Tutorial for applying head/tail breaks in ArcGIS

Sirui Wu
Department of Technology and Built Environment, Division of Geomatics
University of Gävle, 801 76, Sweden
Email: wsr102209@163.com

Introduction

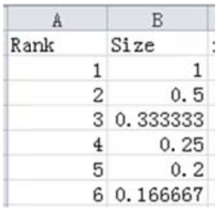
Power laws and lognormal distributions are called heavy-tailed distributions, which imply that there are far more small things than larger ones. To better illustrate the underlying scaling pattern of far more small things than large ones, a new classification scheme namely head/tail breaks (Jiang, 2013) has been developed. It divides things around an average, according to their geometric, topological, and/or semantic properties, into a few large things (in the head) and many small things (in the tail), and recursively continue for the large things or those in the head, until the notion of far more small things than large ones is violated. The applications of using head/tail breaks have been found of vital importance in mapping, map generalization and perception of beauty (Jiang 2014). This tutorial is intended to provide a step by step guide for applying head/tail breaks method in ArcGIS. In order to have a comprehensive understanding of using head/tail breaks method, conventional classification methods, Jenks natural breaks (Jenks, 1967), will also be conducted in this tutorial as a comparison.

The remainder of the tutorial is organized as five parts: Creating your own artificial data; importing data into ArcGIS; Jenks natural breaks; head/tail breaks and visualized comparison. Please make sure you had already installed ArcGIS software and Microsoft Excel is also available. No specific version of software is required in this tutorial but this tutorial suggests that you can use the ArcGIS 10.0 or latest version.

Create data

In this tutorial you are required to create your own data that follows heavy-tail distribution. Rank Size distribution, known as one of the typical heavy-tail distributions, is a very suitable distribution for testing head/tail breaks method. Now, go through following steps to learn how to create rank size distribution.

1. Start a new blank workbook in Microsoft Excel. Create value in column A (namely rank) with 1, 2, 3, 4… until 1023, column B (named size) with 1/A (Figure 1).

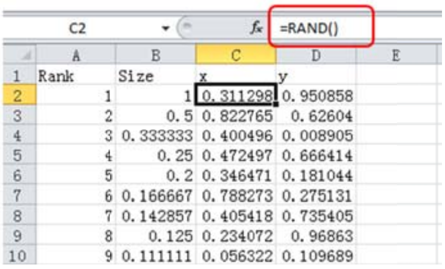


A	B
Rank	Size
1	1
2	0.5
3	0.333333
4	0.25
5	0.2
6	0.166667

Figure1: Create Rank and size in column A and column B.

2. Add two more columns, named *x* and *y*, assigning some random numbers to *x* and *y*. Note: you can use "***=RAND ()***" function in column C2 and click *Enter* to generate random values (Fig 2).

Fill all columns from column C2 to column C1024, i.e. create a random column then drag downwards when the cursor turns into a plus sign.



C2 =RAND()

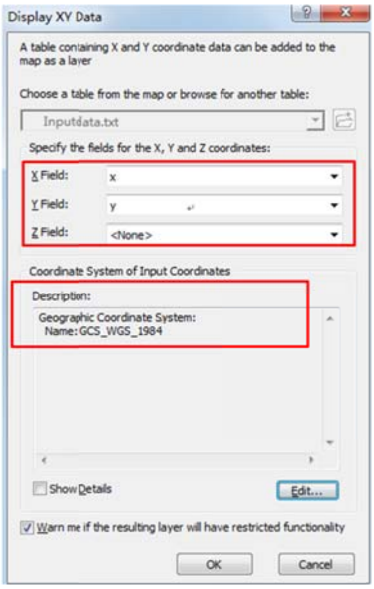
	A	B	C	D	E
1	Rank	Size	x	y	
2	1	1	0. 311298	0. 950858	
3	2	0. 5	0. 822765	0. 62604	
4	3	0. 333333	0. 400496	0. 008905	
5	4	0. 25	0. 472497	0. 666414	
6	5	0. 2	0. 346471	0. 181044	
7	6	0. 166667	0. 788273	0. 275131	
8	7	0. 142857	0. 405418	0. 735405	
9	8	0. 125	0. 234072	0. 96863	
10	9	0. 111111	0. 056322	0. 109689	

Figure 2: Use ***RAND ()*** function to create random values.

3. Fill all these four columns from row 2 to row 1024 and make sure there is no gap left. When you finish above steps, create a blank .txt file and paste all data from this Excel file to that file, saved as ***Inputdata.txt.***

Import data

1. Start ArcGIS and import previous data created in Microsoft Excel by clicking the button *Add data*
2. When ***Inputdata.txt*** data has been imported successfully, *right click* ***Inputdata.txt*** > *Display XY data*. A new window will appear as following (Figure 3).



Display XY Data

A table containing X and Y coordinate data can be added to the map as a layer

Choose a table from the map or browse for another table:

Inputdata.txt

Specify the fields for the X, Y and Z coordinates:

X Field: x

Y Field: y

Z Field: <None>

Coordinate System of Input Coordinates

Description:

Geographic Coordinate System:
Name: GCS_WGS_1984

□ Show Details

☑ Warn me if the resulting layer will have restricted functionality

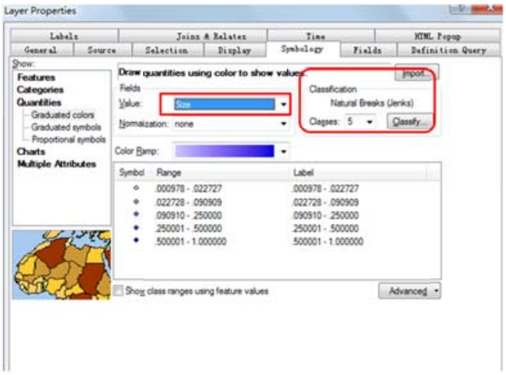
Figure 3: Set WGS1984 as input coordinate system.

3. Make sure *x* and *y* are corresponding to X Field and Y Field, respectively. Choose *GCS_WGS_1984* Geographic Coordinate System as input coordinate system. Click *OK* >*YES.*
4. *Right click* ***Inputdata.txt Events*** > *Data* > *Export Data*, named as ***naturalbreaks.shp.***
5. Export one more shpfile followed by step 4 and named as ***headtailbreaks.shp.***

Now you have successfully prepared these two test data in ArcGIS. Following parts will guide you how to process data with two different classification methods.

Jenks natural breaks

1. *Right click* ***naturalbreaks.shp*** *> Properties > Select Symbology > Quantities*. *Select* ***Size*** *as Value* (Figure 4). You may notice that natural Breaks (Jenks) have been automatically selected as a default and there are five classes. Then, click *Apply > OK.*



Draw quantities using color to show values.

Fields

Value: Size

Normalization: none

Classification

Natural Breaks (Jenks)

Classes: 5

Color Ramp:

Symbol	Range	Label
○	.000978 - .022727	.000978 - .022727
●	.022728 - .090909	.022728 - .090909
●	.090910 - .250000	.090910 - .250000
●	.250001 - .500000	.250001 - .500000
●	.500001 - 1.000000	.500001 - 1.000000

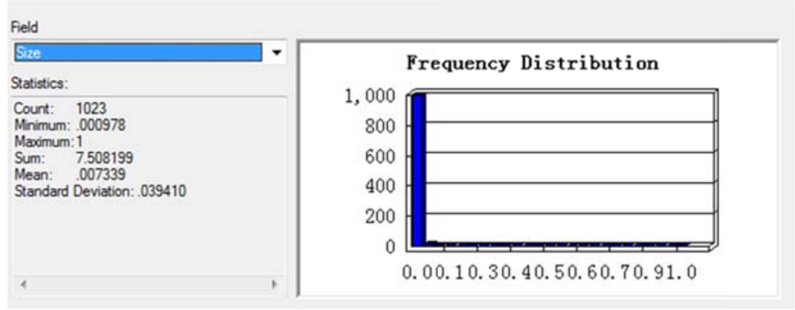
□ Show class ranges using feature values

Figure 4: Jenks natural breaks can be automatically calculated in ArcGIS.

Head/tail breaks

Unlike Jenks natural breaks that can be automatically classified in ArcGIS, head/tail breaks need to be manually calculated. Now we need to calculate the mean of given data *m1* and repeat this step until there is no long heavy-tail distribution.

1. *Right click* ***headtailbreaks.shp*** > Open attribute Table > *Right click* **size** column in attribute table > *Statistics*. Take notes for *count*, *sum* and *mean* information (Figure 5) and close it.



Field

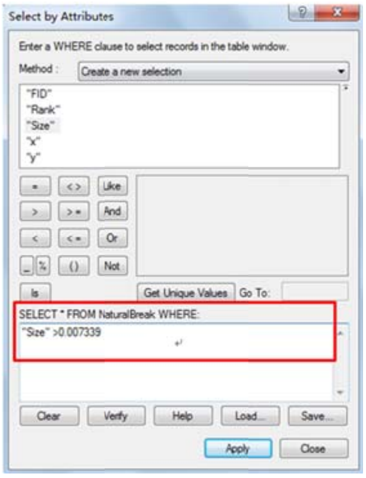
Size

Statistics:

Count: 1023
Minimum: .000978
Maximum: 1
Sum: 7.508199
Mean: .007339
Standard Deviation: .039410

Figure 5: The statistics of headtailbreaks data.

2. Click *Select by attributes* under the attribute table > Type ***"Size > 0.007339"*** as a Query under the *SELECT*FROM HeadTailBreaks WHERE* (Figure 6). Click *apply > Close*. This Query helps you to select those data whose size values are greater than the first mean *m1*.



Enter a WHERE clause to select records in the table window.

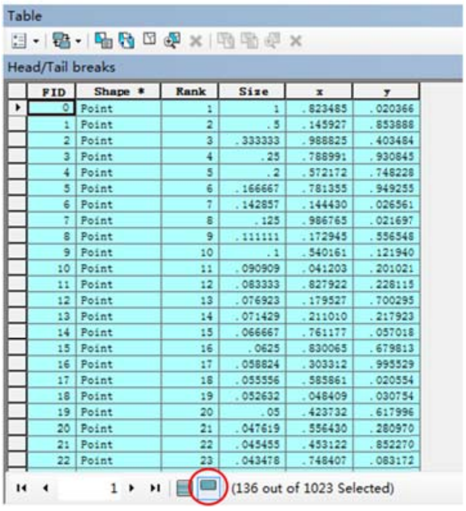
Method : Create a new selection

SELECT * FROM NaturalBreak WHERE:

"Size" >0.007339

Figure 6: Using ***Select by attributes*** function to select required data.

3. Once you have done this step, a new table will be created. This new table only contains features whose size values are greater than the first mean *m1*. You can check this new table by clicking *Show selected records* button (Figure 7). It can be seen below that there are 136 points which have been selected as a new data.



FID	Shape *	Rank	Size	x	y
0	Point	1	1	.823485	.020366
1	Point	2	.5	.145927	.853888
2	Point	3	.333333	.988825	.403484
3	Point	4	.25	.788991	.930845
4	Point	5	.2	.572172	.748228
5	Point	6	.166667	.781355	.949255
6	Point	7	.142857	.144430	.026561
7	Point	8	.125	.986765	.021697
8	Point	9	.111111	.172945	.556548
9	Point	10	.1	.540161	.121940
10	Point	11	.090909	.041203	.201021
11	Point	12	.083333	.827922	.228115
12	Point	13	.076923	.179527	.700295
13	Point	14	.071429	.211010	.217923
14	Point	15	.066667	.761177	.057018
15	Point	16	.0625	.830065	.679813
16	Point	17	.058824	.303312	.995529
17	Point	18	.055556	.585861	.020554
18	Point	19	.052632	.048409	.030754
19	Point	20	.05	.423732	.617996
20	Point	21	.047619	.556430	.280970
21	Point	22	.045455	.453122	.852270
22	Point	23	.043478	.748407	.083172

1 (136 out of 1023 Selected)

Figure 7: Click ***show selected records*** button to check selected data.

4. The given data will be divided into head part (136 points) and tail part according to the first mean value *m1*. Calculate the *m2* for those values greater than *m1* and obtain *m2*. Again, calculate the *m3* for those values greater than *m2* and obtain *m3*. Repeat these steps until head part are no longer heavy-tail distribution.

   <u>Tips: How to define head and tail part?</u>
   Head part: values which are greater than the mean value (not include the mean value).
   Tail part: values which are equal to or less than the mean value (include the mean value).
   More tips can be found in http://en.wikipedia.org/wiki/Head/tail_Breaks (***method*** part).

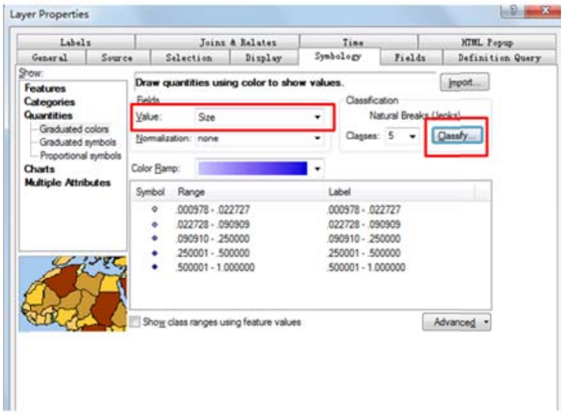
5. When you finished calculating mean value, record all mean values and calculate relevant statistics information showed as following (Table1).

Table 1: The statistics information of using head/tail breaks method
(#= number, %= percentage).

#SelectedPoints	Mean	#Head	%Head	#Tail	%Tail
1023	0.007339	136	15	903	85
136	0.040394	24	18	112	82
24	0.157332	6	25	18	75
6	0.408333	2	33	4	67
2	0.75	1	50	1	50

According to the above table, you can find that percentage of head part is increasing while the percentage of tail part decreased. In this tutorial, head/tail breaks calculation stopped when the percentage of head part is equal to the percentage of tail part.

6. Right click ***headtailbreaks.shp*** > *Properties* > *Symbology* > *Quantities* > Select ***Size*** as **value** again > Click *classify* (Figure 8).



Draw quantities using color to show values.

Value: Size

Normalization: none

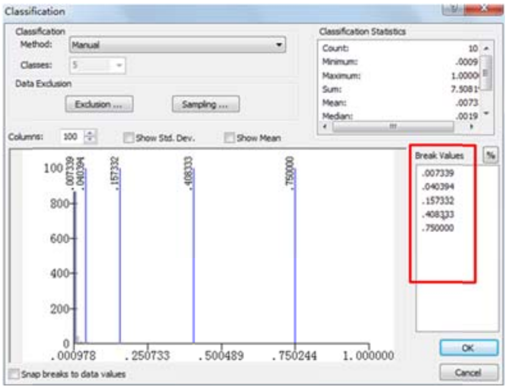
Natural Breaks (Jenks)

Classes: 5

Symbol	Range	Label
◦	.000978 - .022727	.000978 - .022727
●	.022728 - .090909	.022728 - .090909
●	.090910 - .250000	.090910 - .250000
●	.250001 - .500000	.250001 - .500000
●	.500001 - 1.000000	.500001 - 1.000000

Figure 8: Click ***Classify*** and manually set break values for classification.

7. A new window appeared when you click *classify* button. Import above mean values as *break values* and click *ok* (Figure 9). Please make sure these five classes have following interval rule: [minimum, m1], (m1, m2], (m2, m3], (m3, m4], (m4, m5]. Then, new class intervals have been defined. Click *Apply* > *OK*.



Method: Manual

Classes: 5

Columns: 100

Count:	10
Minimum:	.0009
Maximum:	1.0000
Sum:	7.5081
Mean:	.0073
Median:	.0019

Break Values
.007339
.040394
.157332
.408333
.750000

Figure 9: Import mean values calculated above as break values.

Visualize two classification methods

When these two methods have been successfully applied, you need to set an equal symbol (such as point) and color for verification. Besides, you should place them equally in the same width and height by using *data frame properties* under the *layout view*. Figure 10 is a layout view of two classification results. You can easily find out that head/tail breaks method reflects a relatively real phenomenon. In other words, the distribution of point using head/tail breaks is much more natural than Jenks natural breaks.

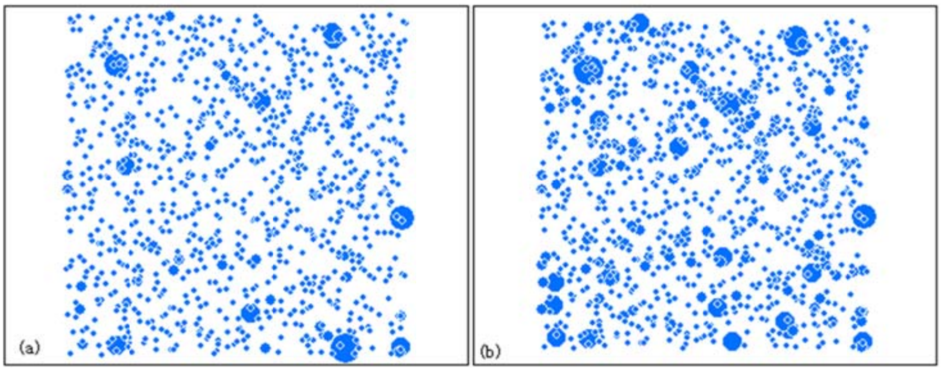


Figure 10: The 1023 points using (a) natural breaks, and (b) head/tail breaks

Acknowledgements

I would like to thank Prof. Bin Jiang and Mr. Ding Ma for their comments and suggestions. A small program developed by Mr. Ding Ma for automatically running the head/tail breaks process is available at https://github.com/digmaa/HeadTailBreaks

References

Jenks G. F. (1967), The data model concept in statistical mapping, *International Yearbook of Cartography*, *7*(1), 186-190.

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. (2014), The fractal nature of maps and mapping, *International Journal of Geographical Information Science*, xx(x), xx-xx, DOI: DOI: 10.1080/13658816.2014.953165