (b) version2

Acknowledgement
I would like to thank Prof. Bin Jiang for his kind guidance and help in preparing this tutorial.

References:

Jiang B. (2013), Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution, *The Professional Geographer*, 65 (3), 482-494.

Jiang B. (2018), A topological representation for taking cities as a coherent whole, *Geographical Analysis*, 50(3), 298-313.

Jiang B. and Miao Y. (2015), The evolution of natural cities from the perspective of location-based social media, *The Professional Geographer*, 67(2), 295-306.

Lynch K. (1960), *The Image of the City*, the MIT Press: Cambridge, Massachusetts.