Study on Spatial Dependence of Information on Internet

Su Jinyan[1]

[1] *sujinyanyc@hotmail.com*
Centre for Documentation and Information, Chinese Academy of Social Sciences, Beijing 100732 (China)

Abstract
The purpose of this paper is to analyse the spatial dependence of information on Internet. Methodological tools taken from spatial economics are employed to test the hypothesis that spatial proximity might to affect the spatial characteristics of information on Internet, though many people consider that the distance is disappearing with the using of Internet. Results show that information on Internet in geographically adjacent regions has similar distribution trends, confirming the presence of spatial dependence.

Introduction
The first law of geography according to Waldo Tobler (1970) is "Everything is related to everything else, but near things are more related than distant things." That is, there is spatial interaction, named spatial dependence or spatial effects. On the one hand, it was argued that the Internet, due to its numerous advantages, would reduce the role played by spatial proximity. For example, Cairncross (1997) and Thomas (2005) considered that the distance was disappearing and the world was becoming more and more flat. Björneborn (2004) found small-world link structures across an academic web space. According to this view, it seems that the geographic distance cannot effect the spatial distribution of information on Internet and thus spatial dependence does not exist when the spatial distribution of information on Internet is analyzed. On the other hand, it is found that the spatial distribution of information on Internet, actually, is influenced by spatial proximity. For instance, Holmberg and Thelwall (2009) show that interlinking between local government bodies in Finland follows a strong geographic, or rather a geopolitical pattern and that governmental interlinking is mostly motivated by official cooperation that geographic adjacency has made possible. For another instance, Kawamura, Otake and Suzuki (2009) analyzed the relationship of Japanese public libraries and found that the connection among libraries which are geographically close to each other or in the same administrative unit is strong.
Center for Spatially Integrated Social Science (CSISS) was funded in 1999 with support from the National Science Foundation in United States of America. Its mission is to recognize the growing significance of space, spatiality, location, and place in social science research. The purpose of this paper is to provide empirical evidence on the spatial dependence of information on Internet, using data collected by search engines. If it is true that the spatial distribution of information on Internet can be influenced by spatial proximity, data observations would not really independent when we analyze spatial distribution issues, thus independence of samples should be reconsidered.

Background
Spatial dependence is a fundamental concept of spatial econometrics, that nearby entities often share more similarities than entities which are far apart. It has been some 30 years since Paelinck and Klaassen (1979) published a small volume entitled spatial econometrics, which arguably was the first comprehensive attempt at outlining the field of spatial econometrics. Its distinct methodology and spatial dependence analysis methods and models were stated detailed by Anselin in his book (1988). The influence of spatial dependence is considered in more and more research areas (e.g. education area and science of science area). The economic geography literature has studied the spatial distribution of Internet adoption showing the emergence of the so-called *Internet Geography*. Billon et al. (2009) provided empirical

evidence on the neighboring effects of Internet adoption as measured by the percentage of firms with their own website in the European regions. An uneven spatial pattern of Internet distribution has been shaping (Malecki & Boush, 2003; Zook, 2006).
Although it seems to be a surge in studies that address spatial aspects of science (e.g. Batty, 2003; Glänzel, Schubert, & Braun, 2002; Katz, 1994), little attention has been paid to the study of the possible relevance of spatial proximity when explaining spatial characters of information. The phrase, *spatial scientometrics*, was pointed out by Frenken, et al. (2009), and spatial distribution, spatial biases and citation impact are the main contents of spatial scientometrics. There are some same studies in China, but these studies are usually in *Regional Economics*. Analysis methods of spatial dependence were introduced by Pang in 2006 (Pang, 2006) and many researchers using the theory of it studied problems bring by spatial dependence in China (e.g. Wu, 2007; Dai & Chen, 2010; Wang, Ding & Zhu, 2010).

Data Collection

We approached our research questions by examining the frequency distribution of keywords over pages in 194 countries in the world, 51 states in America and 33 provinces in China (including Hong Kong and Taiwan, excluding Macao).

Selection of Keywords

Allowing for the differences among research fields, keywords of three fields in science and social science are selected. They are Nano field in *Physics*, AIDS field in *Medicine* and digital library field in *Library and Information Science*. Three or four keywords of each field are chose. The selected keywords in English are Nanometer, Nano technology, Nano materials, AIDS, HIV, acquired immune deficiency syndrome, human immunodeficiency virus, digital library, electronic library and virtual library. These keywords are translated in ten languages, English, Chinese, Spanish, Japanese, Portuguese, German, Arabic, French, Russian and Korean, and thus there are 94 keywords in total. According the data of Internet World Stats, these languages are top ten languages in Internet, and Internet users of them is 82.2 per of total Internet.

Choice of Search Engines

All searches are carried out through commercial search engines, because commercial search engines are now playing an increasingly important role in Web information dissemination and access, and no crawler from a single researcher or even a rather large research team can cover the vast Web space as these search engines can (Vaughan & Thelwall, 2004). Windows Live and Yahoo! are chose in this study, because they are all major commercial search engines which not only have high search market share points, but also have Application Programming Interface key provided now (API). With API key, URLs which searched by search engine can be download in bulk. The maximum number of results retrieved by the APIs is less than 1000. We don't choose Google due to Google didn't provide API key from 2006. File types of information on Internet are limited in txt, pdf, doc, ppt and xls. Though html is the most common file type, it is not included. At last, 46628, 62984 and 52468 URLs respectively in Nono area, AIDS area and digital library area are collected, compmosing the samples of this research. Certainly, the result could be influenced by search engines, because of file types we chose and the maximum number of results retrieved by the APIs.

Methodology

Spatial Dependence Analysis Methods

Gittleman and Kot (1990) proposed to use Moran's I to test for spatial dependence. Moran's I tests for global spatial dependence in group-level data. When rates in nearby areas are similar, Moran's I will be large and positive. A significant and positive value of the statistic will indicate the existence of positive spatial dependence, while a significant and negative value of Moran's I will reflect the presence of a pattern of spatial association between dissimilar values. Morans's I global test of spatial dependence defined as:

$$Moran's\ I = \frac{n}{s_0} \times \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (1)$$

Where, in the present context, y_i stands for the number of information on Internet in region *i*, and, $\bar{y}$ is the sample average. w_{ij} is spatial weight matrix, namely the weight between observation i and j, and s_0 is the sum of all w_{ij} s, $s_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$. The spatial weight matrix is the basement to analysis spatial dependence, which has zero diagonal elements and is usually row-normalized. In particular, we considered w based on the 4 nearest neighbors, calculated using the geographic distance between the corresponding regional centroids.

Spatial Dependence Analysis Tools

ArcGIS and GeoDa are analysis tools when we study the spatial dependence of information on Internet. ArcGIS which we make basic maps is a Windows suite consisting of a group of geographic information system (GIS) software products produced by Esri. GeoDa is a free software package which can be downloaded at http://geodacenter.asu.edu/. It was initially developed by the Spatial Analysis Laboratory of the University of Illinois under the direction of Luc Anselin. GeoDa has powerful capabilities to perform spatial analysis, multivariate exploratory data analysis, and global and local spatial autocorrelation.

Results

Spatial Distribution of Information on Internet

The purpose of this section is to investigate in detail the spatial distribution of information on Internet, measured by number of it in each spatial unit. As can be observed, there are relatively significant disparities in the number of information in different countries (cf. Figure 1). Specifically, the regions with deep red color tend to be located in some of rich countries. Light yellow color regions most locate in Africa, which indicates that most countries of lower number of information on Internet locate in Africa. Obviously, there is an uneven spatial distribution and spatial agglomeration phenomena.

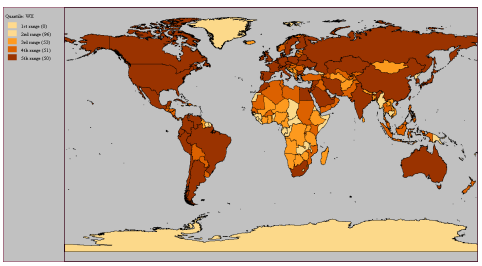
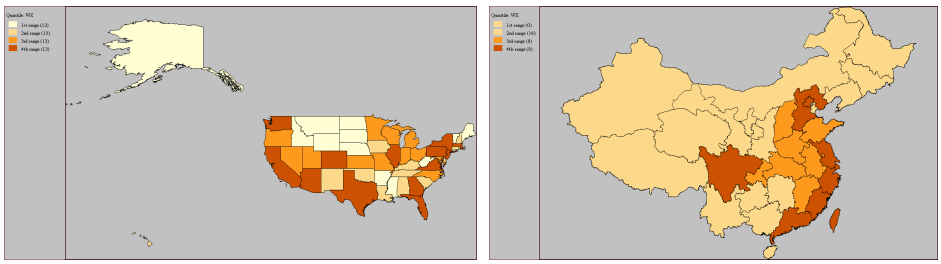


Figure 1. The Spatial Distribution of the Information on Internet in the World (194 Countries)

The spatial distribution of information on Internet in America and China presents in Figure 2. The spatial distribution in America likes a basin (cf. Figure 2(a)), in China a ladder (cf. Figure 2(b)). Major deep color regions distribute on east coast and west coast of the United States and on east coast of China. Light color regions distribute on the central section of America, and on west of China.



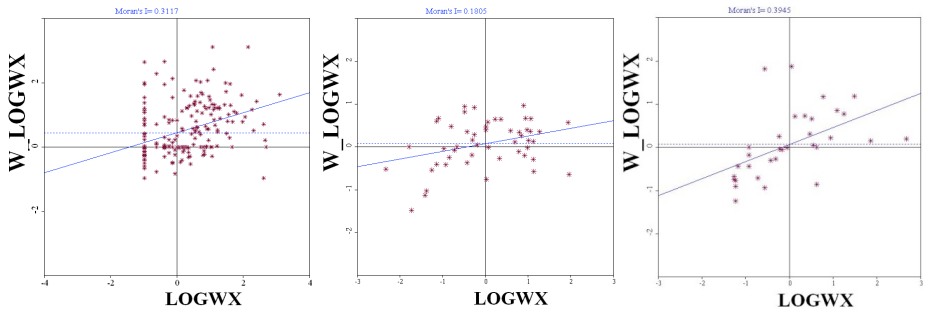
(a) America (b) China

Figure 2. The Spatial Distribution of the Information on Internet in America and China

It is worth mentioning that this initial distribution pattern suggests that the number of information on Internet in different regions is not randomly distributed across space. On the contrary, there seems to be a positive spatial relationship between adjacent areas.

Spatial Dependence Analysis of Information on Internet

As for spatial dependence of information on Internet in the world, the result of the global Moran's I test of it provides us with a value of 0.3117, which is significant at the 5 percent level (cf. Figure 3(a)). *LOGWX* is the number of URLs of each spatial unit. *W_LOGWX* is the spatial lag variable of *LOGWX*. This is clear evidence of the existence of a pattern of positive spatial association in this context, which is consistent with the initial impression drawn from Figure 1. Therefore, we can conclude that information on Internet located in spatially adjacent countries tend to have similar trends. As can be seen from Figure 3(a), most of countries are located in quadrant I. This phenomenon indicates that spatial dependence among countries with high number of information on Internet is stronger than these countries in low levels.



(a) World (194 Countries) (b) America (51 States) (c) China (33 Provinces)

Figure 3. Moran's I Scatterplot of the Information on Internet

Moran's I value of America and China are showed in Figure 3 (b) and (c). Obviously, spatial dependences of information on Internet in America and in China are all existence. The only difference is the value of Moran's I. They are 0.1805 in America and 0.3945 in China. We can conclude that spatial dependence in China is stronger than America. In America, there is geographical agglomeration, that most states are located in quadrants I and III, a few in quadrants II and IV. Regions located in quadrant I are high level cluster and in quadrant III are low cluster. In China, thirty provinces, about 91 percent, locate in quadrants I and III, only two in quadrant II and one in quadrant IV, so spatial dependence in China is much more

serious than in America, even in the world. It is consistent with Figure 3, most provinces with high number of information on Internet located in east of China, most low level provinces located in west of China.

Conclusions

The main goal of this paper has been to analyze the spatial dependence of information on Internet, paying particular attention to the spatial distribution of information on Internet. Our sample consists of 194 countries in the world, 51 states in America and 33 provinces in China. We have employed a set of methodological tools taken from spatial economics that allows us to capture the spatial characteristics of information on Internet and to analyze the influence of spatial proximity. Our findings show that spatial dependence really exists when we deal with issues about spatial distribution of information on Internet. Therefore, independence of sample should be considered when we study problems relate to spatial distribution.

References

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic.

Batty, M. (2003). The geography of scientific citation. *Environment and Planning*, (35), 761-765.

Billon, M., Ezcurra, R., & Lera-Lopez, F. (2009). Spatial effects in website adoption by firms in european regions. *Growth and Change*, 40(1), 54-84.

Cairncross, F. (1997). *The Death of Distance: How the Communications Revolution Will Change Our Lives*. Massachusetts: Harvard Business Press.

Dai, Y., & Chen, C. (2010). Spatial econometric analysis of regional differences in china's construction industry developmen. *Statistics & Information Forum*, 25(5), 53-58.

Frenken, K., Hardeman, S., & Hoekman, J. (2009). Spatial scientometrics: towards a cumulative research program. *Journal of Informetrics*, 3(3), 222-232.

Gittleman, J. L., & Kot, M. (1990). Adaptation: statistics and a null model for estimating phylogenetic efects. *Systematic Zoology*, (39), 227-241.

Glänzel, W., Schubert, A., & Braun, T. (2002). A relational charting approach to the world of basic research in twelve science fields at the end of the second millennium. *Scientometrics*, 55(3), 335-348.

Holmberg, K., & Thelwall, M. (2009). Local government web sites in Finland: a geographic and webometric analysis. *Scientometrics*, 79(1), 157-169.

Internet World Stats. (2009). *Top ten languages in the Internet*. Retrieved October 30, 2009 from: http://www.internetworldstats.com/stats7.htm.

Katz, J. S. (1994). Geographical proximity and scientific collaboration. *Scientometrics*, 31(1), 31-43.

Kawamura, S., Otake, Y. H., & Suzuki, T. (2009). The structure of the hyperlink network formed by the web pages of Japanese public libraries. *Journal of the American Society for Information Science and Technology*, 60(6), 1159-1167.

Björneborn, L. (2004). Small-World Link Structures across an Academic Web Space - A Library and Information Science Approach., Royal School, Denmark.

Malecki, E., & Boush, C. (2003). Telecommunications infrastructure in the Southeastern United States: Urban and rural variation. *Growth and Change*, 34(1), 109-129.

Paelinck, J., & Klaassen, L. (1979). *Spatial Econometrics*. Massachusetts: Saxon House.

Pang, J. (2006). *A Textbook of Spatial Economics*. Beijing: Economic Science Press. (In Chinese).

Thelwall, M., Binns, R., Harries, G., Page-Kennedy, T., Price, L., & Wilkinson, D. (2002). European union associated university websites. *Scientometrics*, 53(1), 95-111.

Thomas L. Friedman. (2005). *The World Is Flat: A Brief History of the Twenty-first Century*. New York: Farrar, Straus and Giroux.

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(2), 234-240.

Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information Processing & Management*, 40(4), 693-707.

Wang, X., Ding, K., & Zhu, X. (2010). Science collaboration network of main research institutes in China: study on web of science. *Studies in Science of Science*, 28(12), 1806-1812.
Wu, Y. (2007). *Spatial Econometric Study on Regional R&D, Knowledge Spillovers and Innovation of China*. Beijing: People's Publishing House. (In Chinese).
Zook, M. (2006). The geographies of the Internet. *Annual Review of Information Science and Technology*, 40(1), 53-78.