A Spatiotemporal Scientometrics Framework for Exploring the Citation Impact of Publications and Scientists

Song Gao
STKO Lab
University of California, Santa Barbara, USA
sgao@geog.ucsb.edu

Yingjie Hu
STKO Lab
University of California, Santa Barbara, USA
yingjiehu@geog.ucsb.edu

Krzysztof Janowicz
STKO Lab
University of California, Santa Barbara, USA
jano@geog.ucsb.edu

Grant McKenzie
STKO Lab
University of California, Santa Barbara, USA
grant.mckenzie@geog.ucsb.edu

ABSTRACT

The research field of scientometrics is concerned with measuring and analyzing science. In practice, this is often done by restricting the impact of publications, journals, and researchers to a mere frequency. However, scientific activities (co-publication, citation, labor mobility) display clear spatiotemporal patterns, and such patterns have rarely been considered in traditional scientometrics. In this work we focus on the study of citations and present a spatiotemporal scientometrics framework to measure the citation impact of research output by taking physical space, place, and time into account. Specifically, we use the statistics of categorical places (institutions, cities, and countries), spatiotemporal kernel density estimations, cartograms, distance distribution curves, and point-pattern analysis to identify spatiotemporal citation patterns. Moreover, we propose a series of s-indices, such as S_institution-index, S_city-index, and S_ country-index to evaluate a scientist's impact as a complement to non-spatial citation indicators, e.g., h-index and g-index. In addition, we have developed an interactive web application which allows users to visually explore research topics, authors, publications, as well as the spread of citations through space and time. Our work offers insights on the role of location in scientific knowledge diffusion.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Scientific Databases—*bibliography database*; I.5.3 [**Clustering**]: Algorithms

Keywords

Spatiotemporal scientometrics, citation impact, s-index

SIGSPATIAL'13, November 05-08 2013. Orlando, FL, USA
Copyright 2013 ACM 978-1-4503-2521-9/13/11$15.00.
http://dx.doi.org/10.1145/2525314.2525368.

1. INTRODUCTION

The power of knowledge and new ideas largely depends on how they spreads out. Some great research findings were forgotten for decades, some spread slowly, while others take the world by storm. Research activities may initially start from a certain region or several places in the world and then spread to other places thus displaying spatiotemporal patterns. In *Nature News*, van Noorden has discussed that only a few number of cities and metropolitan areas in the world are listed either by ranking the number of publications, or by measuring the citation impact, or by geo-locating the addresses of high-impact journals [34]. In many research areas, the number of citations is an important criterion to estimate the impact of a scientific publication. Thus, citation-based metrics have been widely used to quantify the research output of individual scientists. For example, Hirsch proposed the *h-index* as a single number to quantify an individual's scientific research output. According to Hirsch, *"a scientist has an index h if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have fewer than h citations each"* [17].

However when measuring the impact of a publication, a journal, or a scientist, counting the number of citations does not take into account the geospatial and temporal impact of the evaluating target. The spatial distribution of citations could be different even for publications with the same number of citations. Similarly, some work may be relevant and cited for decades while other contributions only have a short term impact. This difference makes it necessary to consider the spatiotemporal influence when evaluating a publication or a researcher. For example, a publication which has 300 citations throughout the world may have a higher influence than a paper with the same number of citations but limited to a single country. Therefore, a spatiotemporal analysis framework can provide an alternative perspective to evaluate a publication or a scientist's influence. How many institutions have done similar work which cites a famous paper? Over how many places have the scientific idea contained in a paper spread out over the years? Is the scientist's influence global or local? Does it cross cultural boundaries. What is the role of physical distance in scientific interactions? Do important activities in scientific interactions, such as co-publications and citations, follow the distance decay function which is a statistical illustration of Tobler First Law of Geography (TFL): *Everything is related to everything else, but near things are more related than distant things* [31]. Can we assume that *publications are more likely to be cited by nearby research institutions than by distant ones*?

In this work, we approach these questions by proposing a theoretical scientometrics framework to evaluate the spatiotemporal citation impact and patterns of scientific publications and researchers.

2. RELATED WORK

In this section we briefly discuss related work required for the understanding of our research.

2.1 Metrics for Evaluating Scientific Output

Many metrics exist for evaluating the value of research output. For publications, the number of total citations is an important indicator. For individuals, there are many h-index-like indicators to quantify the cumulative impact. In [5], nine different variants of the h-index have been compared. Egghe has introduced a g-index as an alternative measure: *"The g-index is the largest number such that the top g articles received at least g^2 citations"* [10]. In addition, the number of *significant papers* is another optional measure. However, the choice of the number of citations which is used to define a *significant paper* is arbitrary. For example, the i10-index which represents the number of publications with at least 10 citations for a research, and has been used in Google Scholar. While it gives a sense of citation impact, using only frequency-based metrics is not sufficient to quantify the spatiotemporal citation impact. It also does not reveal the patterns in which research ideas spread.

2.2 Geospatial Scientometrics

Previous studies show that important scientific activities (e.g., co-publication, citation, labor mobility) display clear spatial patterns [12]. The publications are highly clustered in a few countries and co-publications tend to occur rather domestically than internationally. The US and UK typically rank the top two in terms of their share in the world's publications and citations. It is unclear to what extent their excellent performance can be attributed to the advantage of English language proficiency. When reviewing the spatial patterns at the city-region level, studies show that most scientific activities are concentrated around major metropolitan areas and a few towns established around major universities. In [24], the authors show that about 35% of the total research output was produced by the 30 largest city-regions, such as London, Tokyo-Yokohama, San Francisco Bay Area, Paris, New York and Boston, in both 1996-1998 and 2004-2006 time periods based on the Science Citation Index (SCI) database. Such comparative research presents the spatial structure and the change of the general system of world cities of knowledge. The publisher *Elsevier* using Scopus data ranked the top 10 cities by looking at the average number of citations that a research paper from a city attracted and analyzed the change of relative citation impact between 2000-2008 [34]. For mapping the places of authors who have published papers in high-impact journals, Bettencourt and Kaur analyzed city addresses appearing in Science, Nature, and the Proceedings of the National Academy of Sciences in 1989, 1999, and 2009 and visualize the results in Google Earth[1].

The geospatial effects on collaborative scientific works have also been addressed by many studies. The spatiotemporal constraints and travel costs are considered as main reasons for the decreasing frequency of research collaboration with regard to an increase in physical distance [20]. By semantically structuring publication data, Hu et al. explored the co-publication and some other collaborative relations among scholars distributed at different locations [18]. Some studies found that international co-publications are cited on average more often than domestic co-publications [26, 13]. Batty has studied the geographical arrangement of the highly cited scientists and found a rank-size law existing in the distributions of highly cited institutions and corresponding places or countries [4]. In [6], density maps in geographical space have been introduced into the field of scientometrics in addition to those in more abstract spaces for bibliometric mapping [33]. They produced kernel density maps of European authors who have published highly cited papers on different subjects based on Scopus and Web of Science bibliographic databases.

In addition, literature has also discussed other socioeconomic factors that affect the scientific interactions. For example, Boschma proposed a proximity framework of physical, cognitive, social, and institutional forms to study the scientific interaction patterns. Researchers studied the relationship between each proximity and citation impact by controlling other proximity variables [7]. Moreover, the change of author affiliations over time adds complexity to the network analysis of universities. An approach with thematic, spatial, and similarity operators has been studied in the GIScience research community [1].

3. METHODS

In this section we detail the used methods to analyze the spatial and temporal patterns of citations.

3.1 Analysis for Publications

We first start with analysis methods for publications.

3.1.1 Categorical Place Impact

An intuitive approach to quantitatively measure geospatial impact is based on the hierarchical structure of categorical places, as well as calculation of how many institutions, towns/cities, states/provinces, sub-regions, and countries the citations come from. Multiple classifications of place hierarchical structures are compared in [27] to classify geospatial granularity. The goal is to determine the level of granularity to which the geographical impact of citations is cognitively understandable to scientists and the public. Here we only consider three levels of categorical places, i.e. institution, city/town, and country, since they are the most common components of addresses used by authors in their papers, and easy to collect via existing digital libraries such as the ACM Portal, Elsevier' Scopus, Arnetminer, and Microsoft Academic Search. An example of the three components in an author's address is shown as follows:

Department of Geography, University of California Santa Barbara, Santa Barbara, CA 93106, United States.

[1] http://www.nature.com/news/specials/cities/best-cities.html

In many cases one can directly use address-parsing methods to identify the hierarchical place components for each address (i.e. city: Santa Barbara and country: United States), and then perform statistic analysis to measure the geospatial impact of publications. In other cases, one may need to use reverse-geocoding transfer location coordinates to a more approachable address or place name, and then parse those into different hierarchical place components and repeat the calculation of categorical place frequencies. Algorithm 1 presents the procedure to identify the hierarchical place structures of citations for a publication using the Google Geocoding API[2]. Furthermore, categorical-place citation impact measurement can be implemented both in digital libraries based on different-order administrative divisions[3], e.g., using the ADL gazetteer content standards [16], and on the Web of Linked Data by integrating the ontology of GeoNames[4] with bibliography online web services.

3.1.2 Density Maps and Spatio-Temporal Density Estimation

Kernel density estimation (KDE) [30] has been widely used in spatial analysis to characterize a smooth density surface that shows the geographic clustering of point or line features. The two-dimensional KDE can identify the regions of citation clusters for each cited paper by considering both the quantity of citations and the area of geographical space, compared to the single-point representation which may neglect the multiple citations in the same location. The advantage of density maps is that raw address data can be used without arbitrary aggregations to show the spatial impact results at the urban scale or among inter-urban groups. To calculate the kernel density, we need to apply geocoding to convert text-based place names or address descriptions into coordinates in geographical space. An estimation of the probability density of citations at the location (x, y) on two-dimensional map is given by

$$D(x,y) = \frac{1}{nh^2} \sum_{i=1}^{n} k(\frac{x-x_i}{h}, \frac{y-y_i}{h}) \tag{1}$$

where n is the number of citations within a spatial neighborhood of the location (x, y) and h is the bandwidth which defines the spatial neighborhood of smoothing. k is the kernel functions, e.g., Gaussion kernel, triangular kernel, Epanechnikov kernel, or quartic kernel. In practice, we need to pay attention to what kind of kernel functions and what bandwidths should be chosen for calculating the density of citations. Generally, the plug-in rule is used for selecting the Gaussion-kernel bandwidth [28], and the Freedman-Diaconis rule for other kernels [11], or using data-driven methods by considering the least squares cross-validation for kernel-bandwidth selection [29].

The temporal information (e.g., the publication year) of citations is important as well to detect the trends. There is a broad range of research on spatiotemporal pattern analysis and visualization techniques that has been discussed in previous ACM SIGSPATIAL GIS papers [23]. Mapping citations at different time periods is an intuitive way to detect the temporal changes of spatial patterns. However, using only predefined time intervals (e.g., five years) may neglect the changes within each temporal snapshot, and the division of citations period is also an arbitrary choice. In this work, we propose to create a three-dimensional map of citations in space-time by using spatio-temporal kernel density estimation (STKDE) which simultaneously captures both the spatial patterns and temporal changes. Such STKDE techniques have been used in crime clustering analysis[8, 25], in space-time trajectory analysis[9], and in space-time visual analytics [2].

ALGORITHM 1: A procedure to identify the hierarchical place structures of citations for a publication.

```
Input: A set of institutions' names or addresses ($A_n$),
       the size $n$ equals to the total number of
       citations for a publication.
Output: The array of frequency numbers ($FreNum_p$)
        for different place granularities.
$FreNum_{pi}$ = 0; /* the frequency of institutions */
$FreNum_{pt}$ = 0; /* the frequency of cities or towns */
$FreNum_{pc}$ = 0; /* the frequency of countries */
forall the institution $A_i \in A_n$ do
    /* identifying place structures with geocoding */
    $PC$ = Geocode($A_i$);
    /* denote $PC$ as the geocoding results of place
    components */
    $CountryArray$.append($PC.country$);
    $Locality$=$PC.locality$;
    $L1$=$PC.administrative\_level\_1$;
    $L2$=$PC.administrative\_level\_2$;
    $L3$=$PC.administrative\_level\_3$;
    if ($Locality \neq NONE$) then
        $CityTownArray$.append($Locality$);
    end
    else
        if ($L3 \neq NONE$) then
            $CityTownArray$.append($L3$);
        end
        else
            if ($L2 \neq NONE$) then
                $CityTownArray$.append($L2$);
            end
            else
                $CityTownArray$.append($L1$)
            end
        end
    end
end
$FreNum_{pi}$ = Frequency($An$);
$FreNum_{pt}$ = Frequency($CityTownArray$);
$FreNum_{pc}$ = Frequency($CountryArray$);
$FreNum_p$ = [$FreNum_{pi}$, $FreNum_{pt}$, $FreNum_{pc}$];
return $FreNum_p$
```

The formula for calculating spatio-temporal density is an extension of two-dimensional KDE in space into three-dimensional STKDE in space and time (x,y,t), as:

$$D(x,y,t) = \frac{1}{nh_s^2 h_t} \sum_{i=1}^{n} k_s(\frac{x-x_i}{h_s}, \frac{y-y_i}{h_s}) k_t(\frac{t-t_i}{h_t}) \tag{2}$$

where $D(x, y, t)$ is the density estimation at a space-time

[2] https://developers.google.com/maps/documentation/geocoding

[3] http://www.alexandria.ucsb.edu/gazetteer

[4] http://www.geonames.org

voxel, n is the number of citations, h_s and h_t are the spatial and temporal bandwidths. And k_s and k_t are kernel functions for multivariate probability density estimation with bandwidths h_s and h_t. In this study, we adopt the Epanechnikov kernel described in [30] as

$$K_d(X) = \left\{ \begin{array}{c} \frac{1}{2}c_d^{-1}(d+2)(1-X^TX), if \quad X^TX < 1 \\ 0, otherwise \end{array} \right. \tag{3}$$

where X is multivariate dataset, c_d is the volume of the unit d-dimensional sphere: c_1=2, c_2=π,c_3=4π/3, etc. Therefore, the spatial kernel k_s and temporal kernel k_t for the citation STKDE are given by

$$K_s(u,v) = \left\{ \begin{array}{c} \frac{2}{\pi}(1-(u^2+v^2)), if \quad (u^2+v^2) < 1 \\ 0, otherwise \end{array} \right. \tag{4}$$

$$K_t(w) = \left\{ \begin{array}{c} \frac{3}{4}(1-w^2), if \quad w^2 < 1 \\ 0, otherwise \end{array} \right. \tag{5}$$

The results of STKDE are volume data, i.e., 3D-grids. The visualization of such STKDE directly would require four-dimensional space because of their volumetric data structure consisting with two-dimensions for the geographic space, one for the time and another one for the density estimation scalar. Such volume visualization is not very common in GIScience but very popular in geophysics, geology and medical science, which are common applications of computer graphic techniques [21]. The three main approaches for volume visualization are (1) direct volume rendering by assigning color and transparency to voxels; (2) isosurface that is the equivalent of isoline connecting points of equal value on a two-dimensional map; and (3) volume slicing by planes. The comparison of three methods and more examples of volume visualization are presented in [8, 25, 9]. To sum up, the use of STKDE offers a novel way to understand the spatio-temporal trends of citation impact.

3.1.3 Cartograms Maps

While the KDE approach can show the "hot regions" of citations, it may still be difficult to identify some regions which have a high number of citations in a global map view. For example, a large number of publications and citations are distributed in European countries whose areas are comparatively smaller than countries such as the United States, China, and Australia. The KDE results may be different if one changes the kernel or the spatial resolution. In addition, as we have discussed above, the geographical distribution of research institutions is heterogeneous in space, and some regions will always display high density of citations. What we would like is a self-explanatory visual representation to highlight the spatial distribution of citations in one global map. Cartograms are maps in which the size of geographic regions such as countries or states appear in proportion not to the areas but to their statistical property, and have been widely used in the representation of census results, election votes, disease incidence, and many other socioeconomic data. In this paper, we introduce Gastner and Newman's diffusion-based method for producing citations cartograms, which bring the linear diffusion functions in physics to the calculation of density-equalizing projections [14]. A common diffusion equation and the velocity field are expressed as below:

$$\Delta^2\rho(r,t) - \frac{\partial\rho(r,t)}{\partial t} = 0 \tag{6}$$

$$v(r,t) = -\frac{\Delta\rho(r,t)}{\rho(r,t)} \tag{7}$$

where $\rho(r,t)$ is a density function and $v(r,t)$ is the velocity at the geographic location r and time t, and $\Delta\rho$ is the gradient of the density field.

The calculation of the cartogram involves solving the partial differential equation (6) for $\rho(r,t)$ and then calculating the corresponding velocity field. More detailed algorithm of solving the equation in Fourier space can be found in [14]. The cumulative displacement vector $r(t)$ (indicating both distance and direction) of any point on the map at time t can be calculated by integrating the velocity field.

$$r(t) = r(0) + \int_0^t v(r,t')dt' \tag{8}$$

In the limit $t \to \infty$, the set of displacements of all points on the original map defines the cartogram. Finally, the cartogram is derived by moving all geometric points of boundaries in such a way that the net flow passing through them is zero at all times with the objective of equalizing density during the diffusion process. Multiple open-source codes and tools of making diffusion-cartogram are available at Dr. Mark Newman's website[5]. Before applying this method, we need to choose the resolution of grids and the starting density ρ which will affect the resulting shapes of cartograms. We can try different kernels for calculating density and different grid size to find a resulting map with good readability of local distortion.

3.1.4 Distance Distribution Curves of Citations

As we have mentioned above, scientific activities such as co-publications between institutions have often been constricted by geographical proximity [20]. Here we are also interested in the role of physical distance (using great-circle distance in two-dimensional geographical space) on citing activities. By plotting the distance distribution curves such as probability density functions (PDF) or cumulative distribution functions (CDF) of citation distance between the cited institution location and the citing location, we can get a sense whether the geographical proximity indeed affects citations or not. Moreover, we quantitatively evaluate the spatial impact of publications. To this end, we scale the CDF by multiplying the total number of citations, which will describe not only the citation probability to be found at distance less than or equal to a value, but also the quantity of citations given by

$$C(D) = N_{total} \times Prob(D <= d) \tag{9}$$

For example, let us assume that two papers have the same probability distribution of citation distance, but varying in the total number of citations. The paper having larger quantity should indicate larger spatial impact at the same distance. In addition, we can compare the pairwise quantiles at critical distance intervals from CDFs, e.g., 50, 95, and 100 (largest), and get the corresponding counts of citations, denoted as (C_p, d_p), i.e., C_p citations occur within the p quantile distance d_p.

3.1.5 Spatial Point Pattern Analysis

A citation associated with an institution location being converted into a geographical coordinate can be taken as a

[5] http://www-personal.umich.edu/ mejn/cart/

point-event in geographical space. Therefore, we can make use of statistical analysis of spatial point patterns to study the distribution of citations for publications, to answer questions about spatial patterns as well. For example, does a distribution of citations exhibit clustered or dispersed pattern? Where are the hot-spot regions in which a publication has a high number of citations and is surrounded by other institutions with high values of citations as well.

There are several approaches for spatial clustering or hot-spot analysis, such as K-means, spatial scan statistics [22], Moran's I, Geary's C, Getis-Ord's General G, and Anselin's LISA methods [3]. They are categorized as global or local indicators for detecting clusters and the results may vary with different definitions of neighbors (distance-based or topology-based) and distance matrix, e.g., event-to-event distance, or quadrant-center-to-event distance.

In this study we firstly introduce the mean of citation-to-nearest-citation distance (MC2NCD) to identify the overall spatial citation pattern for publications. The MC2NCD is actually an example of nearest-neighbor analysis in spatial statistics given by

$$\overline{D_{min}} = \frac{1}{N}\sum_{i=1}^{N} d_{min}(r_i) \tag{10}$$

where $\overline{D_{min}}$ is MC2NCD for a publication, N is the total number of citations and $d_{min}(r_i)$ is the nearest-neighbor distance for a citation at the location r_i.

Compared with the expected value under completely spatial randomness (CSR) distribution, we define an average nearest-neighbor-distance (ANND) index given by

$$ANND = \frac{\overline{D_{min}}}{\overline{D_e}} \tag{11}$$

$$\overline{D_e} = \frac{0.5}{\sqrt{N/A}} \tag{12}$$

where $\overline{D_e}$ equals the expected mean of nearest-neighbor distance in the CSR simulation process such that the points are assumed to locate anywhere within the study area; and A corresponds the area of minimum enclosing bounding-box or a convex hull around all citation-points, or it can be a geographical context-awareness value of land-area without considering ocean regions.

If the ANND value is less than 1, the citation pattern exhibits clustering; and if the index is greater than 1, the trend is toward dispersion; while ANND equals 1 and it should be a random spatial distribution. To test the hypothesis, we can calculate the z-score static for the ANND index defined as

$$z = \frac{\overline{D_{min}} - \overline{D_e}}{\sigma} \tag{13}$$

where σ is the expected standard deviation of mean-nearest-neighbor distance under the CSR process. The z-scores and p-values returned by the spatial-point-pattern analysis algorithm tell us whether we can reject that null hypothesis of CSR or not.

Furthermore, several multi-distance-based statistical tests such as K-, F- and G-functions have been proposed for the quantitative analysis of spatial point patterns compared with the null hypothesis of CSR [19]. Ripley's K-function illustrates how the spatial clustering or dispersion of point-event changes when the neighbor-distance varies. Applied to citation-point-pattern analysis, the observed frequency distribution of citations within multi-distance bands is compared to a theoretical Poisson distribution. Both F- and G-function are nearest-neighbor-based approaches. But F-function is based on the distances between randomly chosen points (not the location of any event) and their nearest-neighbor events, while the G-function is based on the distances between the nearest-neighbor events.

In this research, we implement the G-function test with Monte Carlo simulations under CSR in Matlab for analyzing citation-event patterns because this technique only needs the locations of citations and neighbor-distance matrix that we already have from previous analysis of MC2NCD, instead of generating other arbitrary points in geographical space. The G-function is defined as

$$\widehat{G}(d) = \frac{\#\{d_{min}(r_i) \leqslant d, i = 1, ..., N\}}{N} \tag{14}$$

where # is the count of citations, so $\widehat{G}(d)$ represents the proportion of citations within the event-to-nearest-event distance $d_{min}(r_i)$ no great than given distance cutoff d.

Different spatial patterns show different shapes of G-function curves. $\widehat{G}(d)$ rises gradually up to the distance at which most events are spaced and then increase rapidly for evenly-spaced events, while $\widehat{G}(d)$ rises rapidly at short distances and then levels off at larger d-values for clustered events. We suggest using such approach to examine the observed spatial distribution patterns of citations comparing with the expected empirical distribution under CSR.

3.2 Geospatial Index for Individual Scientists

All the techniques introduced above are focusing on the analysis of geospatial citation impact for a publication. In this section, we apply the analysis for an individual scientist's cumulative geospatial impact. Just as the popular *h*-index or *g*-index to quantify an individual's overall scientific impact, we are interested in meaningful and easily computable indicators as a series of *geospatial-indices*. In addition, the spatiotemporal framework above actually contains two categories: space-based and place-based methods. In human discourses, people usually refer to place descriptions for the social and culture understanding of the world rather than spatial coordinates, because space is abstract while the place is more tangible [32, 15]. Therefore, we would like to propose three categorical-place-based *s-indices*: *S_institution-index, S_city-index, and S_country-index* to characterize the geospatial impact of individual scientists. Based on our exploratory analysis of publication citations, we find that generally the magnitude of institutions and cities to which the publications have been cited are similar, and both of them are larger than the number of citing countries. This makes sense since cities and institutions are in a finer geospatial scale, and the total counts of institutions and cities are obviously larger than the number of countries in the world; also scientific publications may not appear in all countries. Thus, we propose different approaches for the s-indices for different granularities of place.

- A scientist has a geospatial-index *S_country* if *S_country* of his or her N_p papers have been cited in at least *S_country* countries.

- A scientist has a geospatial-index *S_city* if *S_city* of his or her N_p papers have been cited in at least $(S\_city)^{\alpha}$

cities/towns.

- A scientist has a geospatial-index *S_institution* if *S_institution* of his or her N_p papers have been cited in at least $(S\_institution)^{\beta}$ institutions.

The values of power-parameter α, β could be varying in different fields of research and their ranges need further empirical studies left for future work. For a given individual researcher, one expects that all these s-indices should increase over time but not all papers will eventually contribute to them. Some papers with limited local citations will not contribute to the growth of the s-index. Such index is a relatively stable measure since it will not be affected by the total number of papers, by self-citations, or by single exceptional papers with very high (or low) citations.

Based on the address-parsing method introduced in algorithm 1, we can calculate the s-indices for individual scientists by identifying the components of citations.

4. EXAMPLES AND EXPERIMENTS

In this section, we apply the proposed methods to understand the spatial and temporal citation patterns to a sample of scientific publications from different domains. We provide a general overview and then focus on Tobler's famous first law of geography paper [31] in detail.

4.1 Datasets

To illustrate the effectiveness of the proposed spatiotemporal scientometrics framework, experiments need to be conducted to analyze existing publication citations. However, popular bibliography database such as ACM Portal, Thomson ISI Web of Science, and Google Scholar Search do not support the massive direct download of citations and author affiliations although they have more complete bibliography data. Therefore, we selected Microsoft's Academic Search (MAS) [6] as the publication data source as it has more complete information about the authors' profiles (especially affiliation information). Although the citations on MAS are by far not as complete as other academic bibliography database, the downloaded citation data can still provide a demonstration for our analysis framework.

Our test dataset contains 20 papers from between 1965-2008 published in different sub-fields, such as geography, ecology, physics, linguistics, and computer science. The criteria for selecting these papers relies on whether most of their citations (approximate 90%) have the first author's affiliated institution information, which is necessary in our experiments. The papers are sorted according to their citations showed in Table 1 and they have 16165 citation records and institution addresses (including duplicate information referring to authors from the same organization) in total. The paper most frequently cited within the set has received 4544 citations.

4.2 GeoSpatial Distribution Results

To better understand the geospatial patterns of citations for these papers, we explore different spatial analysis techniques introduced above to present the diffusion of scientific ideas. Using the categorical-place measurement, we can detect in how many institutions, cities/towns, and countries a publication has been cited. For the most cited papers, we find that they also widely spread out to many places (Figure 1). Some countries like US, UK, Germany, and China are more likely populated because of their higher scientific productivity (see Table 1). Note that all papers in our sample are in English. One could argue that papers written in other languages may have limited international scope but still be highly cited domestically. In addition, based on our experiments, the number of institutions to which the citations have spread could be derived from the number of diffused countries by $0.476 \times N_{country}^{1.71}$ with the goodness of fit $R^2 = 0.97$, which is similar to the result of cities $0.335 \times N_{country}^{1.76}$ with $R^2 = 0.98$. Such exploratory study offer insights on how to set the value range of power-parameters α and β in categorical-place s-indices.

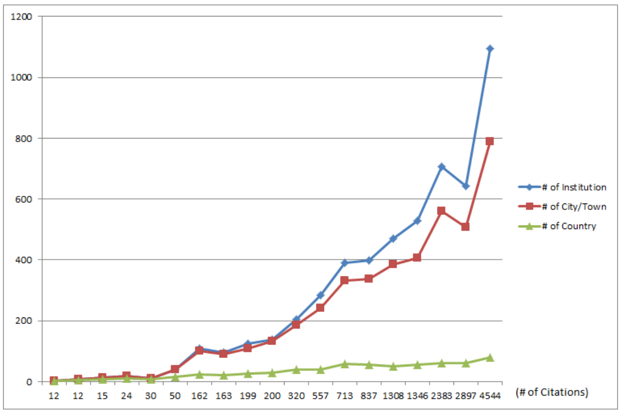


Figure 1: The numbers of citations and the corresponding counts of places.

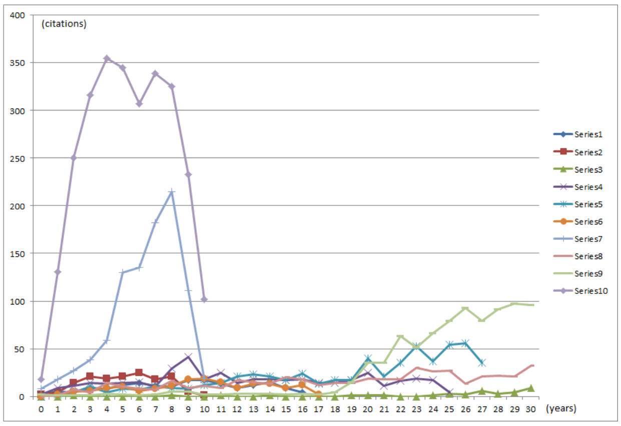


Figure 2: The temporal trends of citations of 10 most cited papers in the dataset.

As we discussed before, papers having the same number of citations may have different spatiotemporal impacts. For example, in the dataset, paper1 (ID=1) and paper2 (ID=2) share the same number of 12 citations, but paper2 has spread over 9 institutions, 8 cities and 6 countries. In this case, we can argue that paper2 has a higher geospatial impact.

When adding time-intervals to the statistic, the temporal variability of citations is also interesting. Figure 2 shows the temporal trends of the citations of the most cited papers

[6] `http://academic.research.microsoft.com`

Table 1: Publication dataset and their total number of received citations, spreading institutions, cities/towns and counties

PaperID	# Citations	# Institution	# City	# Country	Top5 Citing Countries
1	12	3	3	2	UK,US
2	12	9	8	6	US,China,South Korea,Singapore,Germany
3	15	13	13	7	US,Germany,UK,Spain,China
4	24	19	18	11	UK,US,China,Singapore,Canada
5	30	11	11	8	US,Argentina,Germany,Italy,Portugal
6	50	41	41	15	Germany, US,Canada, Japan,Switzerland
7	162	109	100	25	US,Canada,Australia,China,Italy
8	163	95	91	22	US,Germany,UK,China,The Netherlands
9	199	125	110	28	US,UK,China,Italy,Brazil
10	200	139	132	29	US,China,UK,Canada,Italy
11	320	206	185	39	US,UK,Italy,Germany,The Netherlands
12	393	152	138	37	Italy,US,Japan,Germany,France
13	557	284	241	41	US,China,UK,Australia,Canada
14	713	391	332	58	US,Brazil,UK,Canada,Italy
15	837	398	337	57	US,Australia,UK,Canada,Spain
16	1308	470	385	51	US,Canada,UK,China,Italy
17	1346	530	407	55	US,China,UK,France,Germany
18	2383	707	560	62	US,UK,Germany,Canada,Australia
19	2897	644	507	61	US,UK,Italy,Germany,Spain
20	4544	1095	788	81	US,China,UK,Germany,Spain

in our datasets. Some papers were cited directly following the publication year and peaked soon, while others may lag behind but will grow in citation numbers eventually.

In addition, we are interested to understand if the physical distance affects the citation frequency. Firstly, we calculate the average-citing distance and largest-citing distance for all papers. The average of large-citing distance among these 20 papers is 16000 kilometers,which approximates the distance between New York and Sydney. Secondly, we test whether the distribution of citing-distance follows the distance-decay functions by measuring the goodness of fit to the power-law or exponential-law functions (Table 2). At least in our empirical studies, we did not find obvious distance-decay characteristics in citing-distance distributions. Thirdly, we compute the mean of citation-to-nearest-citation distance (C2NCD) and the largest-C2NCD, as well as the average nearest-neighbor-distance (ANND) index, to determine that the citations exhibit clustered or dispersed spatial patterns.

4.3 Case Study with One Paper

Next, we focus on the analysis of the impact of one paper to demonstrate the methods introduced in Section 3.1. we have chosen the paper written by Waldo Tobler in 1970 [31], which is known as The First Law of Geography. It has 670 citation records in the database of Microsoft Academic Search (> 2200 according to Google Scholar) and we have collected 557 records after data cleaning for those lacking location information. The citations are spread out over 284 institutions, 241 cities/towns and 41 countries (Figure 3). The concentration of citations can be found in California, and around the northeast of the US, and in central Europe as well as in China. In the next step we generated the cartogram by country as shown in Figure 4. Such a distortion of boundaries and areas of land helps us to highlight countries with relatively high citations for each paper at first sight, e.g., the enlarged Switzerland and Portugal, versus Spain. In addition, by adding the time (citation-year) to the analysis, Figure 5 depicts the spatiotemporal distributions of citations and the resulting STKDE visualization in a space-time cube. This novel type of analysis offers the possibility to simultaneously detect the citation patterns through space and time.

To explore the the role of physical proximity in citations, we plotted the scaled cumulative distribution functions (CDF) of citation distance, and the histogram of citation distance (see Figure 6). The TFL paper has 279 citations occurring within the average-citing-distance 8755 km and all within the largest-citing-distance of about 16000 km. We also notice that the distance-distribution of citations does not follow the decay function since it has a higher frequency at larger distances.

We also tested whether the spatial patterns of its citations is clustered. Two approaches can be used: the ANND index and the G-function, and both methods need to be compared with the Monte Carlo simulations under CSR. Firstly, the ANND index for this paper is 0.091 and the corresponding z-score is -38.65 (Figure 7). Given such a value, there is a less than 1% likelihood that this clustered pattern could be the result of spatial randomness. Secondly, by plotting the G-function with the nearest-neighbor distance of citations, we

Table 2: The results of different citing distances, nearest-neighbor distances and ANND index

PaperID	(#,ACD)	(#,LCD)	(#,MC2NCD)	(#,LC2NCD)	Distance Decay	ANND
1	(11,282)	(12,5571)	(9,0)	(11,4550)	No	0.8487
2	(6,4059)	(12,16019)	(6,123)	(11,2437)	No	0.4959
3	(8,14762)	(15,18755)	(9,244)	(14,4905)	No	2.6696
4	(12,8486)	(24,12936)	(12,231)	(23,1523)	No	1.267
5	(17,9131)	(30,10754)	(25,0)	(29,1586)	No	0.9735
6	(26,568)	(50,10177)	(24,71)	(49,3917)	No	0.4632
7	(82,3568)	(162,16861)	(81,1)	(161,3763)	No	0.28
8	(82,8642)	(163,16798)	(100,0)	(162,3779)	No	0.2645
9	(101,5259)	(199,18710)	(111,0)	(198,2250)	No	0.2141
10	(100,3953)	(200,18666)	(100,4)	(199,2345)	Yes	0.2357
11	(161,5676)	(320,17043)	(177,0)	(319,2342)	No	0.1548
12	(203,14017)	(393,18710)	(325,0)	(392,3479)	No	0.1104
13	(279,8755)	(557,16908)	(399,0)	(556,1793)	No	0.091
14	(357,6255)	(713,18620)	(464,0)	(712,1684)	No	0.066
15	(421,14906)	(837,18178)	(610,0)	(836,1850)	No	0.06
16	(654,5488)	(1308,17042)	(1243,0)	(1307,1298)	No	0.037
17	(674,6277)	(1346,18630)	(1077,0)	(1345,3581)	No	0.0363
18	(1206,4326)	(2383,18742)	(2067,0)	(2382,1141)	No	0.0215
19	(1449,5081)	(2897,19023)	(2658,0)	(2896,1586)	No	0.0172
20	(2272,5829)	(4544,18710)	(4150,0)	(4543,1752)	No	0.0111

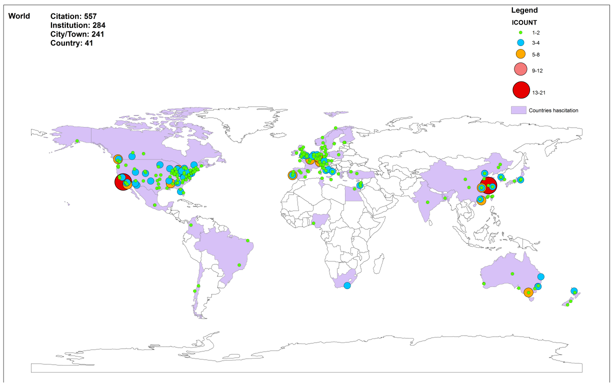


Figure 3: Spatial distribution of citations.

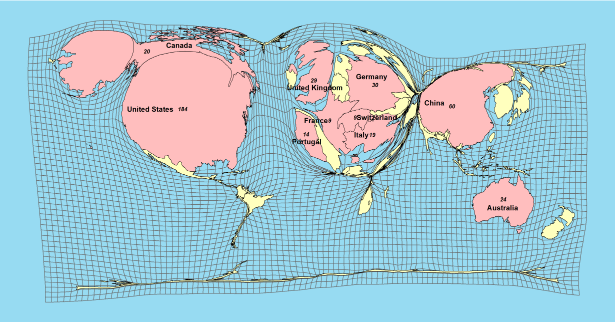


Figure 4: Cartogram of citations by country.

can make the same conclusion that this clustered pattern is significantly different from the Possion point process under CSR.

5. APPLICATION

To give a dynamic visualization of the geospatial impact of scientific outputs, we have developed an interactive web application, called *Citation Map*[7], which allows users to visually explore spatial patterns of citations. By integrating Microsoft's Academic Search and OpenStreetMap, the mashup-application allows users to search publications and their corresponding citations through topic keywords or authors' names, to geolocate publications using the first author's institution, to dynamically map citation information all over the world, to discover the top-ten authors who have cited a publication most frequently, and to share publication and citation information through social media (Figure 8).

6. DISCUSSIONS

This analysis framework relies on the first author's location where the scientific work is assumed to be created or diffused. However, this assumption could bring some potential biases. Firstly, the first author' location might not be the actual place where the research has been conducted considering the contributions of other co-authors or grant issue. Secondly, the authors may change their institutions when they visit or move to another institution and some authors

[7] `http://stko-work.geog.ucsb.edu:8080/map`

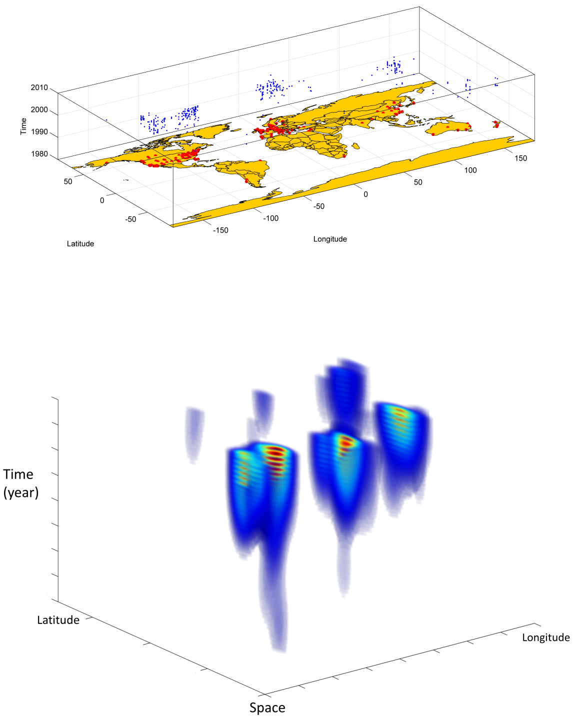


Figure 5: Spatio-temporal distribution of citations and slicing volume-rendering visualization.

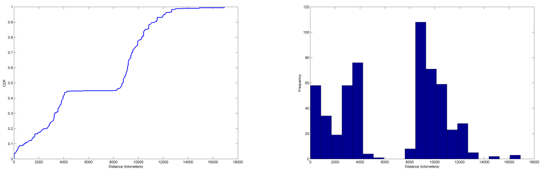


Figure 6: Cumulative citation-distance distribution and Histogram of citation-distance (binned at 1km).

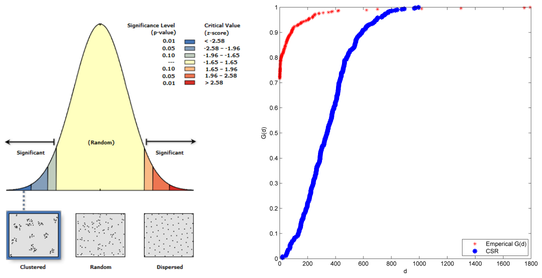
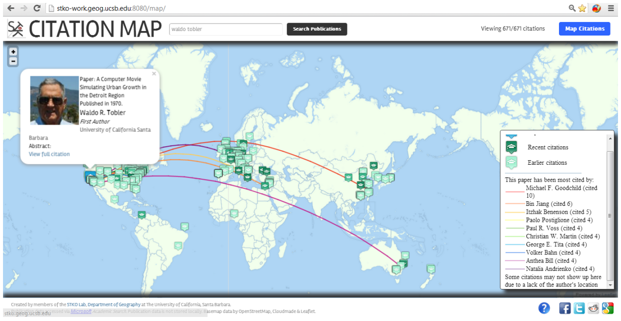


Figure 7: Cumulative citation-distance distribution and Histogram of citation-distance (binned at 1km)



Paper: A Computer Movie Simulating Urban Growth in the Detroit Region
Published in 1970.
Waldo R. Tobler
First Author
University of California Santa Barbara
Abstract:
View full citation

This paper has been most cited by:
Michael F. Goodchild (cited 10)
Bin Jiang (cited 6)
Itzhak Benenson (cited 5)
Paolo Postiglione (cited 4)
Paul R. Voss (cited 4)
Christian W. Martin (cited 4)
George E. Tita (cited 4)
Volker Bahn (cited 4)
Anthea Bill (cited 4)
Natalia Andrienko (cited 4)
Some citations may not show up here due to a lack of the author's location

Created by members of the STKO Lab, Department of Geography at The University of California, Santa Barbara.
Publication data accessed via Microsoft Academic Search. Publication data is not stored locally. Basemap data by OpenStreetMap, Cloudmade & Leaflet.
stko.geog.ucsb.edu

Figure 8: The interface of the CitationMap website.

have simultaneously different addresses, or joint appointments in different countries. Thirdly, the spread of citations to other institutions could also appear when the co-authors cite their own previous work in different organizations.

In addition, given the volume limitation of the citation dataset in our experiments, it is insufficient to draw a more general conclusion about the value distributions of the geospatial measures for papers and s-indices for individual scientists.

Last but not the least, each bibliographic database covers only part of papers from a scientist and part of all citations of a paper, and therefore integrating multiple bibliographic sources may yield more realistic and general impact patterns.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a spatiotemporal scientometrics framework to explore the citation impact of publications as well as individual researchers. Compared with existing scientometrics approaches which often focus on the number of citation, this framework takes into account the distribution of citations in space, places, and time. We proposed a combination of categorical places, spatio-temporal kernel density estimations, cartograms, distance distribution curves, and point-pattern analysis to identify geospatial citation patterns of publications. Based on our empirical experiments, unlike the co-publication activity, we did not find the distance-decay characteristics occurring in the citation patterns.

Moreover, We propose three s-indices ($S\_institution - index$, $S\_city - index$, and $S\_country - index$) to evaluate an individual scientist's geospatial impact, which complement traditional non-spatial measures such as h-index and g-index. An interactive web application has been developed, which visualizes the geospatial distribution of research topics, authors, publications, as well as the spread of citations through space and time.

In the future work, we will collect more citation data and test the framework in different research domains. We assume that other spatial citation indices will emerge.

8. REFERENCES

[1] P. Agarwal, R. Béra, and C. Claramunt. A social and spatial network approach to the investigation of research communities over the world wide web. In *Geographic Information Science*, pages 1–17. Springer, 2006.

[2] G. Andrienko, N. Andrienko, U. Demsar, D. Dransch, J. Dykes, S. I. Fabrikant, M. Jern, M.-J. Kraak, H. Schumann, and C. Tominski. Space, time and visual analytics. *International Journal of Geographical Information Science*, 24(10):1577–1600, 2010.

[3] L. Anselin. Local indicators of spatial association-LISA. *Geographical analysis*, 27(2):93–115, 1995.

[4] M. Batty. The geography of scientific citation. *Environment and Planning A*, 35(5):761–764, 2003.

[5] L. Bornmann, R. Mutz, and H.-D. Daniel. Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, 59(5):830–837, 2008.

[6] L. Bornmann and L. Waltman. The detection of "hot regions" in the geography of science-A visualization approach by using density maps. *Journal of Informetrics*, 5(4):547–553, 2011.

[7] R. Boschma. Proximity and innovation: A critical assessment. *Regional studies*, 39(1):61–74, 2005.

[8] C. Brunsdon, J. Corcoran, and G. Higgs. Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems*, 31(1):52–75, 2007.

[9] U. Demšar and K. Virrantaus. Space–time density of trajectories: exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science*, 24(10):1527–1542, 2010.

[10] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[11] D. Freedman and P. Diaconis. On the histogram as a density estimator: L2 theory. *Probability theory and related fields*, 57(4):453–476, 1981.

[12] K. Frenken, S. Hardeman, and J. Hoekman. Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics*, 3(3):222–232, 2009.

[13] K. Frenken, R. Ponds, and F. Van Oort. The citation impact of research collaboration in science-based industries: A spatial-institutional analysis. *Papers in Regional Science*, 89(2):351–271, 2010.

[14] M. T. Gastner and M. E. Newman. Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences of the United States of America*, 101(20):7499–7504, 2004.

[15] S. Harrison and P. Dourish. Re-place-ing space: the roles of place and space in collaborative systems. In *Proceedings of the 1996 ACM conference on Computer supported cooperative work*, pages 67–76. ACM, 1996.

[16] L. L. Hill. Core elements of digital gazetteers: placenames, categories, and footprints. In *Research and Advanced Technology for Digital Libraries*, pages 280–290. Springer, 2000.

[17] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46):16569, 2005.

[18] Y. Hu, K. Janowicz, G. McKenzie, K. Sengupta, and P. Hitzler. A Linked-Data-driven Semantically-enabled Journal Portal for Scientometrics. In *International Semantic Web Conference (2013; forthcoming).*

[19] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan. *Statistical analysis and modelling of spatial point patterns*, volume 70. Wiley-Interscience, 2008.

[20] J. S. Katz. Geographical proximity and scientific collaboration. *Scientometrics*, 31(1):31–43, 1994.

[21] A. Kaufman. Volume visualization. *The Visual Computer*, 6(1):1–1, 1990.

[22] M. Kulldorff. A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496, 1997.

[23] S. K. Lodha and A. K. Verma. Spatio-temporal visualization of urban crimes on a GIS grid. In *Proceedings of the 8th ACM international symposium on Advances in geographic information systems*, pages 174–179. ACM, 2000.

[24] C. W. Matthiessen, A. W. Schwarz, et al. World cities of scientific knowledge: Systems, networks and potential dynamics. an analysis based on bibliometric indicators. *Urban Studies*, 47(9):1879–1897, 2010.

[25] T. Nakaya and K. Yano. Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS*, 14(3):223–239, 2010.

[26] F. Narin, K. Stevens, and E. S. Whitlow. Scientific co-operation in Europe and the citation of multinationally authored papers. *Scientometrics*, 21(3):313–323, 1991.

[27] D. Richter, K.-F. Richter, and S. Winter. The impact of classification approaches on the detection of hierarchies in place descriptions. In *Geographic Information Science at the Heart of Europe*, pages 191–206. Springer, 2013.

[28] D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*, volume 383. Wiley, 1992.

[29] S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 683–690, 1991.

[30] B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC Press, 1986.

[31] W. R. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46:234–240, 1970.

[32] Y.-F. Tuan. *Space and place: humanistic perspective.* Springer, 1979.

[33] N. J. Van Eck and L. Waltman. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538, 2010.

[34] R. Van Noorden. Cities: Building the best cities for science. *Nature*, 467(7318):906, 2010.