

# The impact of scale on spatial connections: an exploratory analysis

Yihong Yuan \*

*Department of Geography, Texas State University, San Marcos, TX 78666, yuan@txstate.edu*

\* Corresponding author

**Abstract:** Quantifying the intensity of spatial connections has been a crucial topic in many research fields, such as urban transportation, migration, and trade. Researchers have proposed various models, such as the gravity model and the radiation model, to quantify the magnitude of spatial connections. Traditionally, modeling the connections (relatedness) between spatial entities is limited to the physical space, but with the rapid growth of information technologies, the scope of spatial connections extends to the virtual space. However, one topic that has not been fully studied is how spatial scale may impact spatial connections in the virtual space and how this influence can be reflected in spatial decay models. In this study, we used two types of datasets (mass media and social media data) to explore the impact of scale on fitting the distance decay coefficient. The results confirmed that spatial scale can impact the magnitude of spatial decay effects in datasets with different characteristics.

**Keywords:** Big geo-data, Mass media, Location-based social media (LBSM), Distance decay.

## 1. Introduction

Modeling the intensity of spatial connections has been a crucial topic in many research fields, such as urban transportation, migration, and trade (Rodrigue, Comtois, and Slack 2013; Lewer and Van den Berg 2008; Chen, Gong, and Xie 2017). Spatial connections measure how two geographic entities relate to each other and cover a broader range of connections than “spatial interactions” (Yuan, Liu, and Wei 2017). From a geographic perspective, a crucial element that impacts the magnitude of spatial connections is the distance decay effect, which is best summarized by Tobler's first law of geography (TFL) (1970), “near things are more related than distant things.” Researchers have proposed various models, such as the gravity model and the radiation model (Rodrigue, Comtois, and Slack 2013; Hardy, Frew, and Goodchild 2012; Simini et al. 2012), to quantify the magnitude of spatial connections, among which the gravity model is commonly used due to its calculation simplicity and the extensibility of the parameters (Sgrignoli et al. 2015; Allen and Arkolakis 2014; Anderson 2011).

Traditionally, spatial interactions or connections occur due to the physical movement of objects or people, such as transporting goods from one city to another. In the information age, the cause of spatial connection goes beyond physical movements (Liu et al. 2014b; Hu 2018). With the rapid growth of information and communication technologies (ICTs), the scope of spatial connections expands from the physical space to the virtual space. For example, Liu et al. (2014b) investigated the relatedness between Chinese provinces based on their co-occurrence on webpages. Yuan, Liu, and Wei (2017) classified the ties and connections between China and other countries based on international news reports. These studies provided

valuable insight for understanding various types of spatial connections in the big data era.

However, one topic that has not been fully studied is how spatial scale may impact spatial connections in the virtual space and how this influence can be reflected on spatial decay models. In geography, the “scale effect” refers to the fact that analytical results may vary with different analytical units (Batty 2008; Chen et al. 2019; Brockmann, Hufnagel, and Geisel 2006). For example, extracting crime hotspots at the census block group level may yield different results from extracting crime hotspots from the census tract level (Chainey, Tompson, and Uhlig 2008). This scale effect in spatial analysis is often referred to as the Modifiable Areal Unit Problem (MAUP) (Openshaw and Vanderknaap 1983; Jelinski and Wu 1996). This scale effect can potentially impact how two geographic entities are related to each other, both in the physical space and in the visual space. For example, previous studies have used online news reports to explore the distance decay effect between Chinese provinces (Yuan 2017); however, the distance friction coefficient may be different when fitting a gravity model to calculate the relatedness of places at different spatial scales (e.g., at a country level or city level).

In this study, we use two types of datasets to explore the impact of scale on fitting the distance decay coefficient. The first dataset is an open-source dataset, “The Global Data on Events, Location and Tone” (GDELT), which is a CAMEO-coded dataset updated daily to capture over a quarter-billion news event records worldwide dating back to 1979 (Schrodt 2012; Yonamine 2013; Ward et al. 2013; Leetaru and Schrodt 2013). The spatial connections based on GDELT data are defined as the frequency of “co-occurrence” of two places. For example, if countries *A* and *B* frequently appear in the same news record, then the two

countries have a strong spatial connection. The second dataset was collected from social media platforms Flickr and Weibo, which include geotagged posts or photos from platform users. The spatial connection between places *A* and *B* based on the social media data is defined as the likelihood of a user visiting two places. For example, if more users visited both countries *A* and *B*, then the two countries have a strong spatial connection (Yuan, Liu, and Wei 2017). We fit a gravity model for both datasets and calculated the distance friction coefficient  $\beta$  for the two datasets both at two spatial scales: 1) country-level, which measures the spatial connections between countries, and 2) provincial-level, which measures the connections between Chinese provinces. The hypothesis is that the spatial scale has an impact on the magnitude of the distance decay coefficient.

## 2. Related Work

With the development of Information and Communication Technologies (ICTs), various new datasets, new strategies, and new analytical methods have been developed to examine traditional geography laws and principles (Yuan, Raubal, and Liu 2012; Gaster 1996). ICTs provide a wide range of spatio-temporal data sources and can capture real-time geographic patterns more efficiently and effectively (Song et al. 2010; Yuan 2009; Miller 2009; Zook, Kraak, and Ahas 2015; Yuan et al. 2019).

As one of the most fundamental theories in geography, Tobler's first law of geography (1970) specifies that near things are more related than things far apart (Sui 2004). Researchers have applied various models, such as the gravity model and the radiation model, to quantify how physical distance affects the connections (relatedness) between two geographic entities. With the rapid growth of ICT datasets, the exploration of spatial connections goes beyond the physical environment and extends into the virtual space (Miller 2004; Hecht and Moxley 2009). For example, Liu et al. (2014a) used gravity models to explore relatedness between Chinese provinces based on the co-occurrence of province names on webpages. Their assumption was that more co-occurrences of two province names implies a stronger connection. Hardy, Frew, and Goodchild (2012) applied an invariant exponential gravity model to investigate how distance impacts the generation volunteered geographic information (VGI) data on Wikipedia. Their results show that distance plays a crucial factor on the diffusion of information in the virtual space. Yuan, Liu, and Wei (2017) examined the magnitude of spatial connection from three datasets: mass media news reports, location-based social media (LBSM) check-ins, and airline transportation data. The results show that, at the country level, mass media data indicate a stronger distance decay effect than social media, but a weaker effect than air transportation data.

Although many previous studies explored the impact of distance on spatial connections in the age of instant access,

there is one crucial factor that has yet to be taken into consideration – the scale. Researchers often need to aggregate the raw data or divide the study area into spatial units, and the size of the spatial unit inevitably impacts the level of detail captured in an analysis (Zhang et al. 2018; Chen, Gong, and Xie 2017). Very few studies have attempted to investigate how spatial scale impacts the connections between geographic entities (Fernández Vázquez and Rubiera Morollón 2012; Lloyd 2014; Strauss et al. 2009).

## 3. Data and Methodology

### 3.1 Data

As mentioned in Section 1, we used two types of data in this study:

The first dataset, GDELT, consists of over a quarter-billion news event records dating back to 1979 and is updated daily (Yuan, Liu, and Wei 2017). The data includes various attributes such as date, time, actors, time, and approximated location of the event. For example, in a news report regarding the 47<sup>th</sup> G7 summit held in Cornwall, England, Boris Johnson told the BBC after meeting U.S. President Joe Biden, “The alliance between the U.S. and the U.K. should be known as the “indestructible relationship”<sup>1</sup>. In this news report, Actor 1 will be “the United States government” and Actor 2 will be “the United Kingdom government”. Table 1 shows the associated geographic locations of Actor 1, Actor 2, and the actual action.

Event Date	Actor 1_Geo	Actor 2_Geo	Action_Geo
2021-06-11	London, England	Washington, District of Columbia, United States	Cornwall, England

Table 1. Example record from GDELT data

The second dataset consists of information from two social media platforms, Flickr and Weibo (a microblogging site in China). The Flickr data were collected globally from 2008-2012 and include more than 30 million geo-tagged Flickr images. The Weibo data were obtained from the official Weibo application program interface (API) between 05/01/2014 and 05/20/2014 and cover three million users. Each record captures the geographic coordinates (e.g., volunteered geographic information from the built-in positioning module of smart phones), date, time, and user identification (ID).

### 3.2 Methodology

To explore the connection between geographic entities, here we construct a gravity model as follows (Equation 1):

<sup>1</sup> The alliance between the US and the UK should be known as the “indestructible relationship

$$I_{ij} = K \frac{P_i P_j}{D_{ij}^\beta} \quad (1)$$

where  $P_i$  and  $P_j$  are the “conceptual sizes” (relative importance) of geographic entities  $i$  and  $j$ ,  $D_{ij}$  represents the distance separating the geographic centroids of  $i$  and  $j$ , and  $I_{ij}$  denotes the interaction/connection between  $i$  and  $j$  (Austin 1963).  $\beta$  (distance friction coefficient) represents the role of distance. Here we construct four gravity models to investigate the role of the friction of distance in determining the spatial connections in the aforementioned dataset. The specific parameters of the four models are defined as follows:

#### Model 1 - GDELT (country-level):

$I_{ij}$  - The frequency of “co-occurrence” of countries  $i$  and  $j$  in news records.

$P_i$  - The total occurrence of country  $i$  in news records.

$P_j$  - The total occurrence of country  $j$  in news records.

#### Model 2 - GDELT (provincial-level):

$I_{ij}$  - 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.

#### Model 3 - Flickr (country-level):

$I_{ij}$  - The number of unique users who have uploaded images in both  $i$  and  $j$ .

$P_i$  - The number of unique users who have uploaded images in country  $i$ .

$P_j$  - The number of unique users who have uploaded images in country  $j$ .

#### Model 4 - Weibo (provincial-level):

$I_{ij}$  - 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  $j$ .

Based on the preceding definitions, we calculated the best fit for coefficient  $\beta$  for the four models. The parameters were fitted through a Poisson regression. Because the correlation coefficient of regression models ( $R^2$ ) is scale-free, the constant  $K$  does not affect our models.

## 4. Results

Table 2 shows the fitted  $\beta$  values for both datasets. The larger  $\beta$  is, the stronger the distance decay effect is for that particular dataset. As can be seen, the mass media data from GDELT and the crowd-sourced LBSM data demonstrate an opposite pattern. For GDELT data, news reports show a stronger distance decay effect at the country scale than at the provincial scale. This is potentially because national news is more influenced by distance than domestic news is in China. For social media data, country-level analysis shows a much weaker spatial decay effect

than provincial level analysis. It is possible that when traveling internationally, users are less bounded by distance and focus more on the popularity of a destination.

	coefficient $\beta$
Model 1 - GDELT (country-level)	$\beta=0.74$
Model 2 - GDELT (provincial-level)	$\beta=0.12$
Model 3 - Flickr (country-level)	$\beta=0.12$
Model 4 - Weibo (provincial-level)	$\beta=1.05$

Table 2. Fitting the distance friction coefficient at the country level and provincial level.

In previous studies, researchers extracted the  $\beta$  value for different types of movements and interactions: 0.2 for Chinese province name co-occurrences on the web (Liu et al. 2014b), 1.59 for bank note circulation (Brockmann and Theis 2008), and 1.75 for human mobility extracted from mobile phone records (Gonzalez, Hidalgo, and Barabasi 2008). There is a general consensus that geographic entities experience a weaker distance delay in the virtual space (e.g., on social media) than in the physical environment. This study provides a unique perspective to investigate this problem and demonstrates that a unified answer may not be sufficient when exploring spatial delay in the information age, and it is necessary to consider the impact of other factors, such as scale, on spatial connections.

## 5. Conclusion

The development of information technologies has introduced exciting changes and new challenges in various research fields. In this study, we examined the impact of spatial scale on distance decay effects. The results confirmed that spatial scale can impact the magnitude of spatial decay effects in datasets with different characteristics. The consensus in previous studies is that the spatial decay effect is weaker in the virtual space than in the physical space; however, this study shows that the conclusion may not be as simple as it seems and may worth a deeper investigation. There are many factors that play into how distance impacts a certain type of spatial connection, and spatial scale is a factor that should not be overlooked. In the next step of this analysis, we will extend the analysis to other countries and include more datasets to test the robustness of the results. Future studies also can look into how various demographic variables, such as income, may impact the magnitude of spatial connections.

## 6. References

- Allen, T., and C. Arkolakis. 2014. Trade and the Topography of the Spatial Economy. *Quarterly Journal of Economics* 129 (3):1085-1140.
- Anderson, J. E. 2011. The Gravity Model. *Annual Review of Economics*, Vol 3 3:133-160.
- Austin, L. C. 1963. Shaping the World-Economy - Tinbergen, J. *Journal of International Affairs* 17 (2):221-221.
- Batty, M. 2008. The Size, Scale, and Shape of Cities. *Science* 319 (5864):769.
- Brockmann, D., L. Hufnagel, and T. Geisel. 2006. The scaling laws of human travel. *Nature* 439 (7075):462-465.
- Brockmann, D., and F. Theis. 2008. Money circulation, trackable items, and the emergence of universal human mobility patterns. *Ieee Pervasive Computing* 7 (4):28-35.
- Chainey, S., L. Tompson, and S. Uhlig. 2008. The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal* 21 (1-2):4-28.
- Chen, L., Y. Gao, D. Zhu, Y. Yuan, and Y. Liu. 2019. Quantifying the scale effect in geospatial big data using semi-variograms. *Plos One* 14 (11):e0225139.
- Chen, Z. L., X. Gong, and Z. Xie. 2017. An analysis of movement patterns between zones using taxi GPS data. *Transactions in GIS* 21 (6):1341-1363.
- Fernández Vázquez, E., and F. Rubiera Morollón. 2012. *Defining the spatial scale in modern regional analysis : new challenges from data at local level*. Heidelberg: Springer.
- Gaster, R. 1996. *Bit by bit : building a transatlantic partnership for the information age*. Armonk, N.Y.: M.E. Sharpe.
- Gonzalez, M. C., C. A. Hidalgo, and A. L. Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453 (7196):779-782.
- Hardy, D., J. Frew, and M. F. Goodchild. 2012. Volunteered geographic information production as a spatial process. *International Journal of Geographical Information Science* 26 (7):1191-1212.
- Hecht, B., and E. Moxley. 2009. Terabytes of Tobler: Evaluating the First Law in a Massive, Domain-Neutral Representation of World Knowledge. In *Conference of Spatial Information Theory*, 88-105: Springer.
- Hu, Y. 2018. Geo-text data and data-driven geospatial semantics. *Geography Compass*:e12404.
- Jelinski, D. E., and J. Wu. 1996. The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecology* 11 (3):129-140.
- Leetaru, K., and P. Schrodt. 2013. GDELT: Global Data on Events, Language, and Tone, 1979-2012. In *International Studies Association Annual Conference*. San Diego, CA.
- Lewer, J. J., and H. Van den Berg. 2008. A gravity model of immigration. *Economics Letters* 99 (1):164-167.
- Liu, Y., F. Wang, C. Kang, Y. Gao, and Y. Lu. 2014a. Analyzing Relatedness by Toponym Co-Occurrences on Web Pages. *Transactions in GIS* 18 (1):89-107.
- Liu, Y., F. H. Wang, C. G. Kang, Y. Gao, and Y. M. Lu. 2014b. Analyzing Relatedness by Toponym Co-Occurrences on Web Pages. *Transactions in GIS* 18 (1):89-107.
- Lloyd, C. D. 2014. *Exploring spatial scale in geography*. Chichester, West Sussex, ; Hoboken, NJ: John Wiley & Sons.
- Miller, H. J. 2004. Tobler's first law and spatial analysis. *Annals of the Association of American Geographers* 94 (2):284-289.
- . 2009. Geographic data mining and knowledge discovery: An overview. In *Geographic Data Mining and Knowledge Discovery (Second Edition)*, eds. H. J. Miller and J. Han, 3-32. London: CRC Press.
- Openshaw, S., and G. A. Vanderknaap. 1983. Small Area Population Forecasting - Some Experience with British Models. *Tijdschrift Voor Economische En Sociale Geografie* 74 (4):291-304.
- Rodrigue, J.-P., C. Comtois, and B. Slack. 2013. *The geography of transport systems*. Third edition . ed. Abingdon, Oxon: Routledge.
- Schrodt, P. 2012. Conflict and Mediation Event Observations Event and Actor Codebook V.1.1b3.
- Sgrignoli, P., R. Metulini, S. Schiavo, and M. Riccaboni. 2015. The relation between global migration and trade networks. *Physica a-Statistical Mechanics and Its Applications* 417:245-260.
- Simini, F., M. C. Gonzalez, A. Maritan, and A. L. Barabasi. 2012. A universal model for mobility and migration patterns. *Nature* 484 (7392):96-100.
- Song, C. M., Z. H. Qu, N. Blumm, and A. L. Barabasi. 2010. Limits of predictability in human mobility. *Science* 327 (5968):1018-1021.
- Strauss, T., J. Lentz, Iowa. Department of Transportation., and Midwest Transportation Consortium. 2009. *Spatial scale of clustering of motor vehicle crash types and appropriate countermeasures*. Ames, IA: Midwest Transportation Consortium, Iowa State University.
- Sui, D. Z. 2004. Tobler's first law of geography: A big idea for a small world? *Annals of the Association of American Geographers* 94 (2):269-277.
- Tobler, W. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46:234-240.
- Ward, M. D., A. Beger, J. Cutler, M. Dickenson, C. Dorff, and B. Radford. 2013. Comparing GDELT and ICEWS event data, ed. D. University.
- Yonamine, J. E. 2013. Predicting future levels of violence in Afghanistan district using GDELT: UT Dallas.
- Yuan, M. 2009. Toward Knowledge Discovery about Geographic Dynamics in Spatiotemporal Databases. In *Geographic Data Mining and Knowledge Discovery (Second Edition)*, eds. H. J. Miller and J. Han, 347-365 London: CRC Press.
- Yuan, Y. 2017. Exploring the Spatial Decay Effect in Mass Media and Location-Based Social Media: A Case Study of China. In *Advances in Geocomputation: Geocomputation 2015--The 13th International Conference*, eds. D. A. Griffith, Y. Chun and D. J. Dean, 133-142. Cham: Springer International Publishing.

- Yuan, Y., Y. Liu, and G. Wei. 2017. Exploring inter-country connection in mass media: A case study of China. *Computers, Environment and Urban Systems* 62:86-96.
- Yuan, Y., Y. Lu, T. E. Chow, C. Ye, A. Alyaqout, and Y. Liu. 2019. The Missing Parts from Social Media-Enabled Smart Cities: Who, Where, When, and What? *Annals of the American Association of Geographers*:1-14.
- Yuan, Y., M. Raubal, and Y. Liu. 2012. Correlating mobile phone usage and travel behavior - a case study of Harbin, China. *Computers, Environment and Urban Systems* 36 (2):118-130.
- Zhang, Q. Y., G. L. Chen, C. Y. Chang, and C. C. Chen. 2018. MATR: A Mobility-Aware Topology Restructuring Scheme for Bluetooth Body Area Networks. *Journal of Internet Technology* 19 (2):391-402.
- Zook, M., M.-J. Kraak, and R. Ahas. 2015. Geographies of mobility: applications of location-based data. *International Journal of Geographical Information Science* 29 (11):1935-1940.