 doi:10.20944/preprints201904.0283.v1

Natural Cities Generated from All Building Locations in America

Bin Jiang
Faculty of Engineering and Sustainable Development, Division of GIScience
University of Gävle, SE-801 76 Gävle, Sweden
Email: bin.jiang@hig.se

Abstract
Authorities define cities – or human settlements in general – through imposing top-down rules in terms of whether buildings belong to cities. Emerging geospatial big data makes it possible to define cities from the bottom up, i.e., buildings determine themselves whether they belong to a city based on the notion of natural cities that is defined based on head/tail breaks, a classification and visualization tool for data with a heavy-tailed distribution. In this paper, we used 125 million building locations – all building footprints of America (mainland) or their centroids more precisely – to derive 2.1 million natural cities in the country (http://lifegis.hig.se/uscities/). These natural cities – in contrast to government defined city boundaries – constitute a valuable data source for city-related research.

Keywords: Head/tail breaks, natural cities, Zipf's law, geospatial big data

1. Introduction

How many cities are there in America? The answer depends on how we define a city. Conventionally, census or statistical authorities legally determine what is qualified to be a city or what buildings to be included in a city, which is essentially a top-down approach. It may need to have a certain number of residents, or within a certain meters for buildings to be part of a city. For example, the United States Census Bureau defines urban area (UA) with population over 50,000, and urban cluster with population less than 50,000. This top-down approach is largely for the purpose of administration and management, but it has been taken for granted for scientific purposes, since there was no alternative. Now, an alternative approach to defining cities – more precisely, human settlements – from the bottom up, has been developed. Here we present a dataset of 2.1 million natural cities in the mainland of the United States of America (http://lifegis.hig.se/uscities/) automatically and naturally generated from a recently released database of 125 million computer-generated building footprints or their centroids to be more precise in the country (Wallace et al. 2018). The number of buildings in this database, created by Microsoft (2018), is three times greater than the number of buildings in OpenStreetMap (Bennett 2010). They may represent the ground truth of buildings in the country.

2. Methodology based on head/tail breaks

Instead of a certain number of residents, we define a city with at least two closely adjacent buildings. How closely adjacent these buildings need to be is collectively determined by all buildings themselves; that is, a bottom-up approach. All building centroid locations or equivalently social media user locations are used to construct a huge triangulated irregular network (TIN) with over 10^6 nodes, and all the TIN edges shorter than their mean – actually high-density edges – constitute so-called natural cities (Jiang and Miao 2015). This way of defining or deriving cities is substantially inspired by the fundamental thinking of the wisdom of crowds (Surowiecki 2004). That is, the diverse and heterogeneous many are often smarter than the few (even a few experts) and collectively the many can make a smart decision on the cities, which we called natural cities. The 125 million buildings constitute huge crowds and collectively decide (1) with whom to be paired – the edges of the TIN – and (2) which buildings constitute natural cities. The second decision process is actually based on head/tail breaks: a classification and visualization tool for data with a heavy-tailed distribution (Jiang 2013, 2015). It is essentially a recursive function, which makes partition of a dataset of some heavy-tailed distribution – around the mean – into the head for those greater than the mean, and the tail for those less than the mean, and continue iteratively for the head until the remaining head is no longer heavy-tailed, e.g., the head percentage is greater than 40 percent. The following function was used:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 April 2019 doi:10.20944/preprints201904.0283.v1

Recursive function of head/tail breaks

```
Function Head/tail Breaks:
     Break a whole into the head and the tail;
     // the head for those above the mean
     // the tail for those below the mean
     While (head <= 40%):
         Head/tail Breaks (head);
End Function
```

To further illustrate head/tail breaks, let us assume a dataset consisting of 10 numbers: 1, 1/2, 1/3, …, and 1/10, which follows precisely Zipf's law (1949): the first largest city is twice as big as the second largest, three times as big as the third largest, and so on. For the 10 numbers the mean is 0.29, which partitions the 10 numbers into two parts: the first three, as the head part (accounting for 30 percent) are greater than the mean, and the remaining seven (70 percent) as the tail that are less than the mean. For the three in the head, the mean is 0.61. This mean value further partitions the three largest numbers into two parts: 1 (33 percent), as the head, is greater than the mean 0.61, and 1/2 and 1/3 (67 percent) as the tail that are less than the mean. In the end, the 10 numbers are divided into three classes: [1], [1/2, 1/3], [1/4, 1/5, …, 1/10].

3. Results and discussion

The major result of this paper is the 2.1 million natural cities derived from the database of 125 million building footprints. Different slightly from those previous studies on natural cities (e.g., Jiang and Miao 2015), we consider individual triangles – instead of previously TIN edges – as the basic units. In other words, we applied head/tail breaks into about 300 million triangles out of the TIN built up from the 125 million buildings, or more precisely their centroids. This head/tail breaks process goes for three iterations, thus leading to three means, but we chose the second mean – about 1000 meters – as the cutoff for deriving 2.1 million natural cities, part of which are shown in Figure 1. All triangles less than the mean constitute individual natural cities, while the remaining triangles greater than the mean are considered to the space between the natural cities, which is defined as the countryside. Thus, under the notion of natural cities, a building belongs either to one of these natural cities, or to the countryside between these natural cities. This bottom-up way of defining natural cities and countryside is very different from conventional cities and countryside. These 2.1 natural cities or settlements have their inherent hierarchy. If we run head/tail breaks, it would lead to 10 hierarchical levels, unlike the conventional hierarchy subjectively or arbitrarily defined by authorities such as mega cities, large cities, middle cities, small cities, towns, and villages, or urban areas and urban clusters. These different names are convenient for administrative and management purposes, but not always so for scientific purposes.

We examined power law distribution using the maximum likelihood method (Clauset et al. 2009), and found that these derived natural cities – by the second mean – follow Zipf's law (1949) very well, with the Zipf's component 1.0, a goodness fit p value of 0.25. In comparison to the natural cities by the first mean, the Zipf's component was around 0.89 with p value of 0.06. This is a clear evidence that the 2.1 million natural cities represent the best result, which could be the ground truth of cities pattern in the country. The reader may ask how each derived natural city differs or resembles UA, which is defined by the government using some very complicated way for the 2010 census, so complicated that makes the comparison virtually impossible. For example, there is no individual people locations available (due to privacy concern) to replicate the delineation process of UA. Given the circumstance, it makes little sense for such a comparison. If city boundaries were defined following some objective rule, like in Sweden (Haldorson and Daher 2016) using a distance between buildings (e.g., 500 meters) as the threshold to determine whether a building belongs to a city, then the natural cities would fit the conventional cities very well only for those largest natural cities. Conventional cities do not refer to all human settlements, only those large ones, while the natural cities do refer to all human settlements. This is another major gap between conventional cities and natural cities. If city boundaries were defined by following some arbitrary rule, like in China, one day government wanted to include a large area as part

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 April 2019 doi:10.20944/preprints201904.0283.v1

of a city, not to say that the government wanted to create a major city from scratch. In this case, natural cities even those largest are hardly comparable to conventional cities.

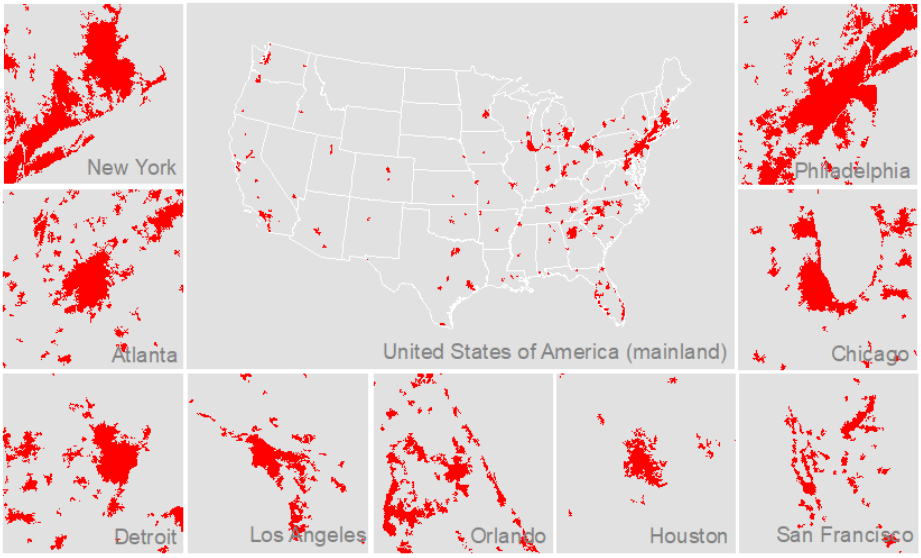

Figure 1: (Color online) Selected natural cities generated from all building locations in America (Note: This figure shows the overall configuration of the natural cities by the 155 largest (the central panel) and a few largest (the surrounding panels). To browse these natural cities, the reader is encouraged to visit: http://lifegis.hig.se/uscities/, with which you can zoom in/out to see all the 2.1 million natural cities at different levels of scale. Alternatively, one can download the data (Jiang (2019) for further research.)

Seen from the discussion here, it is very clear that there is no universal way of defining conventional cities, not only across countries, but also within a country. In contrast, natural cities are universally and globally defined, so they are more scientific – reflecting better the cities pattern – than conventional cities. That is the reason that previous studies (e.g., Jiang et al. 2015) found that natural cities fit to Zipf's law perfectly, not only at country scale, but also at continental and global scales. That is, statistically the largest city is twice as big as the second largest, three times as big as the third largest, and so on. Furthermore, with help of natural cities, it was found that not only city sizes, but also city numbers follow Zipf's law. That is, statistically the number of cities in the largest country is twice as high as that in the second largest country, three times as high as that in the third largest country, and so on (Jiang et al. 2015). To this point, it should be noted that Zipf's law as a universal law should be used to compare natural cities and conventional cities, not at an individual level, but collectively, to see how natural cities fit better Zipf's law than conventional cities. This fact that natural cities better fit Zipf's law is obvious, since natural cities refer to all human settlements, while conventional cities refer to only those large ones.

4. Conclusion and further discussion

This paper describes a dataset of 2.1 million natural cities automatically and naturally generated from the 125 million building footprints. The database of the building footprints certainly represents – to a great extent – the ground truth of building distributions in the country, so does the derived dataset of natural cities or human settlements as the ground truth of cities pattern. With this short paper, and the shared dataset in particular, we intend to further disseminate the concept of natural cities, which represents a new way of thinking – bottom-up in nature – for urban studies and cities related research. The shared data constitute an important data source or evidence for city related research in the future. For example, there are far more small cities than large ones, so natural cities are fractal, according to the third definition of fractal (Jiang and Yin 2014). Even according to the second definition of fractal (Mandelbrot 1982), these natural cities are fractal, for the city boundaries look much more fragmented or irregular than those defined by standard top-down approaches. These derived cities should be more correctly considered to be as living structures – as defined and discovered by Alexander (2002–2005)

– for well predicting human activities (Ren et al. 2019). As the dataset is publicly available, interested readers are encouraged to use it in future research.

Acknowledgement
XXXXXXXX

References:

Alexander C. (2002–2005), *The Nature of Order: An essay on the art of building and the nature of the universe*, Center for Environmental Structure: Berkeley, CA.

Bennett J. (2010), *OpenStreetMap: Be your own cartographer*, PCKT Publishing: Birmingham.

Clauset A., Shalizi C. R., and Newman M. E. J. (2009), Power-law distributions in empirical data, *SIAM Review*, 51, 661–703.

Haldorson M. and Daher K. B. (2016), Översyn av metod och definition för:SCBs avgränsningar av koncentrerad bebyggelse (Review of method and definition for: Statistics Sweden's delimitations of concentrated buildings), https://www.scb.se/Statistik/_Publikationer/MI0810_2015A01_BR_MIFT1601.pdf (accessed on April 15, 2019)

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. (2015), Head/tail breaks for visualization of city structure and dynamics, *Cities*, 43, 69-77. Reprinted in Capineri C., Haklay M., Huang H., Antoniou V., Kettunen J., Ostermann F., and Purves R. (editors, 2016), *European Handbook of Crowdsourced Geographic Information*, Ubiquity Press: London, 169–183.

Jiang B. (2019), Natural cities in America, DOI: 10.13140/RG.2.2.30027.23847, https://www.researchgate.net/publication/332465822_Natural_cities_in_America

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. Reprinted in Plaut P. and Shach-Pinsly D. (editors, 2018), *ICT Social Networks and Travel Behaviour in Urban Environments*, Routledge.

Jiang B. and Yin J. (2014), Ht-index for quantifying the fractal or scaling structure of geographic features, *Annals of the Association of American Geographers*, 104(3), 530–541.

Jiang B., Yin J. and Liu Q. (2015), Zipf's Law for all the natural cities around the world, *International Journal of Geographical Information Science*, 29(3), 498–522.

Mandelbrot B. (1982), *The Fractal Geometry of Nature*, W. H. Freeman and Co.: New York.

Microsoft (2018), Computer generated building footprints for the United States https://github.com/Microsoft/USBuildingFootprints/

Ren Z., Seipel S. and Jiang B. (2019), Capturing and predicting human activities using building locations in America, A manuscript under review.

Surowiecki J. (2004), *The Wisdom of Crowds: Why the Many Are Smarter than the Few*, ABACUS: London.

Wallace T., Watkins D., and Schwartz J. (2018), A map of every building in America, *The New York Times*, Oct. 12th, 2018.

Zipf G. K. (1949), *Human Behaviour and the Principles of Least Effort*, Addison Wesley: Cambridge, MA.