# Exploring the spatial decay effect in mass media and location-based 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@txstate.edu

**Abstract** The rapid development of big data analytics provides tremendous possibilities to investigate large-scale patterns in both the spatial and temporal dimensions. In this research, we utilize a unique open dataset, the Global Database on Events, Location, and Tone (GDELT), and a geotagged social media dataset (Weibo) to analyze connections between Chinese provinces. Specifically, this study constructs a gravity model to compare the distance decay effect between the GDELT data (i.e., mass media data) and the Weibo data (i.e., location-based social media (LBSM) data). The results demonstrate that mass media data possess a weaker distance decay effect than LBSM data for Chinese provinces. This study generates valuable input to interpret regional relations in a fast-growing, developing country—China. It also provides methodological references to explore urban relations in other countries and regions in the big data era.

**Key words:** Mass media; GDELT; Location-based social media (LBSM); Gravity model; Distance decay.

## 1. Introduction

The rapid development of techniques and theories in the big data era has introduced new challenges and opportunities to analyze a large amount of social media data available online (Gao et al. 2012; Eagle et al. 2009; Liben-Nowell et al. 2005; Wu et al. 2014). The widespread use of smart phones, which are equipped with sensors allowing users to instantly locate themselves, inserts another crucial aspect into this development: location. Researchers have defined location-based social media (LBSM) as “Social Network Sites (SNS) that include location information” (Roick and Heuser 2013; Elwood et al. 2012). These data are user-generated, geolocated, and contain varying types of contextual information (e.g., text, videos, and images.) Therefore they can be utilized as potential resources to characterize spatial interactions and social perceptions of place (Malleson and Birkin 2014; Hasan et al. 2013; Gao et al. 2012).

The past few decades also have witnessed many tremendous changes in the traditional media industry. In particular, the rapid development of new media, such as video games and online news columns, also has allowed a new paradigm to emerge in human behavior modeling and pattern recognition. For researchers in this field, these newly-available data sources also offer ever-increasing opportunities to conduct data mining and knowledge discovery tasks (Li and Liu 2003; Masand et al. 1992; Mazzitello et al. 2007). However, compared to social media, traditional mass media is characterized by the significance and aggregated nature of associated events (Liebert and Schwartzberg 1977; Klapper 1968). As such, mass media data often are suitable for investigating the aggregated pattern of an urban system. However, very few empirical studies have compared corresponding spatial patterns extracted from mass media data and LBSM data.

The distance decay effect has been a hot topic in many research fields, such as migration 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 affects the magnitude of interaction between geographic entities. Among all alternative models, researchers commonly use the gravity model (Sen and Smith 1995) due to its effectiveness in predicting the degree of interaction, the simplicity of its equation, and its ability to deal with flows in both directions (Hardy et al. 2012). In this research, we apply an open source dataset, the Global Database on Events, Location, and Tone (GDELT) to analyze connections between Chinese provinces in terms of 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 this dataset has not been investigated sufficiently. This paper compares the magnitude of the distance decay effect in the GDELT data with a dataset from a Chinese social media website (Weibo.com) based on the gravity model. We focus on demonstrating the effectiveness of utilizing both mass media and LBSM data to reveal geographic patterns. This can be considered 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 policymakers to interpret the dynamic nature of inter-region relations in different datasets.

## 2. Datasets

This research primarily utilizes two datasets: 1) the main dataset GDELT, consisting of over a quarter-billion event records from 1979 to the present capturing what has happened/is happening worldwide in multiple columns, such as the source, actors, time, and approximated location of recorded events; 2) a complementary dataset from the Chinese social networking site, Weibo[1], to compare the distance decay effects in mass media and LBSM. The remainder of Section 2 illustrates the two datasets in detail.

### (1) The main dataset: GDELT

[1] www.weibo.com

This research utilizes a CAMEO-coded[2] open dataset, GDELT (Schrodt 2012). For consistency, we use the data from 01/01/2014 to 05/20/2014 in this analysis. Previous studies utilized this dataset as a valuable resource for modeling societal-scale behavior and beliefs across all countries of the world (Leetaru and Schrodt 2013; Shook et al. 2012; Yuan and Liu 2015). For example, a study conducted by Jiang and Mai (Jiang and Mai 2014) analyzes the strength of links between countries based on the GDELT dataset. They also explored bilateral and multilateral events in certain countries from the same dataset. Other research by Yonamine (2013) predicts the level of conflict in Afghanistan by incorporating multiple sociopolitical factors, such as drug prices, unemployment levels, and ethnic diversity. Yet other researchers have looked into the connections and differences between GDELT and other event-based datasets, such as the Integrated Crisis Early Warning System (ICEWS), which is an early-warning system designed to help policymakers predict a variety of international events and crises. They conducted a side-by-side comparison regarding the data quality, quantity, design scheme, and many other features of these two datasets, and concluded that although the efficiency of each dataset mainly depends on the research questions to be answered, GDELT has many more events per country per unit of time than ICEWS.

In GDELT, for instance, for 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 (i.e., "sold") is demonstrated in Table 1. Here we consider records only when two actors are explicitly identified and geotagged in China.

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

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 |

Note that GDELT measures the detailed level of the spatial information by a field named Geo_Type. This field specifies the geographic resolution of each location and holds one of the following values: 1=COUNTRY (country level), 2=USSTATE (a United States (U.S.) state), 3=USCITY (a U.S. city or landmark), 4=WORLDCITY (a city or landmark outside the U.S.), 5=WORLDSTATE (an Administrative Division 1 outside the U.S. – roughly equivalent to a U.S. state) (Leetaru and Schrodt 2013; Yuan and Liu 2015). Because this research is conducted at the province level, we should consider only records with Geo_Type = 4 or Geo_Type = 5 (Figure 1, map created in Mercator projection).

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

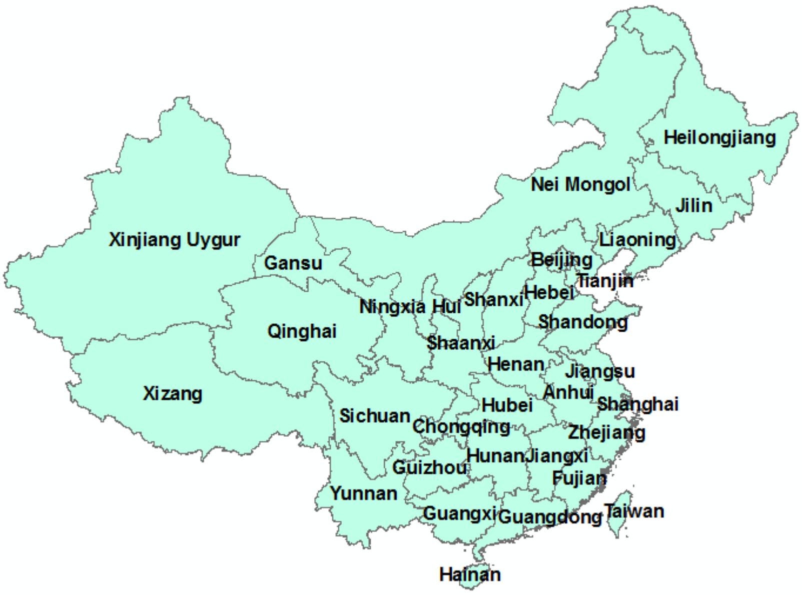


Figure 1. Chinese provinces (the Paracel islands are omitted for simplicity)

### (2) Complementary datasets.

Besides the main GDELT dataset, we also utilize a complementary dataset to compare the distance decay effects in mass media and LBSM. This dataset covers three million users in the Chinese social networking site, Weibo, a micro-blogging website functionally similar to Twitter. The records were obtained from the official Weibo application program interface (API) between 05/01/2014 to 05/20/2014. Each record captures such attributes as the geographic coordinates (e.g., volunteered geographic information from the built-in positioning module of smart phones), date, time, and user identification (ID)[4].

[4] Here user IDs are long integers generated by Weibo.com and are not directly connected to any personally identifiable information (PII), unless the users volunteer to make such information publicly accessible.

## 3. Methodology and preliminary results

This section presents the model-construction procedure and our preliminary results. As discussed in Section 1, the major objective of this research is to compare the magnitude of the distance decay effects in mass media and LBSM by investigating how distance affects inter-regional connection. Although researchers have applied various techniques to analyze spatial interaction in many research fields, such as transportation (e.g., traffic flows between locations) and migration (e.g., relocation flows between countries) (Rodrigue et al. 2013; Lewer and Van den Berg 2008), the gravity model furnishes one of the most commonly used descriptions of this phenomena because of its effectiveness in predicting the degree of interaction and its algebraic simplicity (Rodrigue et al. 2013; Hardy et al. 2012; Sen and Smith 1995). This study also adopts the gravity model. The analysis involves the following three steps.

- ***Data preprocessing***

First, we calculate the frequency of "co-occurrence" between each pair of Chinese provinces in GDELT. The frequency is noted as $I_{ij}$, which denotes the frequency of provinces $i$ and $j$ appearing as the two actors' locations in the same news record. We also processed the Weibo data to identify the Chinese province associated with each geotagged post.

- ***Model construction***

As illustrated in Section 1, this research utilizes the gravity model to examine the distance decay effect (Equation 1):

$$I_{ij} = K \frac{P_i^{\beta_1} P_j^{\beta_2}}{D_{ij}^{\beta_3}} \tag{1}$$

where $P_i$ and $P_j$ are the "conceptual sizes" (relative importance) of provinces $i$ and $j$, $D_{ij}$ represents the great circle distance separating the geographic centroids of $i$ and $j$, and $I_{ij}$ denotes the interaction/connection between $i$ and $j$. $\beta_1$ and $\beta_2$ indicate how the "conceptual sizes" of two countries contribute to the interaction term $I_{ij}$ (Austin 1963). $\beta_3$ (distance friction coefficient) investigates the role of distance. Here we construct two gravity models to investigate the role of the friction of distance in the two datasets (GDELT and Weibo). The specific parameters are:

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

$P_i$ The total occurrence of province *i* in news records.

$P_j$ The total occurrence of province *j* in news records.

**Weibo**: $I_{ij}$ The number of unique users who have checked-in at their locations in both provinces *i* and *j*.

$P_i$ The number of unique users who have checked-in at their locations in province *i*.

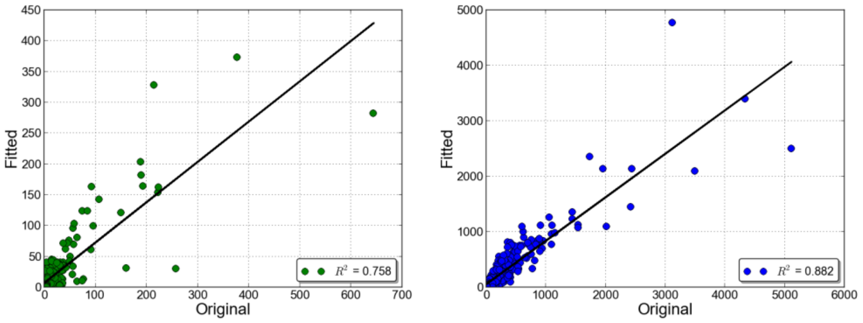
$P_j$ The number of unique users who have checked-in at their locations in province *j*.

Note that the connection between two provinces in the Weibo data can be defined from various perspectives; for example, the "co-appearance" of two province names in the same Weibo post. However, the primary functionality of Weibo.com is to share moments of one's personal life, and hence users rarely publish posts that explicitly discuss two province names. Also, this research focuses on comparing spatial patterns from mass media and geotagged social media; thus, we define connections based on the individual footprint of Weibo users. In other words, the Weibo dataset measures physical movements while the GDELT dataset measures information interactions. However, one of the objectives of this research is to investigate the representativeness of Weibo data in modeling physical mobility, because other studies demonstrated the presence of strong spatial biases when utilizing LBSM data to model spatial behavior. Flickr is an example, where photo-uploading activities do not exhibit a significant distance decay effect because people tend to upload photos when they travel to faraway destinations.

Based on the preceding definitions, we calculated the best fit for coefficients $\beta_1$, $\beta_2$, $\beta_3$ based on Poisson regression. Table 2 indicates the fitted values and the goodness of fit ($R^2$) for both datasets. Because $R^2$ is scale-free, the constant $K$ does not affect our models.

Table 2. Fitted $\beta$ values and pseudo-$R^2$ of Poisson regression models.

| | $\beta_1$ | $\beta_2$ | $\beta_3$ | **Pseudo-$R^2$** |
|---|---|---|---|---|
| **GDELT** | 0.826 | 0.826 | 0.0516 | 0.758 |
| **Weibo** | 0.799 | 0.799 | 1.054 | 0.882 |



(a) (b)

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

Results for the two datasets demonstrate distinct patterns for the distance decay effect. For the GDELT dataset, the distance decay effect (based on great circle distances between provinces) is weak ($\beta_3$=0.0516, pseudo-$R^2$=0.758), whereas the Weibo dataset shows a much stronger distance decay effect ($\beta_3$=1.054, pseudo-$R^2$=0.882). Compared to the $\beta_3$ values obtained for 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 reveal a weaker distance decay effect than LBSM data for Chinese provinces. Additionally, $\beta_1$ and $\beta_2$ values also indicate interesting patterns regarding the role of "conceptual sizes" ($P_i$, $P_j$) in determining the magnitude of this interaction. $P_i$, $P_j$ play a more important role in GDELT data. Note that, unlike the gravity of trade where the interaction term is directional, the interaction term $I_{ij}$ here is not bilateral in either dataset. For example, "the frequency of 'co-occurrence' of provinces *i* and *j* in news records" is equivalent to "the frequency of 'co-occurrence' of provinces *j* and *i* in news records"; thus, the roles of provinces *i* and *j* are exchangeable, resulting in identical $\beta_1$ and $\beta_2$ estimates.

Finally, the different aspects of uncertainty involved in this study are noteworthy, including but not limited to:

- *Natural variability in human activities*: Although human mobility seems to be highly predictable (Yuan et al. 2012; Song et al. 2010; Gonzalez et al. 2008), randomness is an inevitable part of human motion.
- *Potential impact of spatial autocorrelation*: Many researchers have made a distinction between spatial association (autocorrelation) and spatial interaction in the geography field. Interaction primarily refers to movement of tangible entities; therefore, it is less related to correlation. However, several previous studies also argued that spatial interaction models are a special case of a general model of spatial autocorrelation (Fischer et al. 2010; Getis 1991). Hence, the impact of spatial autocorrelation on the gravity models constructed in this research needs to be examined in future research (Chun and Griffith 2011).

- *Inaccuracy/imprecision due to the limitation of available data*: Positional inaccuracy, sampling resolution, and imprecision contribute to the uncertainty of a data source. For instance, the precision of geotagged Weibo posts is strongly related to the strength of an available Global Positioning System (GPS) signal. Also, the sampling resolution is unevenly distributed, as many users may not post their locations to Weibo on a regular basis.
- *Imperfection of models and algorithms*: As Box and Draper (Box and Draper 1987) state: "Essentially, all models are wrong, but some are useful." The results of this study are also highly impacted by the chosen models and algorithms. For the GDELT dataset, actors are georeferenced automatically based on various text mining and machine learning algorithms, naturally introducing potential inaccuracy into the location data. In this research, the gravity model is adopted to interpret inter-region connections in China; however, the application of different models inevitably has an impact on the uncertainty of results. For example, an interesting future direction could be to simulate and estimate the conceptual sizes ($P_i$ and $P_j$) from an inverse gravity model, and compare the results to the current model where $P_i$ and $P_j$ are pre-defined.

## 4. Conclusion

The study summarized in this paper employed GDELT and Weibo data to examine the connection between Chinese provinces. We examined the distance decay effects in two types of datasets (mass media and LBSM) for inter-regional patterns. The fitted $\beta_3$ values demonstrate that mass media data (GDELT) indicate a weaker distance decay effect than geotagged social media data (Weibo) for Chinese provinces. Unlike the Flickr dataset discussed in Yuan and Liu (2015), which indicates a very weak distance decay effect (as users are more likely to post photos when they travel faraway), the geotagged Weibo data still demonstrate a strong distance decay effect. This finding suggests an interesting direction for future research, namely to examine the distance decay effect in various social network sites based on their functionalities. We also demonstrated the effectiveness of applying GDELT and big data analytics to investigate informative patterns in interdisciplinary studies. One potential explanation for the low $\beta_3$ value in the GDELT data is China's status as a developing country with a strong central government. Its capital, Beijing, has a significant impact on all other provinces, regardless of the distance separating them. However, this research does not aim to provide an in-depth interpretation of these findings from a political perspective. Rather, it proposes a method to discover 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 also can look into the correlation between connection strength and various demographic variables, such as population and economic status.

## References

Austin LC (1963) Shaping the World-Economy - Tinbergen,J. J Int Aff 17 (2):221-221

Box GEP, Draper NR (1987) Empirical model-building and response surfaces. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley, New York

Brockmann D, Theis F (2008) Money circulation, trackable items, and the emergence of universal human mobility patterns. Ieee Pervas Comput 7 (4):28-35

Chun Y, Griffith DA (2011) Modeling Network Autocorrelation in Space–Time Migration Flow Data: An Eigenvector Spatial

Filtering Approach. Ann Assoc Am Geogr 101 (3):523-536. doi:10.1080/00045608.2011.561070

Eagle N, Pentland A, Lazer D (2009) Inferring friendship network structure by using mobile phone data. P Natl Acad Sci USA 106 (36):15274-15278. doi:10.1073/pnas.0900282106

Elwood S, Goodchild MF, Sui DZ (2012) Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. Ann Assoc Am Geogr 102 (3):571-590. doi:Doi 10.1080/00045608.2011.595657

Fischer M, Reismann M, Scherngell T (2010) Spatial Interaction and Spatial Autocorrelation. In: Anselin L, Rey SJ (eds) Perspectives on Spatial Data Analysis. Advances in Spatial Science. Springer Berlin Heidelberg, pp 61-79. doi:10.1007/978-3-642-01976-0_5

Gao H, Tang J, Liu H (2012) Exploring Social-Historical Ties on Location-Based Social Networks. Paper presented at the 6th International AAAI Conference on Weblogs and Social Media, Dublin, Ireland,

Getis A (1991) Spatial Interaction and Spatial Autocorrelation: A Cross-Product Approach. Environ Plann A 23:1269-1277

Gonzalez MC, Hidalgo CA, Barabasi AL (2008) Understanding individual human mobility patterns. Nature 453 (7196):779-782. doi:Doi 10.1038/Nature06958

Hardy D, Frew J, Goodchild MF (2012) Volunteered geographic information production as a spatial process. Int J Geogr Inf Sci 26 (7):1191-1212. doi:Doi 10.1080/13658816.2011.629618

Hasan S, Zhan X, Ukkusuri SV Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In: UrbComp 13, Chicago, 2013. ACM,

Jiang L, Mai F (2014) Discovering Bilateral and Multilateral Causal Events in GDELT. Paper presented at the International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction,

Klapper JT (1968) Effects of Mass-Media-Depicted Violence - a Review of Research Findings. Am J Orthopsychiat 38 (2):310-&

Leetaru K, Schrodt P (2013) GDELT: Global Data on Events, Language, and Tone, 1979-2012. Paper presented at the International Studies Association Annual Conference, San Diego, CA, April 2013

Lewer JJ, Van den Berg H (2008) A gravity model of immigration. Econ Lett 99 (1):164-167. doi:DOI 10.1016/j.econlet.2007.06.019

Li X, Liu B (2003) Learning to classify texts using positive and unlabeled data. Paper presented at the Proceedings of the 18th international joint conference on Artificial intelligence, Acapulco, Mexico,

Liben-Nowell D, Novak J, Kumar R, Raghavan P, Tomkins A (2005) Geographic routing in social networks. P Natl Acad Sci USA 102 (33):11623-11628. doi:DOI 10.1073/pnas.0503018102

Liebert RM, Schwartzberg NS (1977) Effects of Mass-Media. Annu Rev Psychol 28:141-173

Liu Y, Wang FH, Kang CG, Gao Y, Lu YM (2014) Analyzing Relatedness by Toponym Co-Occurrences on Web Pages. Transactions in GIS 18 (1):89-107. doi:Doi 10.1111/Tgis.12023

Malleson N, Birkin M (2014) New Insights into Individual Activity Spaces using Crowd-Sourced Big Data. Paper presented at the ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference, Stanford, CA, May 27-31, 2014

Masand B, Linoff G, Waltz D (1992) Classifying news stories using memory based reasoning. Paper presented at the Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, Copenhagen, Denmark,

Mazzitello KI, Candia J, Dossetti V (2007) Effects of mass media and cultural drift in a model for social influence. Int J Mod Phys C 18 (9):1475-1482

Rodrigue J-P, Comtois C, Slack B (2013) The geography of transport systems. Third edition . edn. Routledge, Abingdon, Oxon

Roick O, Heuser S (2013) Location Based Social Networks - Definition, Current State of the Art and Research Agenda. Transactions in GIS:763-784

Conflict and Mediation Event Observations Event and Actor Codebook V.1.1b3 (2012) http://eventdata.psu.edu/cameo.dir/CAMEO.Manual.1.1b3.pdf.

Sen AK, Smith TE (1995) Gravity models of spatial interaction behavior. Advances in spatial and network economics. Springer-Verlag, Berlin ; New York

Shook E, Leetaru K, Cao G, Padmanabhan A, Wang. S (2012) Happy or not: Generating topic-based emotional heatmaps for Culturomics using CyberGIS. Paper presented at the IEEE 8th International Conference on EScience,

Song CM, Qu ZH, Blumm N, Barabasi AL (2010) Limits of predictability in human mobility. Science 327 (5968):1018-1021. doi:DOI 10.1126/science.1177170

Wu L, Zhi Y, Sui ZW, Liu Y (2014) Intra-Urban Human Mobility and Activity Transition: Evidence from Social Media Check-In Data. Plos One 9 (5):e97010. doi:ARTN e97010DOI 10.1371/journal.pone.0097010

Yonamine JE (2013) Predicting future levels of violence in Afghanistan district using GDELT. UT Dallas,

Yuan Y, Liu Y Exploring inter-country connections in mass media: a case study of China. In: International Conference on Location-based Social Media, Athens, GA, 2015.

Yuan Y, Raubal M, Liu Y (2012) Correlating mobile phone usage and travel behavior - a case study of Harbin, China. Computers, Environment and Urban Systems 36 (2):118-130