Computers, Environment and Urban Systems 62 (2017) 86–96

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Exploring inter-country connection in mass media: A case study of China

Yihong Yuan [a,*], Yu Liu [b], Guixing Wei [a]

[a] Department of Geography, Texas State University, San Marcos, TX 78666, USA
[b] Institute of Remote Sensing and Geographic Information Systems, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 28 May 2015
Received in revised form 14 September 2016
Accepted 27 October 2016
Available online 9 November 2016

Keywords:
Time series
Inter-country relations
Spatio-temporal data mining
Mass media events
GDELT

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 model the image of China in mass media, specifically, how China has related to the rest of the world and how this connection has evolved upon time. The results of this research contribute to both the methodological and the empirical perspectives: We examined the effectiveness of the dynamic time warping (DTW) distances in measuring the differences between long-term mass media data. We identified four types of connection strength patterns between China and its top 15 related countries. With that, the distance decay effect in mass media is also examined and compared with social media and public transportation data. While using multiple datasets and focusing on mass media, this study generates valuable input regarding the interpretation of the diplomatic and 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.

1. Introduction

The development of the Internet not only provides a variety of avenues for communication, but it also generates rich data sources from which researchers can analyze human behavior patterns from both individual and aggregated perspectives (Eagle, Pentland, and Lazer, 2009; Liben-Nowell, Novak, Kumar, Raghavan, and Tomkins, 2005). Compared to social media, traditional mass media is characterized by the significance and aggregated nature of associated events (Liebert and Schwartzberg, 1977), including both positive events (e.g., holiday festivals) and negative events (e.g., street riots). As such, mass media data are more suitable for investigating the aggregated pattern than the individual elements of a society, such as how these events affect different geographic regions from a societal level (Batty, 2013). These data are capable of capturing the social, economic, cultural, and, political aspects of a society synergistically. In addition, mass media data are often collected and generated in a longer time span, and are updated on a daily (or even hourly) basis, therefore they are appropriate for modeling both long-term trends and patterns (e.g., the evolvement of a city or a country in decades) and short-term patterns (e.g., real-time update of a riot). Although the rapid development of techniques and theories in the big data era have introduced new opportunities to analyze the large amount of mass media data available online, most existing research focuses on the methodological perspective of event mining such as spatio-temporal pattern recognition (Huang, Zhang, and Zhang, 2008). Relatively few empirical studies have investigated using mass media data to explore the connections between different geographic regions, or exploring how these connections evolve over time (Cohen and Cohen, 2009; Liu, Wang, Kang, Gao, and Lu, 2014b).

Spatial connection (relatedness) has been a hot topic in many research fields, such as transportation, immigration, and political geography. This concept was implied in Tobler's first law of geography (TLF) (Tobler, 1970) "near things are more related than distant things", and covers a broader range of connection than "interaction." Much literature introduces auxiliary information and various data sources to analyze the strength and formation of such connection between geographic entities, such as using traffic flows to measure the strength of ties in origin-destination (OD) locations, or using immigrant flows to measure the bilateral relations between countries (Rodrigue, Comtois, and Slack, 2013; Lewer and Van den Berg, 2008). Undoubtedly, these datasets provide empirical evidences to quantify spatial connection and related geographic phenomenon. However, they also have internal limitations: first, aggregated socio-economic data, such as transportation and trade flows, are often single-faceted, so conceptually they only provide limited aspects of the spatial connection between two entities. Second, these data, although collected during a long time span (i.e., longitudinal data), are not updated in real time; therefore, they often have a coarse temporal resolution (e.g., published on a yearly basis). Normally, there is a time lag between the actual events' happening and when the data become available from authorities. The availability of different datasets also varies in each country based on local policies. Hence, such datasets

* Corresponding author.
E-mail address: yuan@txstate.edu (Y. Yuan).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.012
0198-9715/

have limited capability in quantifying the spatial connection between countries in a more synergized way, and they are difficult to be updated in real-time.

In this study, the open-source dataset "The Global Data on Events, Location and Tone" (GDELT) is employed to analyze the ties and connections between China and other countries as described 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 investigated sufficiently. First, we examine the magnitude of the spatial decay effect in GDELT data and two complementary datasets (Flickr data and Airline Carrier data) based on the gravity model to test the different roles of distance in various circumstances. Second, we construct time series models to analyze the pattern of inter-country connections with respect to time. The dynamic time warping (DTW) distance is adopted due to its capability of dealing with displacements in time series. Based on the constructed DTW distances, we employ clustering analysis to investigate the internal patterns of the series. Although geographic information systems (GIS) are designed to deal with geo-spatial data, many studies attempted to raise awareness of the importance of time in GIS studies. Both "space" and "time" dimensions are considered as fundamental components to interpret a geographic phenomenon or process in GIS (Longley, Goodchild, Maguire, and Rhind, 2005). Hence, in this research we explore techniques for utilizing GDELT to model inter-country relations from both *spatial* and *temporal* perspectives. Here we use "spatial connection" to refer to a broader meaning of relatedness than spatial interaction.

Our ultimate objective is to explore the feasibility of employing mass media data to quantify spatial connection between countries, and how this connection evolves upon time. This research does not aim to provide in-depth interpretation of these connections from a political perspective. Instead, we focus on demonstrating the effectiveness of utilizing mass media data to reveal new insights in a long-lasting question in geography: spatial interaction and connection between countries. Due to the nature of news reports, these data represents more synergized information (i.e., covering multiple aspects of international relations such as economic, political, cultural, military, etc.) than a single-facet dataset (e.g., transportation flow). Our methodology can be considered as a data pre-processing strategy for pattern recognition and interest identification in multiple application areas, such as public relation, sociology and political geography. The results offer valuable insight for policy makers to interpret the long-term geographic dynamics of decision making in international relations, as well as provide references and inputs for analyzing news media in the big data era.

This paper is organized as follows: Section 2 describes related studies in the areas of human activity modeling, event data, and time series analysis. Section 3 illustrates the fundamental research design including the GDELT dataset and methodology. Section 4 presents the data analyses, results, and discusses various aspects of the output in detail. We conclude this research and present directions for future work in Section 5.

2. Related work

2.1. Mass media data in the big data era

The development of World Wide Web (WWW) and Information and Communication Technologies (ICTs) has created a wide range of new spatio-temporal data sources (e.g., social networking check-in data, online news columns), which have allowed for a new paradigm in data analysis and pattern recognition (Wu, Zhi, Sui, and Liu, 2014; Noulas, Scellato, Lambiotte, Pontil, and Mascolo, 2012). A large number of previous studies have focused on utilizing "individual-oriented" datasets such as social networking data, call detailed records (CDRs) and trajectories from Global Positioning System (GPS) devices to analyze correlation and interactions from individual, urban and country level (Liu, Sui, Kang, and Gao, 2014a; Noulas et al., 2012; Yuan, Raubal, and Liu, 2012).

In communication studies, mass media is often defined as the media technologies which target large audiences in a variety of forms, such as radio, newspaper and television (Mazzitello, Candia, and Dossetti, 2007; McQuail, 1979). As discussed in Section 1, compared to "individual-oriented" social media, traditional mass media often concentrates on the significance and aggregated nature of associated events. These datasets focus on specific influential events instead of volunteered individual activities and are, therefore, particularly suitable for analyzing large scale and long-term patterns (Lawson-Borders, 2003). In an attempt to analyze the relationship between social media and traditional mass media, Salman, Ibrahim, Hj-Abdullah, Mustaffa, and Mahbob (2011) discussed the fact that the internet changed how people consume information, but they argue that traditional mass media still plays an important role in interpreting human activities due to its large-scale and selective nature. The development of theories and techniques for big data analytics also offers tremendous flexibility for investigating large-scale events and patterns that emerge over space and time (Schrodt and Gerner, 2000).

Currently, many researchers are analyzing mass-media's impact on both short-term socio-economic events and long-term trends. For instance, Lavrenko (2000) investigated the impact of mass media on economic indicators such as stock market and Romer, Jamieson, and Aday (2003) analyzed the reaction of the population to violence and crime reports. On the other hand, complex systems like cities and countries often evolve gradually over a longer time span. The status of social indicators at different stages can be quantified as time series data. Since mass media news is normally quickly generated with clear timestamps, these data can be particularly useful in tracing and analyzing the emergence of complex systems (Ramsay, Matowe, Grilli, Grimshaw, and Thomas, 2003). For instance, Jiang and Mai (2014) provided an exemplary case study by analyzing the effects of events in a given country on other countries based on GDELT data. They also defined the strength of the causal links between countries based on their findings. This type of analysis provides valuable insight when analyzing the ebbs and flows of international relations. Yonamine (2013) also utilized GDELT data to construct a model which can predict conflict levels in Afghanistan. The proposed model allows for incorporating various socio-political factors such as drug prices, unemployment levels, and ethnic diversity.

This research intends to improve upon these previous studies using GDELT data. Although the processed data size is only in the scale of "millions", the raw data was selected and processed from numerous news media worldwide in real-time; hence, the methodology presents a novel perspective to model international relations in the big data era. It not only considers the role of different countries in mass media, but also focuses on the yearly change of inter-country connections based on time series analysis and the correlation between inter-country connections and their spatial locations. Also, as this article focuses on China, we have an opportunity to examine the changes of international relations in a rapidly-developing country.

2.2. Spatial connection and distance decay effect

As discussed in Section 1, the first objective of this research is to investigate how distance effects inter-country connections in mass media. 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, its simplicity of equation, and its ability to deal with flows in both directions (Rodrigue et al., 2013; Hardy, Frew, and Goodchild, 2012).

In addition, as implied in Tobler's first law of geography (Tobler, 1970), spatial relatedness (connection) is considered as a broader

term than "interaction" when describing the relation between two spatial entities. Researchers have identified different categories of relations that can belong to spatial relatedness, such as spatial proximity, attributive similarity and spatial interaction (Liu et al., 2014b). Gravity models are often utilized to model spatial connection and to discuss the effects of distance on certain types of spatial association. For instance, Hardy et al. (2012) investigated how gravity models can help determine the role of distance in volunteered geographic information (VGI) production and Liu et al. (2014b) used them to explore relatedness between Chinese provinces.

The term "connection" in this research is similar to "relatedness" in previous research, i.e., inter-country connection may capture spatial proximity, interaction and other types of relations. However, we explore both one-way and two-way connections which focus on different aspects of "relatedness". Due to the two-way nature of distance decay effect, the analysis in Section 4.1 considers the "relative importance" (e.g., the total number of records in GDELT) of both countries when normalizing the data. However, the time series analysis in Section 4.2 focuses on how each country plays a different role for China, and how this role changes over time (e.g., for China, the United States may have the highest frequency of "co-occurrence" in GDELT data and thus plays the most important role; however, this observation is not necessarily true vise verse).

2.3. Time series data and dynamic time warping (DTW)

Methodologically, researchers have also applied various techniques to model and compare time series data in mass media. Based on the model construction, further analyses can be conducted such as forecasting, classification, and clustering (Brockwell and Davis, 2002). One important research question in time series analysis is finding whether two time series represent similar behavior (Gunopulos and Das, 2001). Classical distance measures, such as Euclidean distance, are not suitable for measuring the distance between time series data. For example, consider two time series A[1,1,1,1,2,10,1,1,1] and B[1,1,1,1,10,2,1,1,1]: the Euclidean distance between A and B is $\sqrt{128}$. This is a fairly large number for two very similar series. Therefore, researchers investigated more advanced algorithms to measure the similarity between two time series or two curves, such as the Discrete Fréchet Distance (Eiter and Mannila, 1994); however, this method is very sensitive to outliers and displacements (Ahn, Knauer, Scherfenberg, Schlipf, and Vigneron, 2010), hence it is not very appropriate for time series data (Yuan and Raubal, 2012).

As discussed in Section 1, this research also focuses on identifying the temporal patterns of China's relations with other countries based on media event data. In addition to the spatial decay effect in Section 2.2, we also explore the similarity/dissimilarity of event time series based on clustering analysis. This research adopts the dynamic time warping (DTW) distance to quantify the similarity between the time series of connection strengths between countries. DTW was first proposed to find an optimal match between two given time-dependent sequences (Sakoe and Chiba, 1978). Compared to traditional distance measures (e.g., Euclidean distance), DTW is particularly suitable for data with possible displacement in the time dimension (Sakoe and Chiba, 1978; Myers and Rabiner, 1981), and has been widely-applied in various research fields such as speech recognition, motion detection, signal processing, and urban studies for matching two time series (da Costa Filho, de Brito Filho, de Araujo, and Benevides, 2009; Yuan and Raubal, 2012; Senin, 2008; Lee, Han, Li, and Gonzalez, 2008).

Note that the connections in this research between China and foreign countries indicate a correlation not causation, and as discussed in Section 1, this research does not concentrate on interpreting these relations from a political or sociological perspective. Instead, it provides new insight from the perspective of utilizing spatial data mining as a pre-processing tool for social studies.

3. Research design

3.1. Dataset

3.1.1. Main dataset: GDELT

This research utilizes an open dataset named GDELT. This CAMEO-coded dataset[1] (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. The average "tone" of an event is also computed by including all documents containing one or more mentions of it. For instance, a news event that "Chinese people celebrate the spring festival" indicates a positive tone, whereas a news report about border conflict may receive a negative tone score. The score ranges from −100 (extremely negative) to +100 (extremely positive) based on the tonal algorithm in Shook, Leetaru, Cao, Padmanabhan, and Wang (2012).

For instance, in a news report entitled "In Malaysia, Obama carefully calibrates message to Beijing", Actor 1 will be "the United States Government" and Actor 2 will be "the Chinese Government". The associated geographic locations of Actor 1, Actor 2 and the actual action is demonstrated in Table 1. One of the two actor fields may be blank in complex or single-actor situations or contain only minimal details such as "Unidentified gunmen". Here we only consider the cases when the two actors are explicitly identified and geo-tagged.

Note that GDELT data has also raised a considerate amount of controversy in the academic field, especially regarding its data quality and ability to reliably geocode the location of events. For example, the news sources for the GDELT before 2013 are mainly from media published in English, which naturally introduce sampling bias from cultural and regional perspectives. Hammond and Weidmann (2014) compared GDELT with two hand-coded datasets on political violence (Sundberg and Melander, 2013; Raleigh and Hegre, 2009), and they concluded that GDELT should be used with caution for geospatial analyses at the subnational level. However, this study only focused on violence-related events. Another study by Ward et al. (2013) uses a commercial dataset - the World-wide Integrated Crisis Early Warning System (ICEWS) - to validate the GDELT dataset on temporal trajectories at a fixed location. The authors found that GDELT and ICEWS both captured major events in Egypt between 2011 and 2012, but GDELT appears to cover more events than ICEWS. The study by Arva et al. (2013) cross-compared ICEWS and GDELT at the country level, and they reported that the GDELT data performs as well or better than the data in the original ICEWS dataset in detecting hotspots of conflict at the country level.

To further cross-validate country-level geocoding in GDELT, we extract all the records in GDELT and ICEWS that list China as one of the two actors from 2005 to 2010. As noted in Ward et al. (2013), GDELT covers substantially more news records than ICEWS especially after 2005. This is also demonstrated by Fig. 1, where the data size discrepancy between the two datasets increased over time. In 2010, ICEW only covers 1/7 of the data quantity as GDELT. We also conduct a Wilcoxon signed rank test for ICEWS and GDELT data in 2005–2010, which compares the percentage of top 20 countries related

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

Table 1
A sample record from GDELT[a].

Event date	Actor 1_Geo	Actor 2_Geo	Action_Geo	Average tone	Geo_Type
2014-04-26	Washington, District of Columbia, United States	Beijing, China	Kuala Lumpur, Malaysia	2.42	3

Note that the details of location information vary in the records. GDELT measures the level of geographic details by fields 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 US state), 3 = USCITY (a US city or landmark), 4 = WORLDCITY (a city or landmark outside the US), 5 = WORLDSTATE (an Administrative Division 1 outside the US – roughly equivalent to a US state) (Leetaru and Schrodt, 2013). Since this research is conducted at the country level, all records with Geo_Type ≥ 1 should be considered.

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

to China (i.e., co-appearing with China) in news records (H_0 – null hypothesis: For these top 20 countries related to China, there is no significant difference between their probability of co-appearing with China in GDELT and ICEWS). The difference is tested insignificant between ICEWS and GDELT datasets ($p < 0.05$) in terms of country-level geocoding result; however, GDELT has a much larger sample set and therefore is more suitable for aggregated-oriented studies. The concerns of previous studies regarding GDELT's data quality is more from the semantics perspective, i.e., whether an event can be categorized as "violence" or "negative"; however, the objective of this study is to investigate inter-country connection merely based on the geocoded field of GDELT instead of the semantics/tone of news records. Our analysis confirmed the geocoding accuracy of GDELT at the national level, as well as the reliability of GDELT in modeling connection between countries by looking into countries related to China. In addition, starting from April 1st 2013, GDELT data include a new column: source, showing the texts from which this data record is generated. Although it is still difficult to verify early GDELT data, it provides a possibility for the research community to evaluate the machine learning algorithms GDELT currently use, and conduct accuracy check for data after 2014.

3.1.2. Complementary datasets

As illustrated in Section 2.2, besides the main dataset GDELT, we utilize two other complementary datasets to compare the spatial decay effects in different circumstances. Our primary objective is to explore the role of distance in datasets with different natures (i.e., mass media, social media and public transportation data).

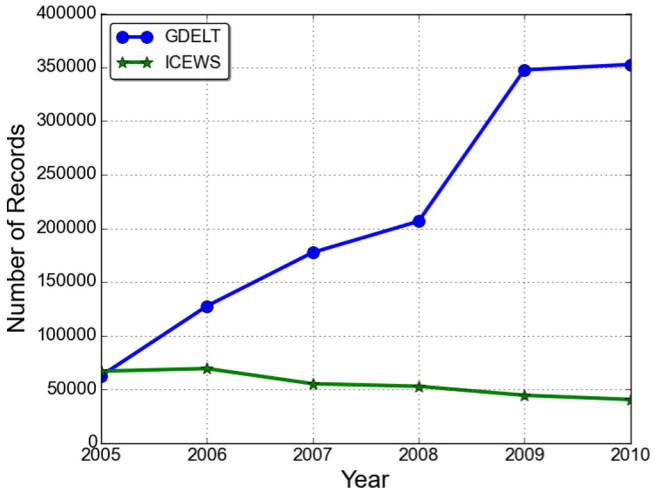

Fig. 1. Number of China-related records in ICEWS and GDELT.

- Flickr data

This dataset covers 31,597,136 geo-tagged Flickr images. The images were randomly sampled globally from the years 2008–2012.[2] Each record captures the geographic coordinates, date, time, unique ID of the uploaded image. The location information of this dataset was acquired through volunteered geographic information (e.g., built-in GPS module of smart phones). Here the interactions between countries are defined as the number of Unique IDs who uploaded images in both countries. The detailed steps of model construction will be illustrated in Sections 3.2 and 4.2.

- Airline carrier data

This dataset includes the number of passengers transported between China and foreign countries in 2008–2012. Note that this data only covers airlines operated in China.[3] However, due to the equality of the freedoms of the air,[4] this data can be considered as proportional to the total number of passengers transported between China and foreign countries. We also utilize the total number of air passengers carried by air carriers registered in each country from the World Bank,[5] which can be considered as the "relative importance" of each country when modeling air transportation interactions in gravity model.

3.2. Methodology

As discussed in Section 2, this research concentrates on the inter-country relatedness between China and foreign countries. The analyses will be conducted from the following three steps:

- Data preprocessing

Due to the large volume of the data records and coded fields in GDELT, it is necessary to preprocess the dataset and select essential information before analysis.

First we extract all news records involving China and another country as two parties. Since this analysis only involves country-level geocoding, all records with country or higher level resolution are included (i.e., Geo_Type ≥ 1). Note that the location of "action" is not considered as a substantial factor here, since an event related to a certain country can happen inside or outside of that country.

Based on the pre-processed data, for each country we calculate the frequencies of "co-occurrence" with China (donated as C) in GDELT. The frequencies are noted as $F_y(i,c)$, which stands for the "co-occurrence" frequency between China and country i in year y.

- Modeling and interpreting spatial decay effects

As illustrated in Section 2.2, the gravity model is utilized in this research to examine the distance decay effect:

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

where P_i and P_j are the "conceptual sizes" (relative importance) of two countries 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. As shown in Section 3.1, we construct three gravity models to compare the best fit of distance friction coefficient β to investigate the underlying

[2] http://labs.yahoo.com/news/yfcc100m/.
[3] Data provided by the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences.
[4] "International Air Services Transit Agreement": https://www.mcgill.ca/files/iasl/chicago1944b.pdf.
[5] http://data.worldbank.org/indicator/IS.AIR.PSGR.

factors of distance decay for three datasets (GDELT, Flickr and Airline carrier data). The specific parameters are defined as follows:

GDELT

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.

Flickr

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.

Airline

I_{ij} The number of passengers carried between i and j.

P_i The number of passengers carried in country i (both domestic and international).

P_j The number of passengers carried in country j (both domestic and international).

Since we provide case studies of China, the model fitting only considers the top 50 countries with the highest I_{ij} between China. 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 the best fit. Since R^2 is scale-free, the constant K does not affect our model fitting. For consistency we use data from 2008 to 2012 for all three datasets in this analysis. As illustrated in previous studies in human mobility, transportation and regionalization (Gonzalez, Hidalgo, and Barabasi, 2008; Liu et al., 2014b), higher β value indicates a stronger distance decay effect. The detailed model fitting results are illustrated in Section 4.1.

- Modeling and interpreting time series data

As discussed in Section 2, a DTW distance is adopted to obtain a robust similarity measure between two series with a high tolerance of element displacement (Brown, Hodgins-Davis, and Miller, 2006). Fig. 2 represents the process of calculating the DTW distance between two example time series (Yuan and Raubal, 2012). First a DTW grid is constructed. Inside each grid cell a distance measure is applied to compare the corresponding elements (here we use absolute differences) of the two time series, and the DTW distance between the two series is considered as a path through the grid which minimizes the total distance.

To explore the similarity and distinctions between countries at a greater detail, we also conduct clustering analysis based on the DTW distance. This research utilizes hierarchical clustering since it can operate directly on the customized distance matrix. In practice, the number of clusters is often determined by specific applications or scenarios. As an exemplary analysis, we adopt the Calinski-Harabasz index (CH index) widely utilized in statistical analyses, which determines the number of clusters based on the combination of within-cluster variance and between-cluster variance. Normally a higher CH index indicates an optimized number of clusters, defined as:

$$\text{CH Index} = \frac{b\_\text{var}/(k-1)}{w\_\text{var}/(n-k)}$$

where between-cluster variance is denoted by $bvar = \sum_{k=1}^{K} n_k||\overline{X}_k - \overline{X}||^2$, within-cluster variance is denoted by $wvar = \sum_{k=1}^{K}\sum_{C(i)=k}||X_i - \overline{X}_k||^2$. $\overline{X}_k$, X_i and n_k are the average value, the ith element and the number of elements in cluster k, and K is the total number of clusters. Section 4 describes the results of distance decay effects, DTW distance calculation and clustering analysis.

4. Analysis results and discussion

4.1. Spatial decay effect

As discussed in Section 3.2, we calculate the best-fit β for three datasets. Figs. 3 and 4 indicate the correlation between the fit β values and the goodness of fit (R^2) of all three datasets.

As can be seen, the three datasets demonstrate distinct patterns for distance decay effect. Distance plays the weakest role in the Flickr dataset ($\beta = 0.12$, $R^2 = 0.9997$). For GDELT dataset the distance decay effect is stronger ($\beta = 0.74$, $R^2 = 0.9252$), whereas the international airline data shows the strongest distance decay effect ($\beta = 1.51$, $R^2 = 0.7926$). Compared the β values obtained by several related studies: 0.2 for Chinese province name co-occurrences on web pages (Liu et al., 2014b), 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 the volunteered geographic information in social networking sites experiences a weaker distance decay than mass media data or public transportation data. This

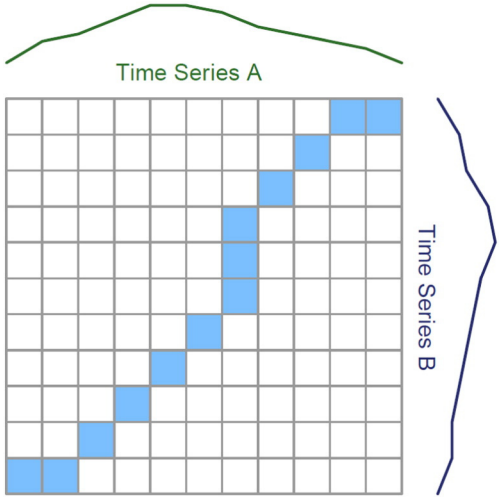

Fig. 2. DTW algorithm.

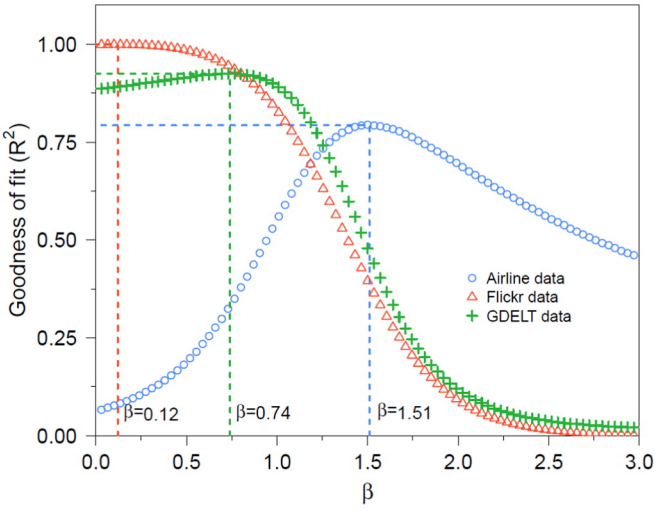

Fig. 3. Fitted β values and the goodness of fit (R^2).

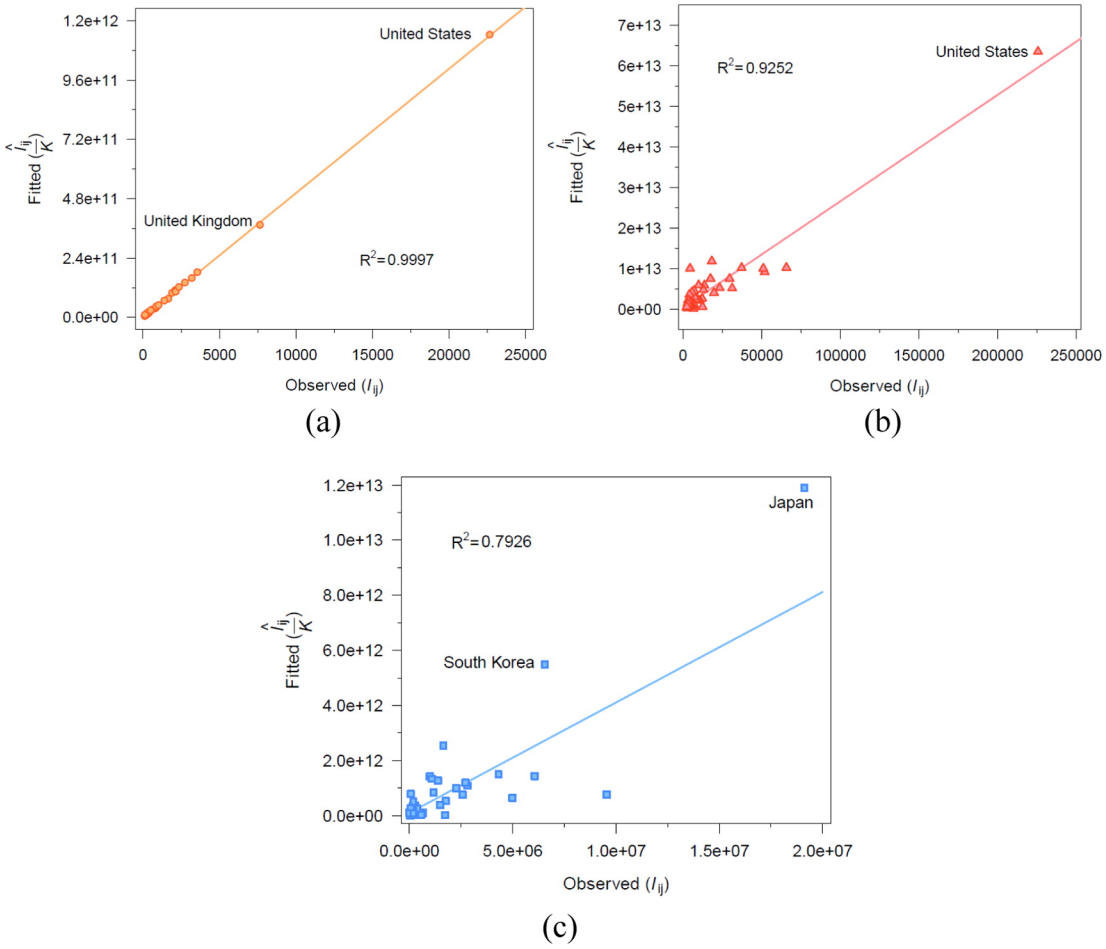

Fig. 4. (a) Observed and fitted I_{ij} (Flickr); (b) Observed and fitted I_{ij} (GDELT); (c) Observed and fitted I_{ij} (Airline).

is potentially due to the data sampling bias in social networking sites (e.g., people are more likely to upload images or post tweets when they travel). Fig. 5 further confirms this pattern by plotting top 15 countries with the highest relatedness ($\frac{I_{ij}}{P_i P_j}$) in three datasets. As seen in Fig. 5a, Switzerland, Australia, and Morocco are generally recognized as popular sites for Chinese tourists. However, the majority of countries in Fig. 5c are nearby or adjacent countries of China, indicating a stronger spatial decay effect in public air transportation. The difference between Fig. 5b and c is also worth noting. As can be seen, GDELT data (i.e., mass media data) demonstrate stronger spatial decay effect than Flickr data (social media), with stronger connection from nearby countries such as Inner Mongolia and South Korea. This is potentially caused by the nature of mass media, where adjacent countries are more likely to be involved in international affairs and then maintain a closer bilateral relation.

4.2. Clustering time series

Besides the spatial component, time series data from news events is another important piece of information for analyzing the correlation between different entities in news reports. Here we first define connection strength as follows:

$$Co_y(i,c) = \frac{F_y(i,c)}{\sum_{j \neq c} F_y(j,c)} \quad (2)$$

where $F_y(i,c)$ is the frequency of co-occurrence between China and I, and $\sum_{j \neq c} F_y(j,c)$ is the total number of records which involves China and another country as two actors. Note that here the connection strength is not normalized by the total occurrence of country I. As mentioned in Section 2.2, unlike the "two-way" spatial decay effect, here we concentrate on the "one-way" effect for China, i.e., how important country I is to China based on the percentage of its co-occurrence with China without considering how important China is for country I. In this way the time series study provides more valuable input for policy makers and planners in China. Hence in this study when referring to "from China's perspective", it indicates that the analysis uses "the total number of records which involves China as one of two actors" as the denominator. We do not aim to provide a perspective from Chinese media's perspective.

Table 2 shows the top 15 countries with the highest connection strength during the study time span 1979–2013 (total number of records 1,271,905 after data pre-processing[6]).

To further explore the changing dynamics of this pattern, we compute the normalized yearly connection strength between China and the top 15 countries. Note that here the time series are normalized to the range [0,1] to capture the relative patterns and eliminate the effect of data magnitude. In relative time series, each country's connection strengths with China are divided by its maximum value between 1979 and 2012. The following series provides an example series between the US and China, which indicates that the connection strength is the highest (1.000) in the year 1982 and the lowest (0.781) in 2006:

- US[0.841,0.902,0.992,1.000,0.983,0.981,0.94,0.917,0.902,0.876,0.883, 0.855,0.839,0.813,0.813,0.836,0.882,0.858,0.852,0.858,0.862,0.831, 0.854,0.831,0.826,0.807,0.794,0.781,0.795,0.811,0.842,0.877,0.906,0.924]

[6] Records involving the province of Taiwan and the special administrative region of Hong Kong are not included due to a data quality issue.

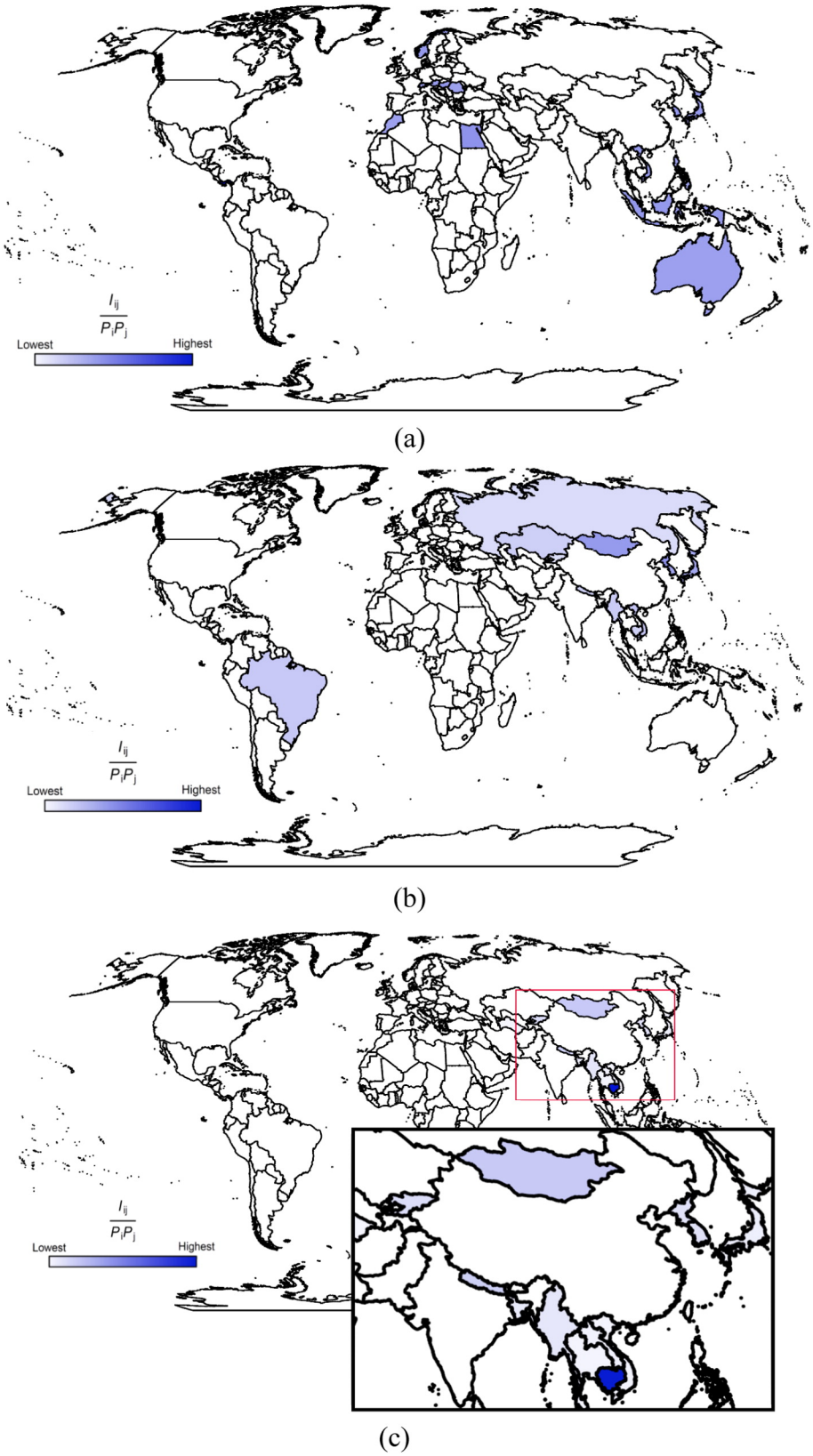

Fig. 5. Top 15 countries with the highest $\frac{I_{ij}}{P_iP_j}$ with China in (a) Flickr data; (b) GDELT data; (c) Airline data with a zoom-in view.

As indicated in Fig. 6, when the number of clusters is equal to four, we obtain the highest CH index; therefore, the time series of 15 countries are divided into four groups in Fig. 7.

Fig. 7a–d indicate four distinct time series patterns based on the result of hierarchical clustering.

- Cluster 1: UK and RP; normalized connection strength first decrease then increase.
- Cluster 2: CA, KS, KN, AS, IN, and IR; normalized connection strength generally increases.
- Cluster 3: RS, FR, US, PK, JA and GM; normalized connection strength

Table 2
Top 15 countries with the highest connection strength with China.

Country	FIPS code[a]	Connection strength $Co(i,c)$	Co-occurrence frequency $F(i,c)$	Distance to China (km)[b]
United States	US	0.179	227,890	11,170
Japan	JA	0.0929	118,194	2099
Russia	RS	0.0808	102,821	5807
South Korea	KS	0.0465	59,165	956
North Korea	KN	0.0424	53,948	812
United Kingdom	UK	0.0414	52,605	8161
France	FR	0.0290	36,880	8238
Iran	IR	0.0238	30,255	5613
Pakistan	PK	0.0236	29,959	3888
India	IN	0.0227	28,935	3784
Australia	AS	0.0219	27,864	8919
Vietnam	VM	0.0193	24,545	2321
Germany	GM	0.0184	23,400	7377
Philippines	RP	0.0152	19,323	2839
Canada	CA	0.0140	17,747	10,476

[a] Federal Information Processing Standards (FIPS) are publicly announced standardizations developed by the United States federal government for use in computer systems.
[b] Distance measured between country capitals.

appears to be stable.
- Cluster 4: VM; normalized connection strength generally decreases.

In addition, Cluster 4 can be considered as an outlier pattern as Vietnam is the only country fall into this category. The four categories provide a more formalized determination of the varying connection strength between China and foreign countries, and Section 4.3 will discuss the indications of these patterns in a greater detail.

4.3. Discussion

The analyses in Sections 4.1 and 4.2 provide valuable insights for quantifying inter-country relatedness from modeling, predicting and clustering perspectives. Although the analyses are conducted from two perspectives (distance decay and time series analysis), they are inherently connected: 1) as mentioned in Section 1, both space and time are fundamental components of GIS (Longley et al., 2005) to a more complete understanding of spatial phenomenon. This research address inter-country connection from both spatial and temporal perspectives; 2) previous research has addressed that the relatedness (connection) of two geographic entities can mainly be explored from two perspectives: interaction and similarity (Liu et al., 2014b). The two analyses in Sections 4.1 and 4.2 can also be viewed as exemplary studies from these two perspectives: spatial interaction from gravity model, and time series similarity from DTW algorithm; 3) Section 4.1 focuses more on the absolute magnitude of inter-country connections (by exploring the frequency of co-occurrence in news, number of photos uploaded, and travel flows); however, Section 4.2 addresses the relative patterns by normalizing the series and removing the effect of magnitude in assessing pattern similarity. Hence, the two sections complement each other by addressing both absolute and relative patterns of inter-country connections.

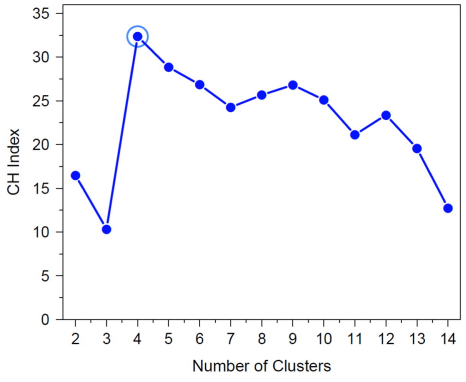

Fig. 6. The correlation between number of clusters and CH index.

From the spatial perspective, the fitted β values confirmed the varying magnitude of spatial decay effects in datasets with different characteristics. Social media data shows the weakest spatial decay effect and air transportation shows the strongest. GDELT data demonstrates an intermediate spatial decay effect between Flickr and airline carrier data. This is further confirmed when investigating the "one-way" connection strength in Table 2. In general, the 15 countries in Table 2 can be grouped into three categories:

- Countries with high level national wealth or political power globally, but not adjacent to or near China: This includes the countries with top 20 Trade-Gross domestic product (GDP) ratio (2003 — 2013) as listed in the Central Intelligence Agency CIA World Factbook (United States. Central Intelligence Agency, 2013.) Countries in this category include United States, United Kingdom, France, Australia, Germany, and Canada.
- Adjacent or nearby Asian countries (Countries that are not necessarily powerful globally, but actively interacted with China from one or multiple perspectives, such as economics, politics, and military forces): Countries in this category include Pakistan, India, North Korea, Vietnam, Iran, and Philippines. Although based on Tobler's first law of geography, spatial adjacency may facilitate international interactions, not all countries adjacent to China can satisfy the criteria of this category (e.g., Nepal, Laos, or Bhutan are not strongly connected to China).
- Countries that combine the characteristics of above two categories (both economically influential from the global level and adjacent to or near China): Countries in this category include Japan, Russia, and South Korea.

In economical geography, Trade-GDP ratio is often utilized as an indicator of openness, which directly impacts the international relations between countries. The second category of countries (adjacent or nearby countries) contribute to the spatial decay effect investigated in Section 4.1. Note that due to the limitation of datasets, the relative importance of countries (P_i, P_j) in gravity models is calculated directly as the total occurrence of each country in GDELT. Researchers have investigated more flexible ways of calculating the relative importance of each entity, such as the inverse gravity model (in which P_i, P_j are estimated dynamically) (Liu et al., 2014b). One of the follow-up studies of this research will cross-validate the gravity models using other solutions (e.g., the inverse gravity models) and test their robustness when more comprehensive data become available.

The results of the clustering analysis in Section 4.2 identified interesting patterns in four clusters:

- In cluster 1, both the United Kingdom and the Philippines demonstrate a connection peak in 1990, which potentially resulted by several major events in Chinese diplomacy. In 1997, the United Kingdom transferred the sovereignty over Hong Kong back to China, which is considered a key event in the history of diplomatic relations for the two countries. Between China and the Philippines, there were a series of territorial disputes in the South China Sea in the 1990s. As shown in Fig. 8, the average tone of the news records dropped during the mid-to-late 1990s.

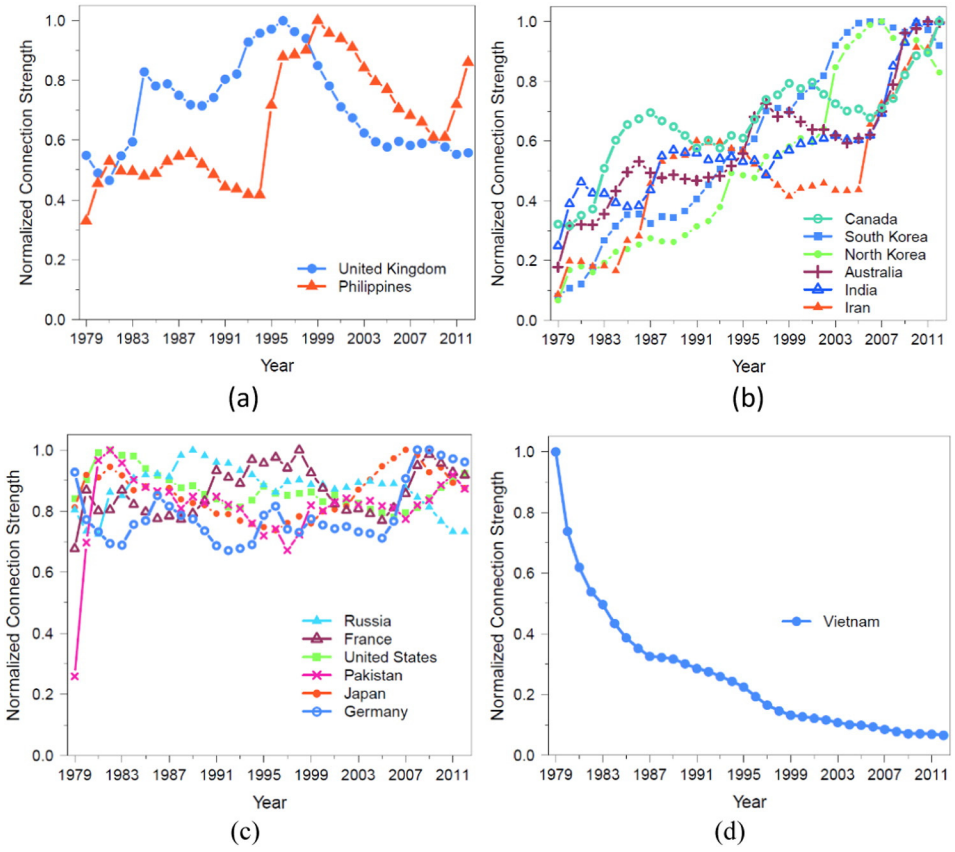

Fig. 7. Four clusters of connection strength. (a) Cluster 1; (b) Cluster 2; (c) Cluster 3; (d) Cluster 4.

Fig. 7a also demonstrated the capability of DTW in matching two similar time series with potential displacement in the time dimension, i.e., the peak time of the two series does not accurately align with each other; however, the clustering analysis based on DTW distance can still recognize similar patterns in this circumstance.

- In cluster 2, all countries indicate an increasing connection with China. In 1979, the Chinese Government published the "Reform & Opening up" policy, which not only resulted in an economy boom within the country, but also stimulated a rapid growth in foreign trade and investment. This potentially explained the increasing patterns for certain countries in Fig. 7b, such as the enhanced trading environment between China and several countries (e.g., South Korea, Australia, and Canada) over the past two decades. Note that the percentage values in Fig. 7 are normalized, hence the increase indicate a higher "share" of connection with China after eliminating the effect of magnitude. To further understand these patterns in cluster 2, we use Australia as a case study. Fig. 9 represents the correlation between GDELT data (normalized connection I_{ij} in Eq. (1)) and trade data (the ratio of exported value to Australia in total export value of China in a 10-year time span) (International Trade Center, 2001–2015). As can be seen, the two data exhibits a linear correlation with Pearson $R^2 = 0.9$. This further supports our hypothesis that the increasing connection between China and Australia may result from the economy boom and "Reform & Opening up" policy in China. This also provides an exemplary analysis for economists and economic

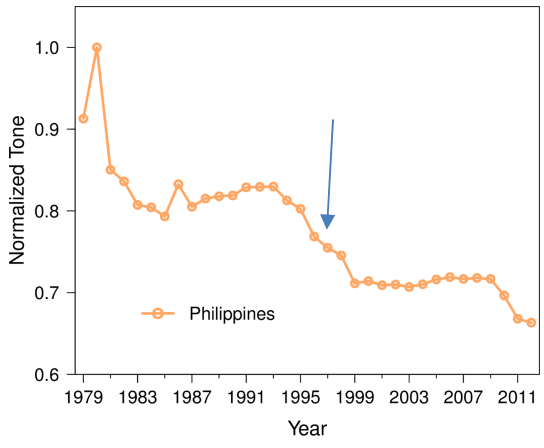

Fig. 8. Yearly average tone between China and Philippines.

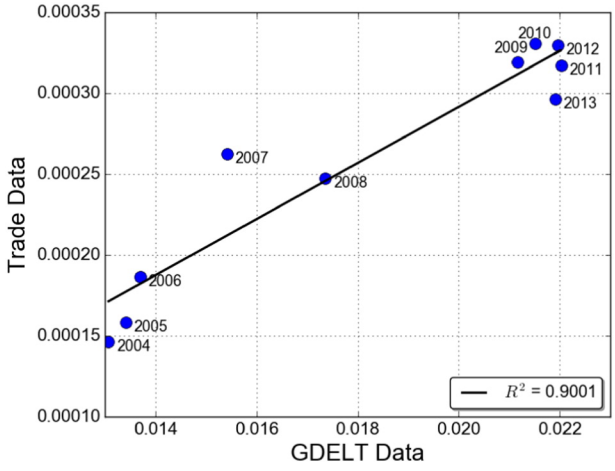

Fig. 9. Correlation between trade data and GDELT data (China vs Australia).

geographers to extract informative bilateral relation change from mass media data.

- Cluster 3 shows the group of countries which demonstrate a more static connection strength such as the connection between the US and China. Note that Pakistan indicates a rapid increase between 1979 and 1981. However, the pattern after 1981 is relatively stable; hence it was grouped into cluster 3 instead of cluster 2. This is consistent with the facts that the economic co-operation between the two countries began in 1979. Many studies have also addressed the continuing collaboration between China and Pakistan since 1980s (Small, 2015).
- Cluster 4 can be considered as an outlier in all 15 countries, since Vietnam is the only country showing a clear decreasing pattern in this study. This can be verified from various international studies focusing on the complicated history between the two countries after the Vietnam War, which is not the major focus of this study. Although today China is still one of Vietnam's largest trading partners, and the economic tie is becoming stronger between the two countries. From China's perspective, the connection strength continues to drop with a smaller and smaller percentage. Fig. 10 indicates that the average tone between China and Vietnam also drops together with the connection strength.

As can be seen, this study is particularly useful for detecting outlier patterns and crucial events in inter-country relations. Additionally, it also reveals interesting patterns regarding "pair-up" patterns. For instance, North Korea and South Korea show a very similar connection pattern with the lowest DTW distance in all country pairs, indicating a possibility that these two countries may interact with each other regarding their connection strength with China.

Note that the top fifteen countries selected in Section 4.2 are based on an aggregated data from 1979 to 2013; however, there are fluctuations in China's international relations during the past three decades. Some countries fell out of the top 15 during certain years, and other countries emerged as China's new partners. Fig. 11 illustrates the top 5 countries that are most connected to China during different time spans. As can be seen, Vietnam dropped out after 1993, which is consistent with the pattern shown in Fig. 7d. On the other hand, the connection between China and South Korea becomes stronger after 1994. This also fits the timeline that the international relations between the People's Republic of China and South Korea were formally established on August 24, 1992 (Holley, 1992).

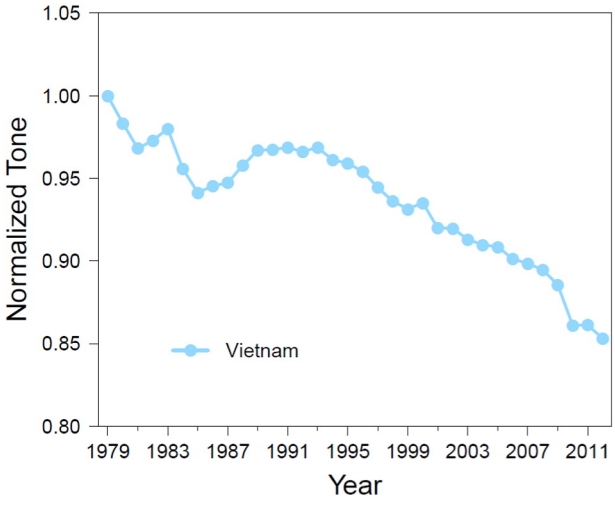

Fig. 10. Yearly average tone between Vietnam and China.

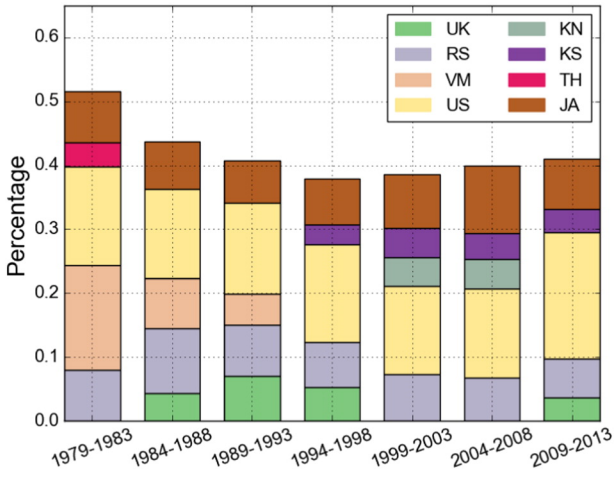

Fig. 11. Top 5 countries connected with China.

This research does not focus on tone analysis due to the concern of the reliability of tone extraction algorithm in GDELT data. For instance, China and Vietnam was in war in 1979, whereas the tone is scored the highest in that year. This is potentially due to the bias introduced by the applied tone algorithm and/or the sampling bias of the GDELT data itself. Many studies in the machine learning field have addressed the needs to improve the accuracy of sentiment analysis and opinion mining (Cardie, 2014; Liu, 2012). One future direction is to cross-validate the extracted tone values in GDELT with more advanced machine learning algorithms.

5. Conclusion

The development of World Wide Web (WWW) has introduced exciting changes in various research fields in the age of instant access. As discussed in the Introduction, the motivation of this research is to explore the usage of mass media in analyzing international relations as a data mining strategy for social sciences. Compared to social media, these data naturally address significant and aggregated events worldwide. They are also collected over a longer time span and are more appropriate for investigating long-term trends and patterns. The objective of this research was to tackle a two-folded question - the feasibility of applying mass media data to analyze inter-country connection with large *spatial* and *temporal* scales. This paper employed the GDELT dataset to examine the connection between China and foreign countries based on time series modeling and clustering analysis. The contributions of this research include:

- The fit β values for three types of datasets (mass media, social media and air transportation) demonstrate that mass media data indicate a stronger distance decay effect than social media (Flickr data) as well as a weaker effect than international air transportation data. This is potentially due to that physical movement has more spatial constrains (more "costly") than virtual information transfer.
- Although DTW has been widely applied in speech recognition and signal processing, this method has rarely been utilized in the area of international relations. The case studies demonstrated the effectiveness of the DTW distance in measuring the similarity of long-term mass media data and identify outlier patterns. Four types of connection strength patterns were identified between China and its top 15 connected countries based on a clustering analysis.
- As shown in Section 4.3, the patterns extracted from GDELT can be validated with secondary datasets, such as trade data and historical event data. It captures ebbs and flows of international relations from various perspectives, such as economic, military, and political (e.g., the return of Hong Kong). This study demonstrates the

effectiveness of applying GDELT and data mining techniques to investigate informative patterns for interdisciplinary researchers. Although this research does not aim to provide in-depth interpretation of the causes and consequences of these inter-nation events from a political perspective, it proposed a method to discover the patterns that can provide insights in different research fields.

Potential future research directions include extending and validating this method to other countries, regions, and machine-coded datasets to test its robustness. GDELT provides a rich data source to analyze inter-region relations at various spatial scales, such as investigating the connection between different provinces in China. Also, further research may involve comparing social media and mass media in an effort to characterize urban-level patterns. Future studies can also look into the correlation between connection strength and various demographic variables such as population, economic status and the tone of each event record.

References

Ahn, H. K., Knauer, C., Scherfenberg, M., Schlipf, L., & Vigneron, A. (2010). Computing the discrete Frechet distance with imprecise input. *Algorithms and Computation, Pt 2*(6507), 422–433.

Arva, B., Beieler, J., Fisher, B., Lara, G., Schrodt, P. A., Song, W., ... Stehle, S. (2013). Improving forecasts of international events of interest. *Annual meeting of the European political science association*.

Batty, M. (2013). *The new science of cities*. Cambridge, Massachusetts: MIT Press.

Brockmann, D., & Theis, F. (2008). Money circulation, trackable items, and the emergence of universal human mobility patterns. *IEEE Pervasive Computing, 7*(4), 28–35.

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to time series and forecasting*. New York: Springer.

Brown, J. C., Hodgins-Davis, A., & Miller, P. J. O. (2006). Classification of vocalizations of killer whales using dynamic time warping. *Journal of the Acoustical Society of America, 119*(3), El34–El40.

Cardie, C. (2014). Sentiment analysis and opinion mining. *Computational Linguistics, 40*(2), 511–513.

Cohen, S. B., & Cohen, S. B. (2009). *Geopolitics: The geography of international relations*. Lanham, MD: Rowman & Littlefield.

da Costa Filho, A. C. B., de Brito Filho, J. P., de Araujo, R. E., & Benevides, C. A. (2009). Infrared-based system for vehicle classification. *Microwave and optoelectronics conference (IMOC), 2009 SBMO/IEEE MTT-S international* (pp. 537–540).

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*(36), 15274–15278.

Eiter, T., & Mannila, H. (1994). Computing discrete Fréchet distance. *Tech. report CD-TR 94/64*. Christian Doppler Laboratory for Expert Systems.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature, 453*(7196), 779–782.

Gunopulos, D., & Das, G. (2001). Time series similarity measures and time series indexing. *SIGMOD Record, 30*(2), 624.

Hammond, J., & Weidmann, N. B. (2014). Using machine-coded event data for the micro-level study of political violence. *Research & Politics, 1*(2).

Hardy, D., Frew, J., & Goodchild, M. F. (2012). Volunteered geographic information production as a spatial process. *International Journal of Geographical Information Science, 26*(7), 1191–1212.

Holley, D. (1992). *South Korea, China Forge Official Ties : Diplomacy: Action increases pressure on North Korea to ease hard-line policies*. Roh plans visit to Beijing. Los Angeles Times.

Huang, Y., Zhang, L. Q., & Zhang, P. S. (2008). A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Transactions on Knowledge and Data Engineering, 20*(4), 433–448.

International Trade Center (2001–2015). International trade in goods-exports 2001–2015. Available from: http://www.intracen.org/itc/market-info-tools/statistics-export-country-product/

Jiang, L., & Mai, F. (2014). Discovering bilateral and multilateral causal events in GDELT. *International conference on social computing, behavioral-cultural modeling, & prediction*.

Lavrenko, V. (2000). Language models for financial news recommendation. *Department of Computer Science*. Amhers, MA: University of Massachusetts.

Lawson-Borders, G. (2003). Integrating new media and old media: Seven observations of convergence as a strategy for best practices in media organizations. *The International Journal on Media Management, 5*(2), 911–999.

Lee, J. -G., Han, J., Li, X., & Gonzalez, H. (2008). Traclass: Trajectory classification using hierarchical region-based and trajectory-based clustering. *International conference on very large data base (VLDB'08). Auckland, New Zealand*.

Leetaru, K., & Schrodt, P. (2013). Gdelt: Global data on events, language, and tone, 1979–2012. *International studies association annual conference. San Diego, CA*.

Lewer, J. J., & Van den Berg, H. (2008). A gravity model of immigration. *Economics Letters, 99*(1), 164–167.

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*(33), 11623–11628.

Liebert, R. M., & Schwartzberg, N. S. (1977). Effects of mass-media. *Annual Review of Psychology, 28*, 141–173.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Chicago, IL: Morgan & Claypool.

Liu, Y., Sui, Z. W., Kang, C. G., & Gao, Y. (2014a). Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PloS One, 9*(1), e86026.

Liu, Y., Wang, F. H., Kang, C. G., Gao, Y., & Lu, Y. M. (2014b). Analyzing relatedness by toponym co-occurrences on web pages. *Transactions in GIS, 18*(1), 89–107.

Longley, P., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2005). Geographical information systems and science. *Chichester*. Hoboken, NJ: Wiley.

Mazzitello, K. I., Candia, J., & Dossetti, V. (2007). Effects of mass media and cultural drift in a model for social influence. *International Journal of Modern Physics C, 18*(9), 1475–1482.

McQuail, D. (1979). The influence and effects of mass media. In D. A. Graber (Ed.), *Media power in politics*. Washington, D.C.: CQ Press.

Myers, C. S., & Rabiner, L. R. (1981). A level building dynamic time warping algorithm for connected word recognition. *Ieee Transactions on Acoustics Speech and Signal Processing, 29*(2), 284–297.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A tale of many cities: Universal patterns in human urban mobility. *PloS One, 7*(5), e37027.

Raleigh, C., & Hegre, H. (2009). Population size, concentration, and civil war. A geographically disaggregated analysis. *Political Geography, 28*(4), 224–238.

Ramsay, C. R., Matowe, L., Grilli, R., Grimshaw, J. M., & Thomas, R. E. (2003). Interrupted time series designs in health technology assessment: Lessons from two systematic reviews of behavior change strategies. *International Journal of Technology Assessment in Health Care, 19*(4), 613–623.

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

Romer, D., Jamieson, K. H., & Aday, S. (2003). Television news and the cultivation of fear of crime. *Journal of Communication, 53*(1), 88–104.

Sakoe, H., & Chiba, S. (1978). Dynamic-programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 26*(1), 43–49.

Salman, A., Ibrahim, F., Hj-Abdullah, M. Y., Mustaffa, N., & Mahbob, M. H. (2011). The impact of new media on traditional mainstream mass media. *The Innovation Journal: The Public Sector Innovation Journal, 16*(3), 1–11.

Schrodt, P. (2012). *Conflict and mediation event observations event and actor codebook V.1.1b3*.

Schrodt, P. A., & Gerner, D. J. (2000). *Analyzing international event data: A handbook of computer based techniques*.

Senin, P. (2008). *Dynamic time warping algorithm review*. Information and Computer Science Department, University of Hawaii at Manoa.

Shook, E., Leetaru, K., Cao, G., Padmanabhan, A., & Wang, S. (2012). Happy or not: Generating topic-based emotional heatmaps for Culturomics using Cybergis. *IEEE 8th international conference on EScience* (pp. 1–6).

Small, A. (2015). *The China-Pakistan axis: Asia's new geopolitics*. London: Hurst & Company.

Sundberg, R., & Melander, E. (2013). Introducing the UCDP Georeferenced Event Dataset. *Journal of Peace Research, 50*(4), 523–532.

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

United States. Central Intelligence Agency (2013). *The CIA world factbook on CD-ROM*. Quanta Press.

Ward, M. D., Beger, A., Cutler, J., Dickenson, M., Dorff, C., & Radford, B. (2013). *Comparing GDELT and ICEWS event data*. D. University.

Wu, L., Zhi, Y., Sui, Z. W., & Liu, Y. (2014). Intra-urban human mobility and activity transition: Evidence from social media check-in data. *PloS One, 9*(5), e97010.

Yonamine, J. E. (2013). *Predicting future levels of violence in Afghanistan District using GDELT*. UT Dallas.

Yuan, Y., & Raubal, M. (2012). Extracting dynamic urban mobility patterns from mobile phone data. *Geographic information science - 7th international conference, 354–367*. Columbus, USA: Lecture Notes in Computer Science, Springer.

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.