A wavelet analysis of scientific papers and time dependent factors: A scientometric look

Khalid Zaman1*

1 Department of Management Sciences, COMSATS Institute of Information Technology, Abbottabad, Pakistan

* Corresponding author: khalidzaman@ciit.net.pk

Abstract

In this study, we use wavelet approach to study the scientific research papers received by two journals i.e., i) Journal of Informetrics and ii) Research Policy, by using monthly observations during the period of 1970:2 to 2010:2 from article received history of both publishers database. Instead of analyzing the time series at their original level, as it is usually done, this study decompose the two variables i.e., scientific papers and the time dependent factors at various scales of resolution using wavelet decomposition and then we study the relationships among the decomposed series matched to their scales. A major finding of the study is that some variations in received articles explained by many other time dependent factors that are previously not captured due to uneven time lags, therefore, an estimate with a time trend is made in this study. Both variables are not constant over frequency bands as it depends on the time scales. The two series (i.e., Springer and time factors) are correlated at higher frequencies, but their correlations measures in time dimension do not last long. However, during two structural breaks i.e., 1990-1995 and 2007-2011, the two time series seems to be correlated for longer period with frequency band higher than 321 days. The results further provides an evidence that the two time series are in phase and the Elsevier scientific papers are time dependent in the long run. Thus, the analysis of wavelet coherence reveal that the two series i.e., scientific papers and time dependent factors correlated during the sample period.

Keywords: Scientific Papers, Time Factors, Springer, Elsevier, Wavelet Approach

ARTICLE INFORMATION

Received: 15 August 2018
Revised: 25 August 2018
Accepted: 25 October 2018
DOI: 10.31580/jei.v5i4.101

Introduction

There was tremendous growth in the scientometric literature in the 1960s (Herubel, 1999) and from this point forward the field of scientometrics developed and differentiated into several specializations. These were brought together by the launch of the first journal devoted to the field, “Scientometrics”, founded and edited by Tibor Braun of the Hungarian Academy of Sciences. One of the most notable developments was citation analysis. Once a laborious manual job few scholars would engage in, the emergence of (print) databases allowed citation patterns to be studied with relative speed and ease. Scientometrics is the quantitative study of research. Scientometrics examines how researchers pursue scientific research, including how and where they publish (Lowry et al., 2007).

Surulinathi et al. (2013) analyze development of Knowledge Management Research in India in terms of publication output as reflected in Scopus database during 1999-2007. The publication behavior indicates that the Knowledge management researchers were lowly selective in publishing. Serenko (2013) conduct a meta-analysis of prior scientometric research of the knowledge management (KM) field. Around one hundred and eight scientometric studies of the KM discipline were subjected to meta-analysis techniques. He overall volume of scientometric KM works has been growing, reaching up to ten publications per year by 2012, but their key findings are somewhat inconsistent. Most scientometric KM research is published in non-KM-centric journals. The KM discipline has deep historical roots. It suffers from a high degree of over-differentiation and is represented by dissimilar research streams. According to Bauer et al. (2013, p. 524), “Science policy and academic environments make heavy use of assessments of scientific performance, regarding various parameters (e.g., funding, tenure, paper publication) as well as dimensions (individual, institutional, national). Expert judgment is an important part of these assessments. However, it is resource-intensive and, therefore, increasingly complemented—sometimes even substituted—by scientometric indicators that seem easily obtained and suggest quantitative precision”. Mukhtar and Nabi (2013) evaluate the output of world literature on intellectual property that are researched for publications and are indexed in the Science Citation Index (SCI) database during the period from 2001 to 2010. The results reveal that the average number of publications published per year was 143.9 papers. The highest number of papers, i.e. 172 were published in the year 2010. The countries involved in producing publications were 74 in number. United States was at the top of producing 572 (39.75%) publications followed by Great Britain with 119 (8.23%) and India stood at the eighth position with 54 (3.75%) publications. Yang et al. (2013) analyzed the growth and development of Municipal solid waste (MSW) research productivity in China in terms of publication output as reflected in science citation index for the period 1997-2011. The study revealed that the output of MSW research in China has rapidly increased over the 15 years in contrast with USA. Chinese authors
contributed 730 publications out of which 708 were journal articles, 17 reviews, 3 editorial materials, 1 correction and 1 meeting abstract, from 421 institutions. Schreiber (2013) opine that the increase of the h-index with time often depends for a long time on citations to rather old publications. This inert behavior of the h-index means that it is difficult to use it as a measure for predicting future scientific output. Wang et al. (2013) investigated the 31-year life time non-self-citation processes of all Thomson Reuters Web of Science journal papers published in 1980. The correlation between non-self-citation counts in each time window and total non-self-citations in all 31 years is calculated, and it is lower for more highly cited papers than less highly cited ones. There are significant differences in citation ageing between different research fields, document types, total citation counts, and publication months. However, the within group differences are more striking; many papers in the slowest ageing field may still age faster than many papers in the fastest ageing field. Bentley and Kyvik (2013) examine self-reported research hours for full-time faculty at research universities in 13 countries (i.e., Argentina, Australia, Brazil, Canada, China, Finland, Germany, Italy, Malaysia, Norway, UK, USA, and Hong Kong, a semi-autonomous special administrative region of China). The study examine the level of variation in individual faculty research time and the factors associated with individual differences. The results suggest that the factors associated with additional research time vary across countries, but individual motivation towards research (relative to teaching) is a significant in all countries. University policies towards research and the research status of individual faculty, are relatively weak predictors of individual research time, though stronger effects are generally found in English-speaking countries. Research hours typically decrease with age, but plateau or increase in the oldest cohorts. Family and gender are weak predictors of research time amongst full-time faculty. Magnone (2013) employed the application of scientometric methodology to the analysis of scientific communication, relationships between number of submissions of scientific articles and calendar events (e.g., festive seasons, weekend vacations, national public holidays, Chinese New Year, Christmas) are examined quantitatively. Data records are collected during twenty-year (1990–2010) and one-year periods – as case study – ended 31 December 2008. The analysis shows that the overall submission rates are strongly influenced by calendar events.

The above discussion confirms the strong correlation between scientometric studies and time dependent factors. In this study an effort has been made to find empirically investigate the comovement between scientometric studies and time depended factors in two largest publishers i.e., Springer and Elsevier, by using wavelet coherence and wavelet coherence phase differences. After introduction which is presented in Section 1 above, Section 2 shows the methodology. Results are discussed in Section 3. Final section concludes the study.

Methodology

In contrast to Fourier analysis which lost time information and requires stationary time series, wavelet methodology captures the simultaneous examination of the behavior of a time-series in both frequency and time space. Although the wavelet theory born in 1980s (Grossmann and Morlet (1984); Goupillaud et al. (1984)), the wavelet applications to economic data is recent (see e.g. Conner and Rositer (2005); Gençay et al. (2005); Crowley (2007); Anguier-Conraria and Wen (2007); Gronwald (2009); Tonn et al. (2010); Anguier-Conraria and Soares (2011); Tiwari (2013) etc). Wavelet analysis has been devised to analyze signals with rapidly changing spectra. In some sense, wavelet analysis is close to the windowed short-term Fourier transform (Gabor (1946)), especially when using the Morlet wavelet, but the major difference is that the size of the window is fixed for the short-term Fourier transform and it is adapted to the frequency of the signal in wavelet analysis. Because of this difference, wavelet analysis has a more accurate time-frequency resolution (see Lachaux et al. (2002)). Thus the wavelet transform is a useful tool for the analysis of a non-stationary time-series at different frequencies over various sliding windows of time. The wavelet transform decomposes a time series in terms of the wavelet functions \( \psi(t) \) that depends on time parameter \( t \). In our work, we use continuous wavelet analysis tools, mainly cross wavelet transform measuring the degree of local correlation between two time series in the time-frequency domain, and wavelet coherence phase differences. Before introducing wavelet coherence, let us provide basic definitions of continuous wavelet transform and Morlet wavelet. The following description is extracted from Grinsted et al. (2004).

Continuous wavelet transform

The continuous wavelet transform, with respect to the wavelet \( \psi \), is a function \( W_s(t) \) defined as:

\[
W_s(t) = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{s}} \psi^* \left( \frac{t - \tau}{s} \right) dt.
\]

Where * denotes the complex conjugation. The parameter \( s \) is the scaling factor that controls the length of the wavelet and \( \tau \) is a location parameter that indicates where the wavelet is centered. Scaling a wavelet simply means stretching it (if \( |s|>1 \), or compressing it (if \( |s|<1 \)). If the wavelet function \( \psi(t) \) is complex, the wavelet transform \( W_s \) will also be complex. The transform can then be divided into the real part \( \text{Re}(W_s) \) and imaginary part \( \text{Im}(W_s) \) or amplitude, \( |W_s| \) and phase, \( \tan^{-1} \left( \frac{\text{Im}(W_s)}{\text{Re}(W_s)} \right) \). The phase of a given time series \( x(t) \) is parameterized in radians, ranging from \(-\pi \) to \( \pi \). In order to separate the phase and amplitude information of a time-series, it is important to make use of complex wavelets. Just like with the Fourier transform, under some regularity conditions, we can construct \( x(t) \) from its continuous wavelet transform.

The morlet wavelet

The minimum requirements on a function \( \psi(t) \) to qualify for being a mother wavelet are that \( \psi(t) \) is a square integrable function and

\[
\int_{-\infty}^{\infty} |\psi(f)| df < \infty
\]

satisfies the admissibility condition:

\[
|f| \rightarrow \infty
\]

where \( |\psi(f)| \) is the Fourier transform of \( \psi \). Therefore, \( \psi(t) \) has to wiggle up and down the t-axis, behaving like a wave, justifying the choice of the term wavelet.

Given a mother wavelet \( \psi(t) \), a daughter wavelet \( \psi_{ab}(t) \) at scale \( a \) and position \( b \) is defined as

\[
\psi_{ab}(t) = \frac{1}{\sqrt{|a|}} \psi \left( \frac{t - b}{a} \right),
\]

where \( a \) and \( b \) are real. This function is square integrable over the range of real time or space. The component \( a^{-1/2} \) is a normalization factor to make sure that wavelet transforms are comparable across scales and time series.

There are several wavelet functions available, such as Morlet, Mexican hat, Daubechies, etc. We choose a complex Morlet wavelet as it yields information on the amplitude and phase, both essential to study synchronism between two different time-series. The Morlet wavelet is defined as the product of a complex exponential wave and a Gaussian envelope:

\[
\psi(t) = \pi^{-\frac{1}{4}} e^{\frac{-|t|^2}{2}} e^{i\omega t},
\]

where \( \omega \) is the non-dimensional frequency.

The wavelet power spectrum

Typically one has to deal with discrete time-series.
The wavelet power spectrum is simply $|W_n(s)|^2$. The wavelet power spectrum characterizes the distribution of the spectral density of a time-series across the two-dimensional time-scale plane, leading to a time-scale (or time-frequency) representation. Although Torrence and Campo (1998) have shown that the statistical significance of wavelet power can be assessed against the null hypothesis that the data generating process is given by an AR(0) or AR(1) with a certain background power spectrum $P_f$, for more general one has to rely on Monte-Carlo simulations. We assess the statistical significance of the wavelet power against the null hypothesis that each variable follows an ARMA(p,q) process, with no pre-conditions on $p$ and $q$. These simulations are done using the amplitude adjusted Fourier-transformed surrogate proposed by Schreiber and Schmitz (1996).

The cross wavelet transform (XWT)

The cross wavelet transform of two time series $x = \{x_n\}$ and $y = \{y_n\}$ is defined as $W_{n}^{xy} = W_{n}^{x}W_{n}^{y*}$. The cross wavelet power is given by $|W_{n}^{xy}|^2$. While the wavelet power spectrum depicts the variance of a time-series, with times of larger variance showing larger power, the cross wavelet power of two time-series depicts the covariance between the time-series at each scale or frequency. The complex argument $\arg(W_{n})$ can be interpreted as the local relative phase between $x$ and $y$ in time frequency space. The theoretical distribution of the cross wavelet power of two time series with background power spectra $P_{x}$ and $P_{y}$ in Torrence and Campo (1998) as

$$D\left(\frac{W_n^{xy}(s)W_n^{y*}(s)}{\sigma_x \sigma_y}, \sigma_x, \sigma_y\right) < p = Z(p) \sqrt{p_{x}p_{y}}.$$

Where $Z(p)$ is the confidence level associated with the probability $p$ for a pdf defined by the square root of the product of two $\chi^2$ distributions.

Wavelet Coherency (WTC)

As in the Fourier spectral approaches, Wavelet coherency (WTC) can be defined as the ratio of the cross-spectrum of each series, and can be thought of as a local correlation, both in time and frequency between two time series. Thus, WTC near one shows a high similarity between the time series, while coherency near zero shows no relationship. While the wavelet power spectrum depicts the variance of a time-series, with times of large variance showing large power, the cross wavelet power of two time-series depicts the covariance between these time-series at each scale or frequency. Following Torrence and Webster (1999) the WTC of two time series is defined as

$$S(s^{-1} |W_n^{xy}(s)|^2) = S(s^{-1} |W_n^{x}(s)|^2) + S(s^{-1} |W_n^{y}(s)|^2).$$

where $S$ is the smoothing operator. Notice that this definition closely resembles that of a traditional correlation coefficient, and it is useful to think of the wavelet coherency as a localized correlation coefficient in time frequency space. Without smoothing coherency is identically 1 at all scales and times. We may further write the smoothing operator $S$ as a convolution in time and scale:

$$S(W_n) = S_{mouse}(S_{time}(W_n)),$$

Where $S_{mouse}$ denotes smoothing along the wavelet scale axis and $S_{time}$ denotes smoothing in time. The time convolution is done with a Gaussian and the scale convolution is performed with a rectangular window (Torrence and Campo 1998). For the Morlet wavelet a suitable smoothing operator is given by

$$S_{time}(W_n) = |W_n(s)\exp(-c_s^2\frac{|s|^2}{2})|,$$

where $W_n(s)\exp(-c_s^2\exp(0.6\pi))$, $c_1$, $c_2$ are normalization constants and $\Pi$ is the rectangle function. The factor 0.6 is the empirically determined scale de-correlation length for the Morlet wavelet (Torrence and Campo, 1998). In practice both convolutions are done discretely and therefore the normalization coefficients are determined numerically. Since theoretical distributions for wavelet coherency have not been derived yet, to assess the statistical significance of the estimated wavelet coherency, one has to rely on Monte Carlo simulations.

Cross wavelet phase angle

The phase for wavelets shows any lead or lag relationships between components, and it is defined as

$$\phi_{x,y} = \tan^{-1} \frac{\text{Real}(W_n^{xy})}{\text{Imag}(W_n^{xy})}, \phi_{x,y} \in [-\pi, \pi],$$

where $\text{Imag}$ and $\text{Real}$ are the imaginary and real parts, respectively, of the smooth power spectrum. Phase differences are useful to characterize phase relationships between two time series. A phase difference of zero indicates that the time series move together (analogous to positive covariance) at the specified frequency. If $\phi_{x,y} \in \left[0, \frac{\pi}{2}\right]$, then the series move in-phase, with the time series $y$ leading $x$. This is indicated by the phase arrows pointing right and upward. On the other hand, if $\phi_{x,y} \in \left[-\frac{\pi}{2}, 0\right]$, then the series move in-phase with time series $x$ is leading. The phase arrows pointing right and downward in this case. We have an anti-phase relation (analogous to negative covariance for cross wavelet transform and negative correlation for wavelet coherency) if we have a phase difference of $-\pi$, then $\phi_{x,y} \in \left[-\pi, -\frac{\pi}{2}\right]$ or $\left[-\frac{\pi}{2}, -\pi\right]$, meaning leading i.e. the arrows pointing left and upward. The series is leading if $\phi_{x,y} \in \left[-\pi, -\frac{\pi}{2}\right]$ or the directional arrows pointing left and downward.

Empirical results and discussion

Using an application of scientometric methodology to the analysis of scientific communication, relationships between number of submitted articles and monthly observations of scientific papers are examined quantitatively. With regard to the aim of understanding the complexities of these relationships, the time series include monthly observations based on received data as reported on the article history of the Springer and Elsevier paper format. Recently, Magnone (2013) examine the scientometric look at calendar events based on the articles received data from Elsevier, however, the present study is...
The data set contains monthly count of the scientific papers over the 1970-2010 period. The scientific papers received from Springer and Elsevier database is in nominal numbers (taken natural logarithm for stationarity process), while monthly observations taken as a time trend between the successive years. Figure 1 shows the number of scientific papers received from Springer (SPRINGER) and papers received from Elsevier database (ELSEVIER) and their corresponding returns (SPRINGER and ELSEVIER).

Some changes in scientific research publications may be explained by many other time dependent factors that are not captured due to uneven variations of publications in desired time lags; therefore, an estimate with a time trend is also made in this study. For this purpose, we investigate the range of frequencies at which research publications both from Springer and Elsevier database often covary with the time lag, we consider the measure i.e. wavelet coherence. The wavelet coherence is defined as the cross spectrum normalized by the two related auto-spectra. It measures the local correlation between two time series as a function of frequency, with a value between 0 and 1. When the value is 1, the two series are perfectly correlated at a given frequency over a specific period whereas the correlation value equals to zero indicates no correlation at a particular frequency. The results of the wavelet coherence between scientific papers received and time lag indicates the weak to strong correlation between the two series at a given frequency. The cone of influence is indicated by the black convex curve stretching from left margin to the right margin. The results conclude that the two series are often locally correlated at higher frequencies, but their correlations measure in time dimension do not last long. However, during two structural breaks i.e., 1990-1995 and 2007-2011, the two time series seem to be correlated during the sample period.

The analysis of wavelet coherence reveal that the two series are often locally correlated at all different frequencies. Although the direction at medium and higher frequencies is not clear, however majority of the arrows at lower frequencies are pointing right and upward which indicates that Elsevier scientific papers leads to the time dependent factors in the long run. Overall during 2001–2011, the two time series seems to reach the coefficient of coherence of 0.9 for longer period of times with frequency band higher than 512 days. The direction of the arrows is clearly right and upward which again provides an evidence that the two time series are in phase and the Elsevier scientific papers are time dependent in the long run. Thus, the analysis of wavelet coherence reveal that the two series i.e., scientific papers and time dependent factors correlated during the sample period.

Conclusion

Publication and citation analyses have become standard tools for research evaluation. However, some methodological problems remain unresolved. The consequent uncertainties have sometimes been reflected in hesitations to apply these tools as standards in policy making processes and research management decisions (Leydesdorff, 2004). Scientometric indicators i.e., publication productivity with a dimension of number of papers, number of researchers, number of years or citedness of papers etc are the area of ongoing debate (Vinkler, 2000). This study investigates the relationship between scientific papers and time dependent factors, using wavelet analysis. This is the first study that uses wavelet approach for scientometric analysis. Indeed, the wavelet analysis enables us to decompose the data into various time scales and to investigate the relationship on a time scale by time scale basis. Overall, it is found that the relationship between the two variables are complex in nature i.e., the relationship is not constant over frequency bands as it depends on the time scales. The analysis of wavelet coherence differences reveals that the two series are correlated during the sample period. Policy-makers in the world have the eminent goal to improve the life of their people, for example, in the medical sector, scientific research leading to new prevention and treatment strategies has the potential of enormous social value for the population. A widely used instrument to enhance research output is to allocate funds to the institutions that provide scientific progress. Because funding resources are limited in most countries, decision-makers have to select areas and projects that are worthy of support, taking into account, besides other factors, the urgency of the anticipated results (Al-Mutawakel et al. 2010).

References


1. It turns out that both series in levels are integrated of order one, following the results of the augmented-Dickey–Fuller (ADF) and Phillips–Perron (PP) tests not reported here.