Demarcating Natural Cities Based on Massive Street Nodes for Validating Zipf's Law
(Extended abstract from full paper http://arxiv.org/abs/1006.0814)

Bin Jiang[(1)] and Tao Jia[(2)]

[(1)] Department of Technology and Built Environment, Division of Geomatics
University of Gävle, 801 76 Gävle, Sweden
Email: bin.jiang@hig.se,

[(2)]Future Position X, Box 975, 801 33 Gävle, Sweden
Email: jiatao83@163.com

OpenStreetMap (OSM) is a wiki-like collaboration, or grassroots movement, to create a free editable map of the world using data from portable GPS units, aerial photography, and other free sources (c.f., Haklay and Weber 2008). It has become a phenomenal success as one of the best examples of volunteered geographic information contributed by individuals and supported by Web 2.0 technologies (Goodchild 2007). Both the OSM data size and the geographic coverage have been growing dramatically over the past few years, thanks to the continuously growing OSM community. It provides an extraordinary geospatial data source for developing web mapping services (e.g., www.fromtomap.org) and conducting related research. For the first time in human history, researchers can acquire an entire world street data set for data-intensive geospatial computing.

In this paper, we will argue that OSM data constitute an extraordinary data source for geospatial research. In many respects, the data sources are much rich and better than census data and satellite imagery. The example we present here is how to demarcate cities or city boundaries based on massive street nodes extracted OSM data for validating Zipf's law. Street nodes include both street intersections and street ends. Conventionally, cities or city boundaries can be obtained from the census data, e.g., metropolitan areas, census-designated places, and urban areas are proxies of cities. However, they are criticized for being arbitrary. More critically, not all population concentrations are counted due to states law. Another possible way of extracting cities is the use of satellite imagery, but it involves sensitivity or uncertainty in image process or information extraction due to for example cloud.

We define cities according to the density of street nodes. This is a natural way – a bottom up approach to defining cities, so the cities are called natural cites. The definition of the natural city is based on the fact that human activities are constrained to streets – no streets no human activity, or alternatively no street nodes no residential places or cities. This way we can extract all places from the largest megacity with 8 million residents to the smallest town with a single person. This is practically possible, since the smallest town must have at least one street junction or street end. Our approach involves no bias or sensitivity imposed by the census data and satellite imagery.

We wrote a little program to extract the 24,657,017 street nodes for the entire country from over 120 gigabytes of street network data of the entire USA. Based on the massive street nodes, we applied the city clustering algorithm (Rozenfeld et al. 2009) to the massive street nodes to obtain individual urban agglomerations or natural cities. This process is very simple and straightforward as described in the following recursive function:

```
Select any street node as current point;
Recursive Function Agglom (current point)
│  Draw a circle with radius (r) around the current point;
│  Search other points within the circle, and add to a point set;
│  If (the point set = empty) Then
│   └ Return;
│  Else
│   │ Pick up any point from the point set as the current point;
│   └ Remove the current point from the point set;
└──Call Function Agglom (current point)
```

The radius in the above function is what we call the clustering resolution. The final result of natural cities relies on the clustering resolution, i.e., the finer the resolution, the more natural cities. Usually the resolutions should be about the size of street blocks, e.g., > 300 meters. Therefore, we chose four resolutions: 400 m, 500 m, 600 m and 700 m for our investigation. The size of natural cities can be measured by the number of clustered street nodes. The size can also measured by the areas of individual natural cities. For this purpose, we need to delimit city boundaries. Figure 1 shows a part of the natural cities from the region near New York.

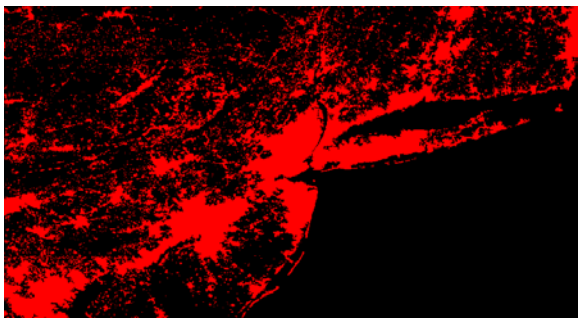

Figure 1: Natural cities near the New York region

The first finding from our study is that there is a long tail for the distribution of natural cities. That is, a vast majority of natural cities are small cities, staying in the tail; while a minority of natural cities are big cities, staying in the head (Figure 2). Given an average size of all cities (m), the corresponding rank R(m) would partition all cities into two categories: those 10% bigger than the average in the head and those 90% smaller than the average in the tail.

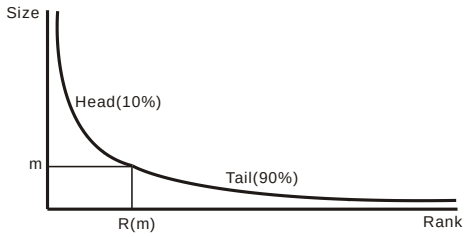

Figure 2: Illustration of a long tail distribution

The second finding is a further refined result from the first. We found that Zipf's law holds remarkably well for all the natural cities with respect to different resolutions: 400 m, 500 m, 600 m, and 700 m. Figure 3 shows a log-log plot, where the straight line stretches over more than two decades.

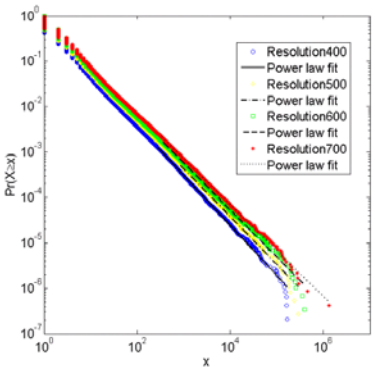

Figure 3: (Color online) Power law distribution with respect to different clustering resolutions
(Note: size x is measured by the number of nodes)

Two contributions of this study that are of interest to the workshop can be summarized as follows. First, we processed over 120 gigabytes of street network data in order to get the 25 million street nodes. Eventually we got over dozens of million of natural cities based on data-intensive computing. It is an extraordinary data source for future research. Second, the findings and insights about Zipf's law constitute another novel aspect of this paper. The validity of Zipf's law has never been examined at such a massive data level and from such a unique geospatial perspective.

References:

Goodchild M. (2007), Citizens as sensors: The world of volunteered geography, *GeoJournal*, 69.4, 211-221.

Haklay M. and Weber P. (2008), OpenStreetMap: user-generated street maps, *IEEE Pervasive Computing*, 7(4), 12-18.

Jiang B. and Jia T. (2010), Zipf's law for all the natural cities in the United States: A geospatial perspective, Preprint, http://arxiv.org/abs/1006.0814

Rozenfeld H. D., Rybski D., Gabaix X. and Makse H. A. (2009), The area and population of cities: new insights from a different perspective on cities, Preprint, http://arxiv.org/abs/1001.5289

A short bio:

Bin Jiang is full professor at University of Gävle, Sweden, affiliated to the Royal Institute of Technology (KTH) at Stockholm via KTH Research School. He received PhD in GeoInformatics from University of Utrecht and International Institute for Geo-Information Science and Earth Observation (ITC), the Netherlands. He worked in the past with The Hong Kong Polytechnic University, and the Centre for Advanced Spatial Analysis of University College London. He is the founding chair of the International Cartographic Association Commission on Geospatial Analysis and Modelling (http://fromto.hig.se/~bjg/ica/). He has been coordinating the NordForsk-funded Nordic Network in Geographic Information Science (http://fromto.hig.se/~bjg/NordGISci/). His research interest is geospatial analysis and modelling, in particular topological analysis of urban street networks in the context of geographic information systems. He has just edited a book entitled "*Geospatial Analysis and Modelling of Urban Structure and Dynamics*" being published by Springer. He is currently an Associate Editor of the international journal *Computer, Environment and Urban Systems*.