Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke

Scaling property of urban systems using an entropy-based hierarchical clustering method

Tao Jia
Royal Institute of Technologhy (KTH)
Valhallavägen 79
Stockholm, Sweden
jiatao83@hotmail.com

Bin Jiang
University of Gävle
Kungsbäcksvägen 47
Gävle, Sweden
bin.jiang@hig.se

Abstract

Urban systems have long been characterized as the scaling property, although there are a lot of argues on the definition and organization of its component or spatial unit. In this paper we propose an entropy-based hierarchical clustering method to aggregate the individual location to form the component or spatial unit. Through the application of the method to three datasets from the different aspects of urban systems, we double check the robustness and consistence of this method. Importantly, it is found that the size of the derived component or spatial unit follows a remarkably power law distribution, which further suggests the scaling property of the underlying urban systems.
Keywords: Scaling property, Urban systems, Power law, Entropy-based hierarchical clustering.

1 Introduction

Urban systems have long been characterized as the scaling property in terms of its component or spatial unit. Here, the component or spatial unit can be referred to individual city, city block [1], named street [2], or axial line [3]. Besides, there are a lot of argues around the definition of the component or spatial unit in the literature [4, 5, 6, 7]. Practically, these different definitions can be summarized into two views: one is to regard it as the arbitrary legal or administrative definition and the other is to derive it with certain spatial rules. However, the fact is that most of the empirical studies have found the scaling property of the urban systems irrespective of the definition [4, 6, 8, 9, 10].

Scaling in general refers to the non-linear relationship between two quantities. When considering it in large urban systems, it means that the small components or spatial units are much more common than the large ones [11]. This kind of structure in essence indicates a heavy tail distribution, or specifically a power law distribution, of the components or spatial units. For example, it is reported that the number of cities is in an inverse power relation with their size, and the size here can be referred to population [8], household income [9], or the spatial extent [6]. In this respect, to verify the scaling property of urban systems, apart from the definition we have to organize the test data in a way. As far as we know, there are three ways to tackle this issue: (1) from the perspective of temporal organization [12], (2) from the perspective of spatial organization in a time slot [10] and (3) from the perspective of both temporal organization and spatial organization in a time slot [4].

This paper proposes an entropy-based hierarchical clustering method to aggregate the individual spatial location to form the component or spatial unit in the urban systems. In this way, it conforms to the second perspective elaborated above in terms of the definition. Moreover, we also delve into the second view proposed above according to the organization of the data. That is, we concentrate on the spatial organization of the data in a time slot. Particularly, three datasets from different aspects of the urban systems are examined, namely the street nodes [13] dataset from the OpenStreet Map (OSM) project [14], which reflects the image of the underlying urban infrastructure supporting the associated urban function; the static points dataset from taxi GPS records, which is considered as a proxy of human activity; and the photo locations dataset from the Flickr website, which is highly associated with the spatial points of interest (POIs). Findings from these datasets are analyzed and reported, which suggests a remarkable scaling property of urban systems.

The rest of this paper is organized as following. We elaborate the entropy-based hierarchical clustering method in section 2. In section 3 we apply this method to three different datasets and present the findings. Finally, we draw a conclusion in section 4.

2 Methodology

The entropy-based hierarchical clustering method is mainly composed of two parts with a recursive character: (1) Decomposition of Triangular Irregular Network (TIN) based on head-tail division rule [5] to obtain a level of clustering; and (2) Selection of the best clustering level with the aid of maximum entropy change.

2.1 TIN Decomposition

It is well known that TIN not just can be used to represent the surface of 3D data vividly but also can be employed to organize the distribution of 2D point data to form a triangular tessellation. This non-overlapping tessellation ensures the plausibility of the spatial relationship among the data points, and hence it hints us to decompose the TIN into several sub

TINs by removing the long edges greater than the threshold value.

To determine the threshold value in this decomposition process, we apply the head-tail division rule which states that given a *variable V, if its values v follow a heavy tail 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* [5]. Following this rule, if the edge length of a TIN follows a heavy tail distribution, then the edges with length greater than the mean value will be removed and the others reserved. However, in practice, a TIN is reserved if it is still connected (which means you can navigate from any node to any other node) even if by removing the long edges. Otherwise, it is divided into several sub TINs. To this end, both the reserved TINs and the sub TINs constitute the current level of clustering (*c.f.* figure 1).

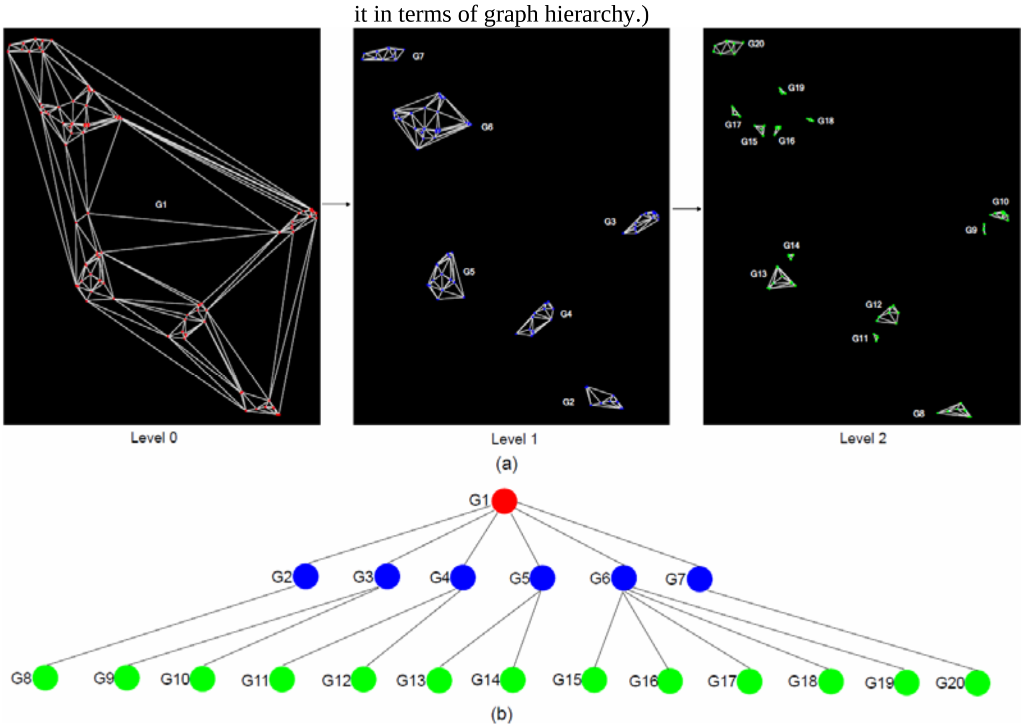
Where I(X) is the information X contains and p is the probability mass function of X. Here, a simple strategy is adopted to calculate the entropy value for each level of clustering. That is, there is little information if all the locations belong to the same cluster, whereas there is the maximum information if each location forms its own cluster. Therefore, the definition is given as (2).

$$H_k = -\sum_{i \in N_k} p_i^k * \log p_i^k \, , \; p_i^k = n_i^k \Big/ \sum_{i \in N_k} n_i^k \qquad (2)$$

Where n_i^k is the number of locations belonging to cluster i in level k, p_i^k is the probability of a location belonging to cluster i in level k, and N_k is the number of clusters in level k.

The maximum entropy Log(N) is reached in theory when

Figure 1: A demonstration of 2-level decomposition of TIN based on head-tail division rule (Note: In this demonstration, a total of 100 levy points are generated; (a) shows the two levels of clusters; (b) shows it in terms of graph hierarchy.)



2.2 Entropy change calculation

To calculate the entropy change between successive levels of clustering, we have to obtain the entropy value for each level of clustering firstly. Entropy is typically referred to the Shannon (1948) information entropy (for more detail, interested readers are suggested to refer [15]), which is mathematically defined as (1).

$$H(X) = E\big(I(X)\big) = -\sum p(x) * \log p(x) \qquad (1)$$

each location forms a cluster. However, it is meaningless since it loses the initial purpose. Therefore, the attention here is focused on the difference between the entropy values of consecutive levels of clustering. And it is defined as $\Delta H_{k-1}^{k} = H_k - H_{k-1}$, where H_k is the entropy value for the kth level of clustering and H_{k-1} is the one for the (k-1)th level. The entropy change allows us to find the level of clustering which has the most dramatic change compared with the preceding one. Intuitively, this level of clustering with the largest entropy change represents the best clustering solution, because it not just achieves the relative high information but also avoids the process of over splitting which obscures the initial purpose. In reality, it reflects the best organization of spatial units from random (maximum entropy) to regular (minimum entropy). To this point, the best level of clustering k can be derived as (3).

$$k = \{\max(\Delta H_{k-1}^{k}) \mid k = 1,2,...\} \quad (3)$$

3 Scaling property in urban systems

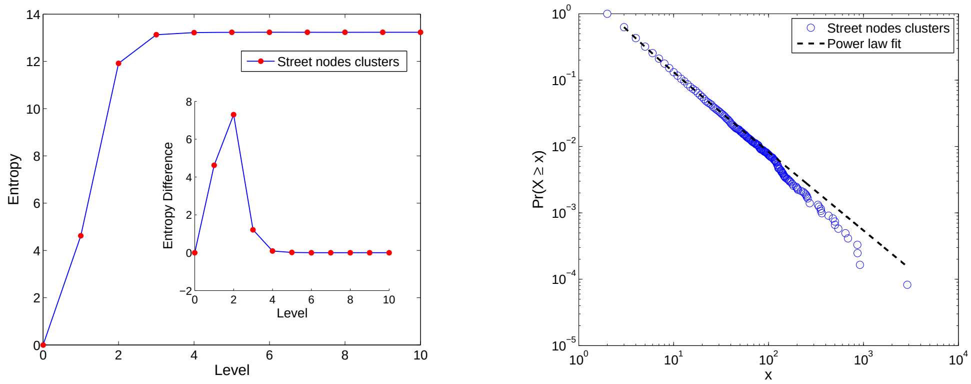
In this section, we apply the entropy-based hierarchical clustering method to three different datasets, namely street nodes from OpenStreet Map (OSM) project, static points from taxi GPS records and photo locations from Flickr website. The three datasets from different aspects of urban systems double-check the robustness and consistence of our method, and importantly the subsequent results suggest a remarkable scaling property of urban systems.

3.1 Street nodes

Street nodes dataset, which has a total number of 105648, is extracted from the OSM project and covers the whole city of Stockholm, Sweden. It includes not only the street intersection points but also the street originals and destinations [13]. From this point, it reflects the image of the underlying urban infrastructure which supports the associated urban function and human activity. By applying our method to the city level street nodes dataset, we can classify the street nodes into different clusters which is similar to the formation of natural cities using a country level street nodes dataset [10].

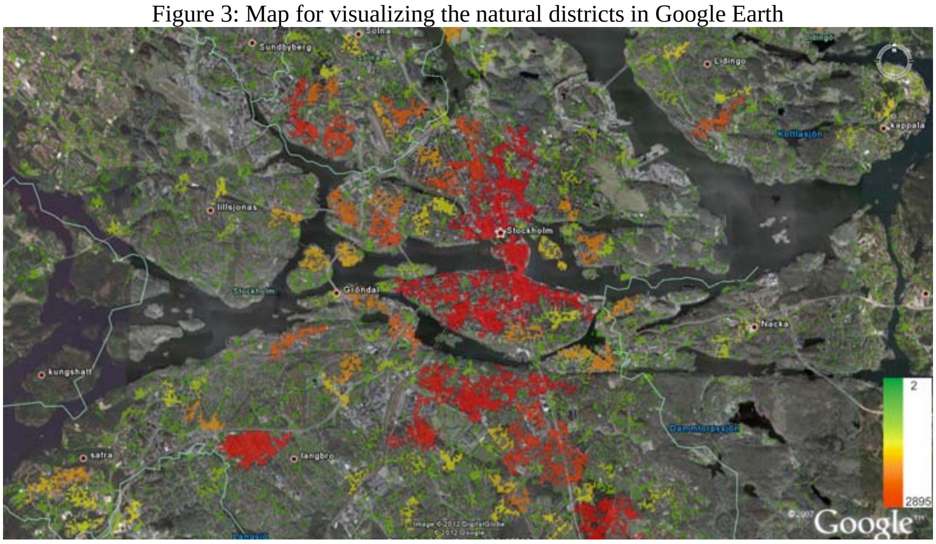
As shown in figure 2a, the clusters in level two have the maximum entropy change, and thus we coin these clusters as the natural districts. The size of these natural districts is further found to demonstrate a remarkable power law distribution with P value reaching 0.2 which pass the KS test

Figure 2: Plot for the result from street nodes dataset
(Note: (a) shows the entropy and entropy change for each level of clustering; (b) shows the power law plot for the clusters in level 2 with alpha equals 2.2)



(a) (b)

Figure 3: Map for visualizing the natural districts in Google Earth



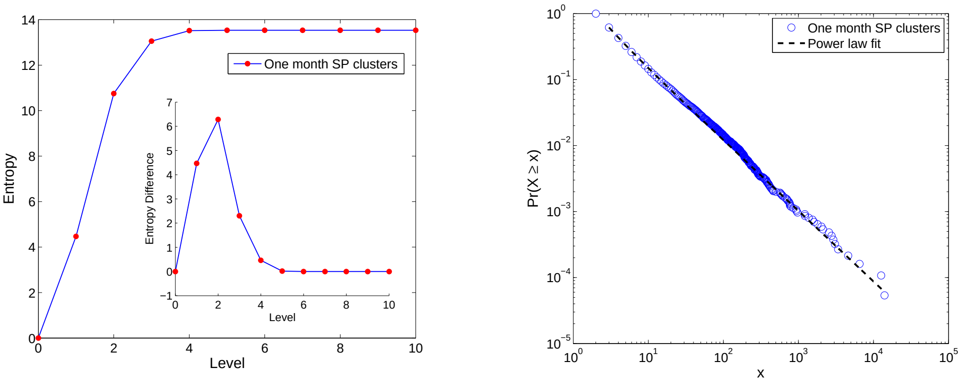
Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke

[16] significantly (*c.f.* figure 2b). This fact tells the scaling property of urban systems in terms of the natural districts. Moreover, we visualize these natural districts in Google Earth shown in figure 3, which clearly depicts their spatial distribution in the city of Stockholm, Sweden. Subsequent consistent check with the real district is lacked for the absence of the real data, but we are pretty sure that it should reflect the true situation.

by applying our method to the static points dataset, we can aggregate static points into different clusters which reflect the hot spots in terms of human activity in the geographic space.

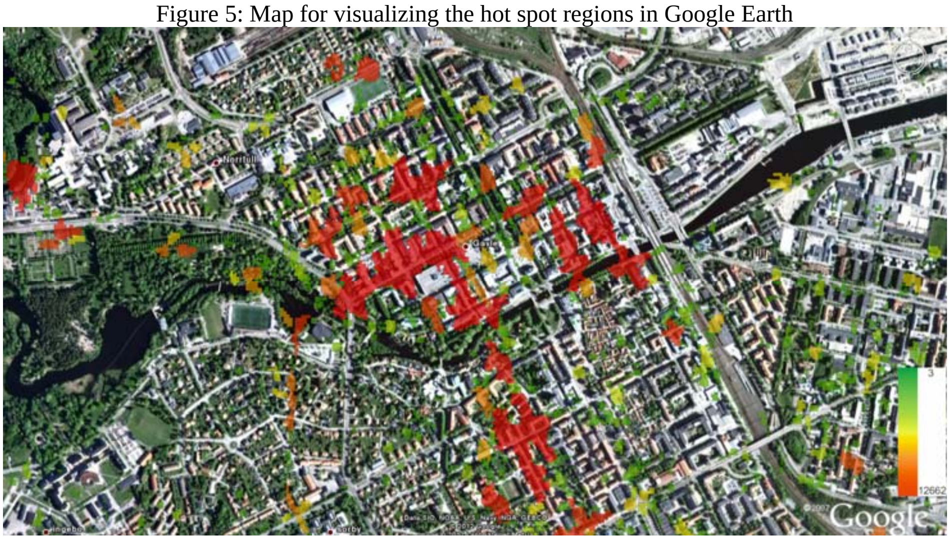
We show the findings in figure 4, which demonstrate the same pattern as we observed in the previous example. That is, we clearly obtain the best clusters with respect to level two that has the maximum entropy change (*c.f.* figure 4a), and the size of the clusters follow a power law distribution with P value equaling to 0.1 which pass the KS test [16] significantly

Figure 4: Plot for the result from static points dataset
(Note: (a) shows the entropy and entropy change for each level of clustering; (b) shows the power law plot for the clusters in level 2 with alpha equals 2.1)



(a) (b)

Figure 5: Map for visualizing the hot spot regions in Google Earth



3.2 Static points

Static points dataset, which has a total number of 316209, is obtained from the taxi GPS records lasting for one month in Gävle, Sweden. As its name indicates, it refers to the location with zero speed along the tracking trajectory. In fact, these points typically represent the locations where the taxicabs stop for a moment in the case of customer destination, traffic intersection, parking lot, etc. From this perspective, they can be roughly accepted as a proxy of human activity data. Again,

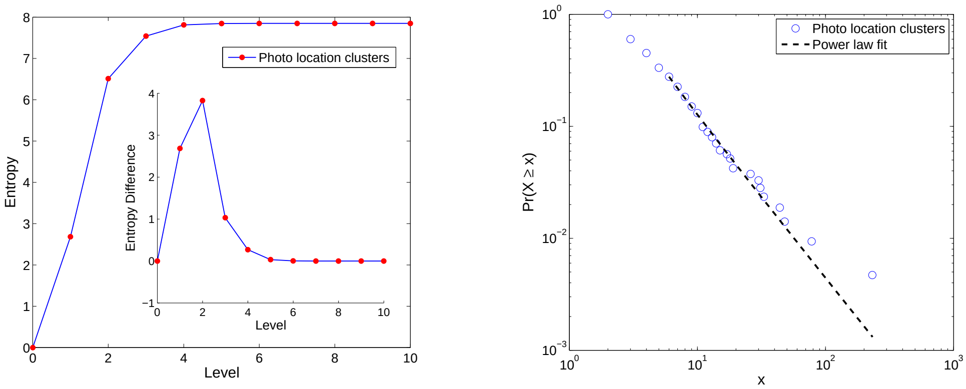
(*c.f.* figure 4b). This finding further strengthens our understanding on the scaling property of urban systems in terms of human activity. As shown in figure 5, the human activity clusters are visualized in Google Earth. From the map, we can observe the hot spot regions in Gävle, Sweden, say the Nygatan Street which gathers several shopping malls, the Gävle train station, etc.

3.3 Photo locations

In this last case, we adopt two months photo locations dataset with a total number of 124,738 in London, which is retrieved from the Flickr website. Some of the Flickr users may upload their pictures to the website together with geotags which tell the exact locations where they have been taken. Although the number of geotagged photos has roughly approached to 172 million, it is still rising with a fast speed. Thus, this massive dataset gives us a fresh look at the dynamics in urban systems in general, and particularly, it allows us to mine both the human activity and the popular landscapes.

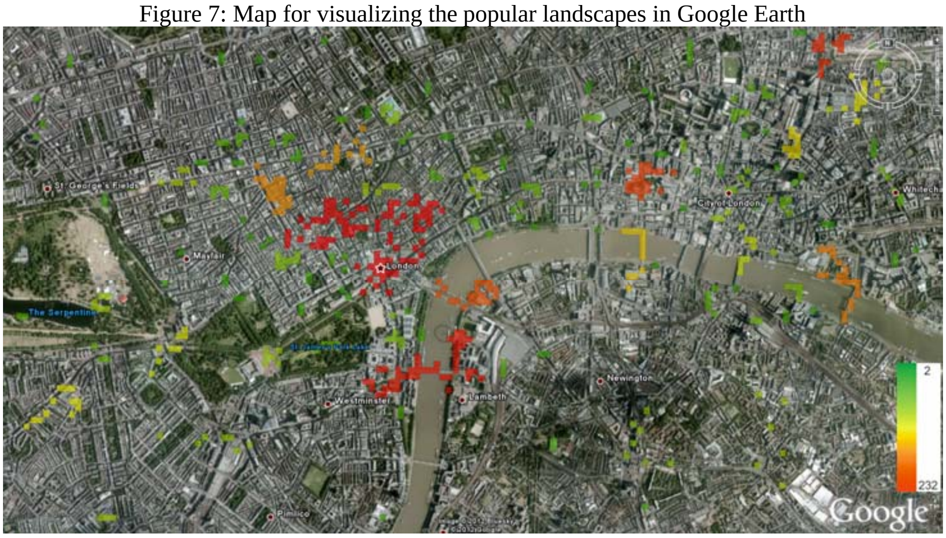
The clusters obtained through the application of our method to the photo locations dataset display the same pattern as the previous examples. That is, the best clustering is derived in level two (*c.f.* figure 6a), and the size of the clusters obey a good power law distribution with P value [16] as high as 0.7 (*c.f.* figure 6b). Besides, these clusters are highly associated with the man-made or natural landscapes. For instance, we can clearly see several red clusters (*c.f.* figure 7) known as famous landscapes in London, such as the Big Ben, the British Airways London Eye, and the Trafalgar Square. Therefore, we can conclude that the landscapes in urban systems display the scaling property, which further assumes the preferential rule in forming their popularity.

Figure 6: Plot for the result from photo locations dataset
(Note: (a) shows the entropy and entropy change for each level of clustering; (b) shows the power law plot for the clusters in level 2 with alpha equals 2.4)



(a) (b)

Figure 7: Map for visualizing the popular landscapes in Google Earth



4 Conclusion

We propose an entropy-based hierarchical clustering method in this paper, and this method can recursively classify the location data with increasing levels and further automatically suggest the best clustering level based on the maximum entropy change. By applying this method to three datasets from different aspects of the urban systems, we test the robustness and consistence of the method, and subsequently we conclude the scaling property of urban systems empirically.

Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke

However, it is important to note that the datasets adopted are organized spatially and do not contain the temporal development information. Actually, the dataset including both the spatial organization and the temporal development is of great significance in verifying our method and exploring the urban systems. In this respect, it suggests our future work should be focused on examining the effect that the scaling property on the evolution of urban systems.

Acknowledgement

We thank the constructive comments from the three anonymous referees.

References

[1] Lämmer S., Gehlsen B. and Helbing D. (2006), Scaling laws in the spatial structure of urban road networks, *Physica A: Statistical Mechanics and its Applications*, 363(1), 89-95.

[2] Jiang B. and Claramunt C. (2004), Topological analysis of urban street networks, *Environment and Planning B: Planning and Design*, 31(1), 151-162.

[3] Hillier B. (1996), *Space is the Machine: A Configurational Theory of Architecture*, Cambridge University Press: Cambridge, UK.

[4] Batty M. (2006), Rank clocks, *Nature*, 444, 592–596.

[5] Jiang B. and Liu X. (2010), Automatic generation of the axial lines of urban environments to capture what we perceive, *International Journal of Geographical Information Science*, 24(4), 545-558.

[6] Rozenfeld H.D., Rybski D., Gabaix X. and Makse H.A. (2011), The area and population of cities: New insight from a different perspective on cities, *American Economic Review*, 101, 2205-2225.

[7] Jia T. and Jiang B. (2011), Measuring urban sprawl based on massive street nodes and the novel concept of natural cities, Preprint, arxiv.org/abs/1010.0541

[8] Rosen K.T. and Resnick M. (1980), The size distribution of cities: An examination of the Pareto law and primacy, *Journal of Urban Economics*, 8(2), 165-186.

[9] Reed W.J. (2003), The Pareto law ofincomes—an explanation and an extension, *Physica A: Statistical Mechanics and its Applications*, 319, 469-486.

[10] Jiang B. and Jia T. (2011), Zipf's law for all the natural cities in the United States: a geospatial perspective, *International Journal of Geographical Information Science*, 25(8), 1269-1281.

[11] Jiang B. (2010), Scaling of geographic space and its implications, in: *Las Navas 20th Anniversary Meeting on Cognitive and Linguistic Aspects of Geographic Space*, Las Navas del Marques, Avila, Spain, July 5 - 9, 2010.

[12] Chen Y.G. (2011), Allomtric scaling analysis for city development: model, method, and applications, arXiv: 1105.5445

[13] Jia T. (2010), *Exploring Massive Volunteered Geographic Information for Geographic Knowledge Discovery*, Licentiate Dissertation, Royal Institute of Technology, Stockholm, Sweden.

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

[15] Shannon C.E. (1948), A mathematical theory of communication, *Bell System Technical Journal*, 27, 379-423, 623-656.

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

Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke