Exploring spatial decay effect in mass media and social media: a case study of China

Yihong Yuan

Department of Geography, Texas State University, San Marcos, TX, USA, 78666.
Tel: +1(512)-245-3208
Email:yuan@txtate.edu

Abstract

The development of theories and techniques for big data analytics offers tremendous possibility for investigating large-scale events and patterns that emerge over space and time. In this research, we utilize a unique open dataset “The Global Data on Events, Location and Tone” (GDELT) to analyze the connections between Chinese provinces in mass media, specifically, this study employ the gravity model to analyze the distance decay effects in both mass media and social media data. The results demonstrates that mass media data indicate a weak distance decay effect for Chinese provinces. This study generates valuable input regarding the interpretation regional correlation for the nation of China. It also provides methodological references for investigating international relations in other countries and regions in the big data era.

Keywords: Mass media; GDELT; Geo-tagged social media data; Gravity model; Spatial decay.

1. Introduction

The rapid development of techniques and theories in the big data era have also introduced new challenges and opportunities to analyze a large amount of mass media data available online (Eagle et al. 2009, Liben-Nowell et al. 2005). Compared to social media, traditional mass media is characterized by the significance and aggregated nature of associated events (Liebert and Schwartzberg 1977). As such, mass media data are often suitable for investigating the aggregated pattern of a society.

In addition, the spatial decay effect has been a hot topic in many research fields such as immigration and transportation (e.g., the decay of traffic flows between locations) (Rodrigue et al. 2013). Researchers have employed different models to investigate how distance decay influences the magnitude of interactions between geographic units. Among all potential models, the gravity model is commonly-used due to its effectiveness in predicting the degree of interaction, simplicity of equation, and its ability to deal with flows in both directions (Hardy et al. 2012).

In this research, the open-source dataset “The Global Data on Events, Location and Tone” (GDELT) is employed to analyze the connections between Chinese provinces in mass media. The fields of communication, history, and political science, among others, have widely explored GDELT’s continuous compilation of print, broadcast, and web news media events (Leetaru and Schrodt 2013, Yonamine 2013), but the spatial element of the data has not been be investigated sufficiently. This paper aims to compare the magnitude of spatial decay effect in the GDELT data and a dataset from the Chinese social media website Weibo[1] based on the gravity model. We focus on demonstrating the effectiveness of utilizing both mass media and social media data to reveal geographic patterns, which can be considered as a data pre-processing strategy for pattern recognition and outlier identification in multiple areas such as urban planning, sociology and political geography. Our results also provide valuable input for policy makers to interpret the dynamic nature of inter-region relations in different datasets.

2. Dataset

2.1 Main dataset: GDELT

This research utilizes an open dataset named GDELT. This CAMEO-coded dataset[2] (Schrodt 2012) is updated daily and consists of over a quarter-billion news event records dating back to 1979. It captures what has happened/is happening worldwide, which can be utilized as a valuable resource for modeling societal-scale behavior and beliefs across all countries of the world (Leetaru and Schrodt 2013). The data include multiple columns such as the source, actors, time, and approximated location of recorded events. For consistency we use the data from 01/2014 to 05/2014 in this analysis.

For instance, in a news report entitled "An artist in Shanghai sold his painted box room to the Sifang Art Museum in Nanjing". The associated geographic locations of Actor 1, Actor 2 and the actual action is demonstrated in Table 1. Here we only consider the records when the two actors are explicitly identified and geo-tagged in China.

Table 1. A sample record from GDELT[3]

Event Date	Actor 1_Geo	Actor 2_Geo	Action_Geo
2014-01-28	Shanghai, China	Nanjing, Jiangsu, China	Shanghai, China

2.2 Complementary datasets.

Besides the main dataset GDELT, we also utilize a complementary dataset to compare the spatial decay effects in mass media and geo-tagged social media data. This dataset covers 3 million users in the Chinese social networking site Weibo[4]. The records were sampled between 05/01/2014 to 05/20/2014. Each record captures the geographic coordinates (e.g., volunteered geographic information from built-in GPS module of smart phones), date, time, user ID, and etc. The detailed steps of model construction will be illustrated in Sections 3.

3. Methodology and preliminary results

- ***Data preprocessing***

---

[1] www.weibo.com

[2] CAMEO - Conflict and Mediation Event Observations (CAMEO) is a framework for coding event data

[3] Due to page limit, only fields related to this research are displayed

[4] www.weibo.com

First we calculate the frequencies of "co-occurrence" between each pair of Chinese provinces in GDELT. The frequencies are noted as $F(i, j)$, which stands for the frequency that provinces I and J function as the two actors' locations in the same news record. We also processed the Weibo data to identify the Chinese province associated with each geo-tagged post.

- ***Model construction***

As illustrated in Section 1, the gravity model is utilized in this research to examine the distance decay effect:

$$I_{ij} = K \frac{P_i P_j}{D_{ij}^{\beta}} \tag{1}$$

where P_i and P_j are the "conceptual sizes" (relative importance) of two provinces i and j in a certain topic, D_{ij} represents the distance between them, and I_{ij} denotes the interaction/connection between i and j. Here we construct two gravity models to compare the best fit of distance friction coefficient β to investigate the role of distance decay in the two datasets (GDELT and Weibo). The specific parameters are defined as follows:

GDELT: I_{ij} The frequency of "co-occurrence" of i and j in news records.

P_i The total occurrence of i in news records.

P_j The total occurrence of j in news records.

Weibo: I_{ij} The number of unique users who have physically appeared in both i and j.

P_i The number of unique users who have physically appeared in i.

P_j The number of unique users who have physically appeared in j.

Note that the connection between two provinces in Weibo data can be defined from various perspectives. For instance, the "co-appearance" of two province names in one weibo (post) is another way to define it. However, the primary functionality of Weibo.com is to share moments of one's personal life; hence, it is very rare that users publish posts that explicitly discuss two province names. Here we define the connection based on individual footprints of Weibo users (i.e., the number of unique users who have posted geo-tagged weibo in both provinces), which also aims to examine if the distance decay effects still exist for check-in data.

Based on the above definitions, we calculate the best fit of coefficient β by evaluating the Pearson correlation (R^2) between fitted and observed I_{ij}. The β value that achieves the highest R^2 is considered as the best fit. Since R^2 is scale-free, the constant K does not affect our models. A higher β value indicates a stronger distance decay effect (Gonzalez et al. 2008, Liu et al. 2014). Figure 1 and Figure 2 indicates the correlation between the fit β values and the goodness of fit (R^2) of both datasets.

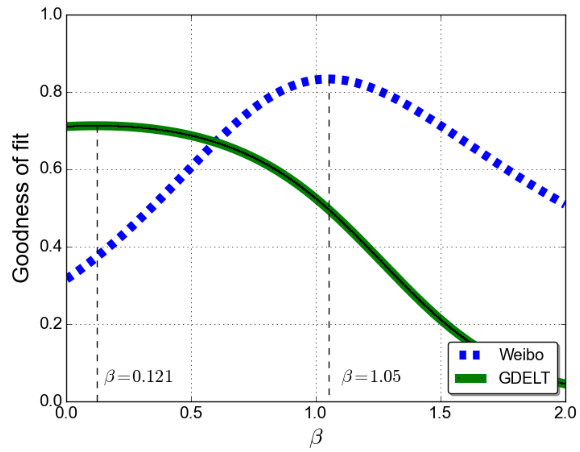
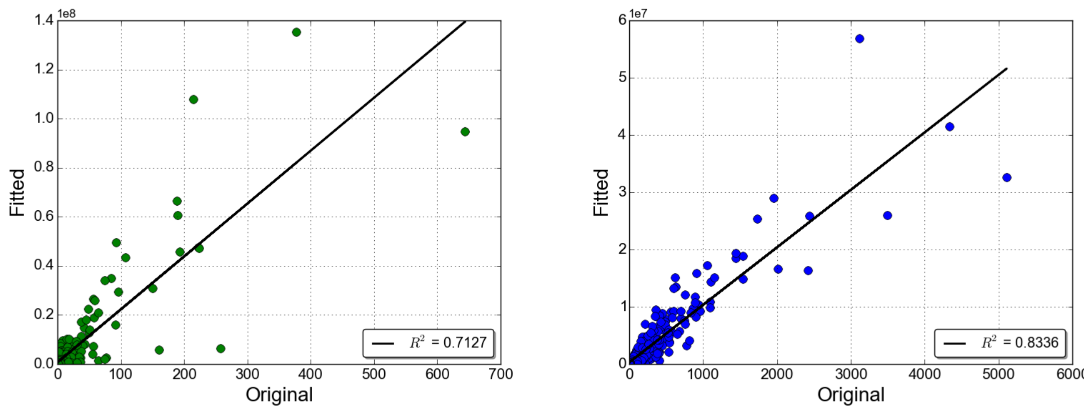


Figure 1. Fitted β values and the goodness of fit (R^2).



(a) (b)

Figure 2. (a) Observed and fitted I_{ij} (GDELT); (b) Observed and fitted I_{ij} (Weibo)

As can be seen, the two datasets demonstrate distinct patterns for distance decay effect. For GDELT dataset the distance decay effect is weaker (β=0.121, R^2=0.7127), whereas the Weibo data shows the strongest distance decay effect (β=1.05, R^2=0.8336). Compared to the β values obtained by several related studies: 0.2 for Chinese province name co-occurrences on web pages (Liu et al. 2014), 1.59 for bank note trajectories (Brockmann and Theis 2008) and 1.75 for individual mobility patterns by mobile phone data (Gonzalez et al. 2008), our study further confirms that mass media data reveals a weak distance decay effect for Chinese provinces.

4. Conclusion

This paper employed the GDELT and Weibo data to examine the connection between Chinese provinces. The contributions of this research include:

- We examined the spatial decay effects in two types of datasets (mass media and social media) for inter-region patterns. The fit β values demonstrate that mass media data indicate a weaker distance decay effect than geotagged social media

(Weibo data) in for Chinese provinces. Unlike the Flickr dataset discussed in Yuan and Liu (2015) which indicates a very weak spatial decay effect (as users are more likely to post photos when they travel faraway), the geo-tagged weibo data still demonstrates a strong spatial decay effect.

- We demonstrated the effectiveness of applying GDELT and big data analytics to investigate informative patterns for interdisciplinary researchers. One potential explanation for the low β value in GDELT data is due to the fact that China is a developing country with a strong central government; hence, the capital Beijing has a significant impact on all other provinces regardless of the distance between them. However, this research does not aim to provide in-depth interpretation of these findings from a political perspective; instead, it proposed a method to discover the patterns that can provide insights in different research fields.

Future research directions include extending this method to other countries and regions to test its robustness. GDELT provides a rich data source to analyze international relations at various spatial scales, such as investigating the connections between different countries. Future studies can also look into the correlation between connection strength and various demographic variables such as population and economic status.

5. References

Brockmann, D. & Theis, F. 2008. Money circulation, trackable items, and the emergence of universal human mobility patterns. *IEEE Pervasive Computing,* 7, 28-35.

Eagle, N., Pentland, A. & Lazer, D. 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America,* 106, 15274-15278.

Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A. L. 2008. Understanding individual human mobility patterns. *Nature,* 453, 779-782.

Hardy, D., Frew, J. & Goodchild, M. F. 2012. Volunteered geographic information production as a spatial process. *International Journal of Geographical Information Science,* 26, 1191-1212.

Leetaru, K. & Schrodt, P. 2013. GDELT: Global Data on Events, Language, and Tone, 1979-2012. *International Studies Association Annual Conference.* San Diego, CA.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P. & Tomkins, A. 2005. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America,* 102, 11623-11628.

Liebert, R. M. & Schwartzberg, N. S. 1977. Effects of Mass-Media. *Annual Review of Psychology,* 28, 141-173.

Liu, Y., Wang, F. H., Kang, C. G., Gao, Y. & Lu, Y. M. 2014. Analyzing Relatedness by Toponym Co-Occurrences on Web Pages. *Transactions in Gis,* 18, 89-107.

Rodrigue, J.-P., Comtois, C. & Slack, B. 2013. *The geography of transport systems,* Abingdon, Oxon, Routledge.

Schrodt, P. 2012. Conflict and Mediation Event Observations Event and Actor Codebook V.1.1b3.

Yonamine, J. E. 2013. Predicting future levels of violence in Afghanistan district using GDELT. UT Dallas.

Yuan, Y. & Liu, Y. 2015. Exploring inter-country connection in mass media: a case study of China. *International Conference on Location-based Social Media.* Athens, Georgia.