INFORMETRICS 89/90
L. Egghe and R. Rousseau (Editors)

CHAOTIC STRUCTURES IN INFORMETRICS

Albert N. TABAH*

Physical Sciences & Engineering Library
McGill University - Montreal, Quebec H3A 2K6 Canada

Aaron Jaan SABER

Department of Mechanical Engineering
Concordia University - Montreal, Quebec H3G 1M8 Canada

Abstract

A sample space of the number of publications in polymer chemistry over a five year period has been subjected to time series data analysis. The technique uses autocorrelation and a subsequent FFT (fast Fourier transform). The analysis shows that the sample space is small but the data appears to have chaotic behaviour. Recommendations for future work and applications are identified.

1. INTRODUCTION

Bibliometric phenomena and principles, such as Bradford's law, are now well understood [1]. However, even though we understand the internal relationships between cited authors and publications, we still don't know the underlying reasons for our empirical laws. In fact, rather than looking at individual subdisciplines, as we so often do with Bradford studies [2], there may be value in examining an entire body of published knowledge to find the underlying mechanisms behind publication characteristics. However, when the entire body is viewed, publications seem to appear randomly. That is, articles, books and other materials in different fields seem to be published at varying rates and in what appears to be a random fashion.

Recent work attributes the appearance of published knowledge either to chance, or Poisson processes or rank order correlations [3]. Nevertheless, some underlying mechanisms in publication characteristics or information diffusion may exist and could be described quantitatively. One may ask whether dynamic mechanisms are at work. Indeed, examinations of nonlinear dynamical systems suggest that seemingly random processes may actually be chaotic phenomena [4] and that some kind of hidden mechanisms may be at work [5]. We have each been asking whether information production and dissemination involves or can be described by dynamic mechanisms, (AT), and whether conservative physical systems can exhibit chaotic behavior (JS). Together, we decided to apply accepted techniques for analyzing time function data to see whether information production could be described as "chaotic".

The study of chaotic behaviour is very new [3], [6] and has been demonstrated to occur in nature [7], [8]. However, to the authors' knowledge, there are heretofore no similar studies in information science.

* Author to whom all correspondence should be sent.

In this paper we demonstrate the behavior of data from a bibliographic database search. The data was correlated and Fourier transformed. We draw some parallels between dynamic chaotic processes and information production and dissemination and point out some areas of information science worth investigating. While some of the material presented below is rather speculative in nature, we nevertheless present some preliminary evidence that there may indeed be some nonlinear dynamic phenomena at work.

We begin with a brief quantitative discussion of chaos.

2. CHAOS

To begin, we define chaos as nonperiodic behavior whose measures center about identifiable conditions. In the literature, these conditions are called "attractors" [9]. As an example, the range of oscillations or the point of rest in pendulum behavior are attractors. Thus, the behavior a system converges toward is a locus of values or a single valued attractor.

A quantitative description of a continuously time-varying system can involve differential equations. For example, in an elementary form a differential equation is

$$\frac{dx}{dt} = f(x) \qquad (1)$$

where t is time and x is a variable that depends on time. Thus, $\frac{dx}{dt}$ represents the state of change of x with respect to time, as described by the function f(x). Often, the solution of the equation appears explicitly or implicitly in the equation itself. In a simple case, that might imply that input and output are not proportional; these equations are called "nonlinear".

Systems that have nonlinear behavior permit slight imbalances to compound over time. Over the long term, results appear irregularly or randomly, and the future is impossible to predict exactly. It has been shown that slightly different initial conditions can cause dramatically altered results [10]. Such sensitive dependence leading to an actual condition satisfies our definition of chaos.

While differential equations describe continuous phenomena, their discrete analogs are difference equations [11]. Instead of describing the continuous changes of a system's state, difference equations model incremental changes. Difference equations are recursive and so feed back initial values to calculate subsequent states. One of the simplest on which a lot of work has been done is the logistic equation that describes the change of an animal population x over time [7] :

$$x_{t+1} = kx_t(1-x_t) \qquad (0 < x_0 < 1) \qquad (2)$$

Applied to a biological population, the value of the constant k models the fate of that population. Thus, if $k < 0$ the population decreases and eventually becomes extinct; if $0 < k < 1$ the population grows, and if $1 < k < 3$ the population reaches equilibrium. However, as k is increased beyond 3 toward 4 there is a dramatic change. At $k = 3.2$ the population starts oscillating between two conditions. As the value of k grows larger, the oscillations double to 4 points, and then to 8, 16, and beyond, until at $k = 3.57$ an infinite number of points is reached. At this level the behavior is indistinguishable from a random process, even though the initial conditions were far from random and x_0 was a fixed value [8].

Plotting x_{t+1} versus x_t in Equation 2 reveals whether there is an attractor and so signals whether the system is chaotic. For example, one of the characteristics of chaotic attractors is that successive points in the plot can diverge exponentially, yet stay within certain confines, namely the limit cycle. Furthermore, there is a class of attractors that delineate the confines of system behavior. These are called strange attractors [7].

The rate of the divergence between points can be quantified with the calculation of the so-called Lyapunov exponent. A positive exponent quantifies the "sensitive dependence on initial conditions" and gives an estimate of how far into the future one can predict the state of the system, whereas a negative Lyapunov exponent gives an estimate of the stability of the attractor [12], [13], i.e. the more negative, the more stable.

3. CRITERIA FOR FINDING CHAOS

Recent work in biological and epidemiological systems has demonstrated that it is possible to take time series data and establish whether it is due to chance, quasiperiodically or chaotic behavior [14], [15].

The following analyses must be followed to demonstrate successfully chaotic behavior in a system [16] :

1. Fourier transform : Plot the time history of the data, optionally calculate the autocorrelation function [17], and then examine its Fourier transform plot. The output must exhibit a broad spectrum of explicit frequencies.

2. Lyapunov exponent : Calculate the Lyapunov exponents for the data. The maximum exponent must be a positive number.

3. Poincaré section : Because systems may be multi-dimensional, and because of the nature of the quickly varying orbits, an intersection is taken through the orbits of attractors to reveal their character. This gives a measure of their "strangeness". The approach involves plotting x_{t+1} versus x_t and taking a slice vertical to its plane. If the so-called Poincaré map repeats itself on any scale of observation (i.e. possesses a fractal nature), chaos is successfully demonstrated.

There is no room here to elaborate on the several ways and limitations for data presentation and analysis. There are, however, two problems that remain with this type of time series data and must be acknowledged :

a. How do we know that we have selected an appropriate set of variables?
b. What is the amount of data required to construct a good model?

Since the time series involved in chaotic systems are not perfectly reproducible, it is vital that these two problems be resolved [18], [19].

4. DATA COLLECTION

To examine the possibility of chaotic behavior in the appearance of published knowledge, we focused on polymer chemistry as a sample space. This is because polymer chemistry is an area that has shown considerable activity over the last twenty years. Furthermore, during a specific period from 1984 to 1988, there seemed to be unusually high activity. On that basis, we collected the number of articles published on polymer chemistry as reflected in the biweakly updates of the Chemical Abstracts data base on STN.

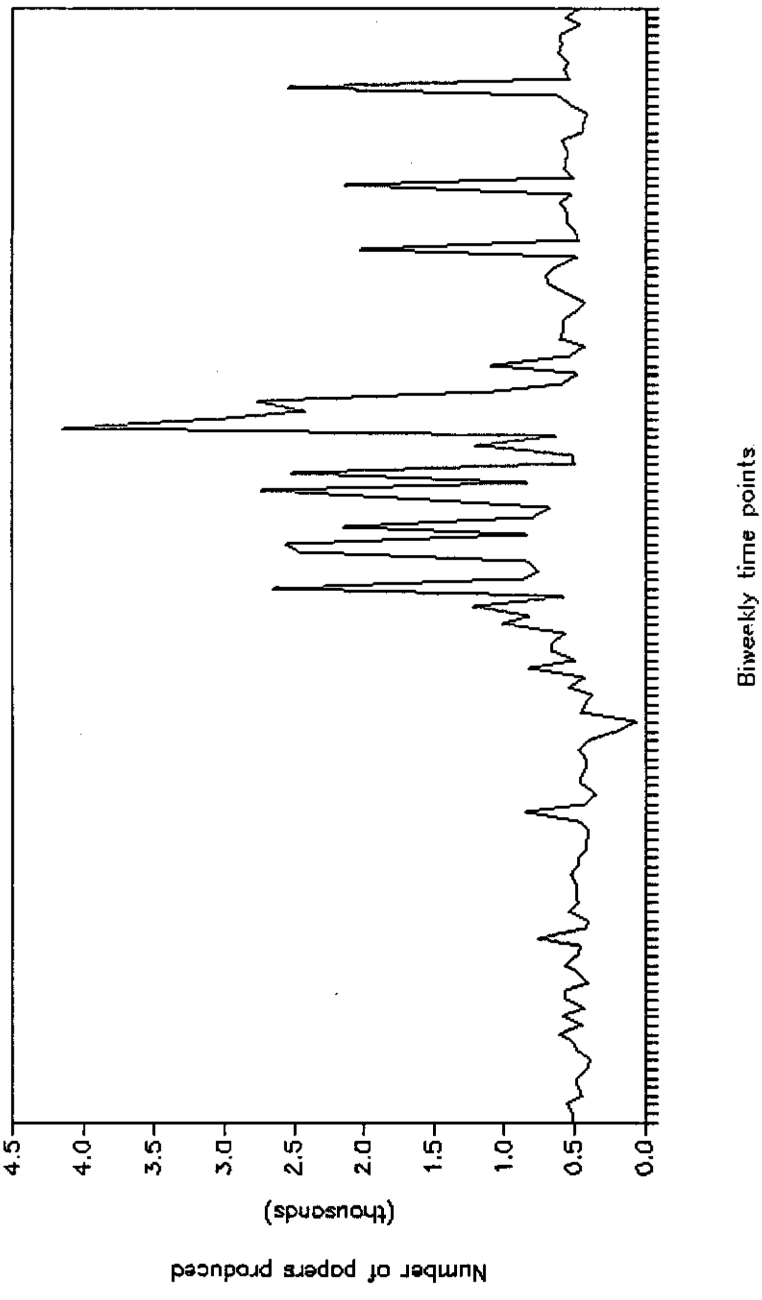

Fig.1 : Number of polymer chemistry articles indexed by Chemical Abstracts between May 1984 and January 1989

The data are illustrated in Figure 1. An average of 782 papers appeared each biweekly period between 12 May 1984 and 6 January 1989. The 12 May 1984 date was selected to coincide with the start of online limiting capabilities for STN biweekly updates. The minimum production during the period was 68 papers and the maximum peaked at 4147 publications. The data show that there are several weeks with unusually high numbers of reported publications that fall below the maximum production.

5. DATA ANALYSIS

The data were analyzed using signal processing techniques. The approach first involved preparation of the autocorrelation of the raw data. This was followed by construction of the fast Fourier transform (FFT) of the result [20]. This approach limits the spurious or random "noise" in the data.

To implement the autocorrelation and the FFT, the data were extended to 256 intervals by assuming null publication data for 6 intervals prior and 124 intervals post the period of interest.

6. RESULTS

The autocorrelation of the data is shown in Figure 2 where the ordinate is the correlation amplitude derived from the scale in Figure 1. Only half of the autocorrelation is shown, since the other half is a reflection. The lack of the noise spike in the autocorrelation reveals that the data is inherently well-organized with few random fluctuations.

The Fourier transform results are illustrated in Figure 3. Like the autocorrelation, only half of the Fourier transform needs to be shown. The first, lowest frequency spike corresponds to the sample period (from 12 May 1984 to 6 January 1989). The data generally decay as frequency increases to that associated with the biweekly appearance of the data.

There are several clear peaks (p_1,p_2,p_3) that seem to be harmonics of one another. This suggests that certain papers appear at regular intervals and may be associated with other papers that appear two and three times as quickly. However, there is also a cluster of three peaks that do not appear to be associated with the harmonic triad. The three peaks, located at position c do not appear to be associated with other information in the spectrum.

The high frequency data are related to the sharp cut-off to zero at the start and end of the data set.

7. CONCLUSIONS

The Fourier transform reveals "a broad spectrum of explicit frequencies", thereby satisfying the first criterion for identifying chaotic behaviour. However, the small sample space limits the confidence level in stating that the behavior is indeed chaotic. Nevertheless, the use of time series data of the type we have considered seems appropriate to chaotic behavior analysis.

8. FURTHER RESEARCH AND APPLICATIONS

Future work will involve both larger sample spaces in specific subject areas, as well as a variety of subjects. Attention will also be paid to the calculation of Lyapunov exponents in order to identify their signs and values.

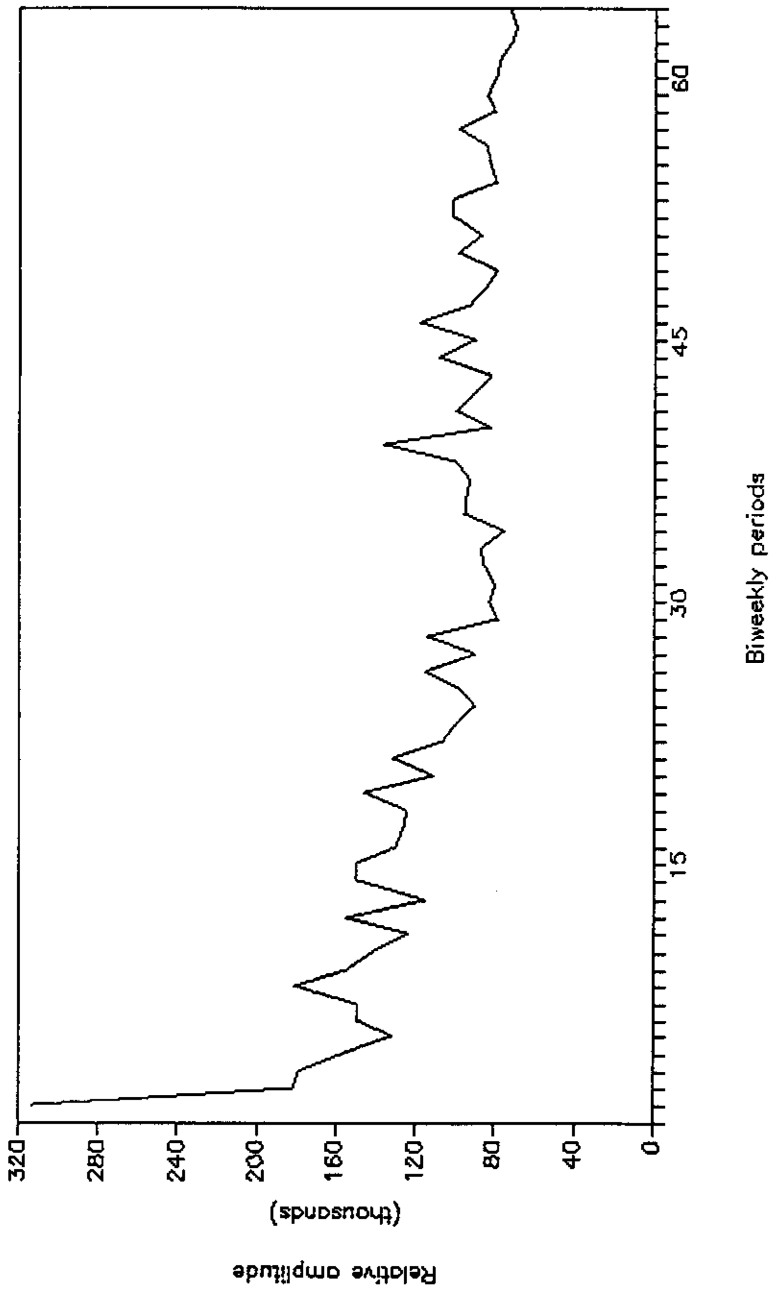

Fig.2 : Autocorrelation function of the polymer chemistry articles indicates relatively non-random and noise-free data. Only half of the spectrum is shown, since the other half would only be a mirror image.

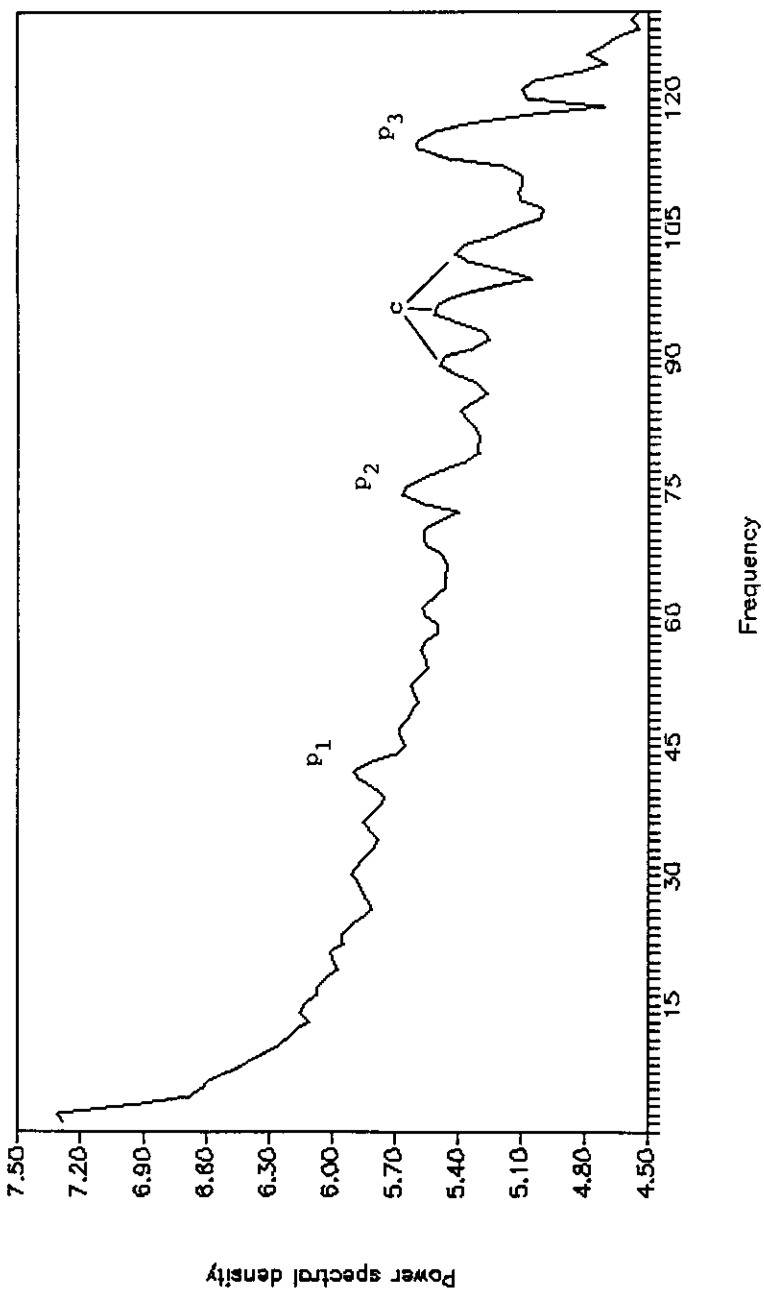

Fig.3 : Fast Fourier transform of the data gives results along expected lines, although some spikes (p1-3 and c) may have a basis in the regularity with which some of the articles or journals use indexes and thus may be harmonics of one another.

In this context, we have identified three interesting areas of study :

1. Epidemic theory :
In a series of remarkable papers in the 1960's Goffman put together a theory of information diffusion based on a mathematical model of infection in a population [21]. Recent work suggests that there are nonlinear dynamic phenomena prevalent in models of recurrent epidemics [12], [22] and that the parallel can once again be extended to information science.

2. The Lotka hypothesis :
Lotka demonstrated that the distribution of author productivity in a given field is inversely proportional to the number of authors writing in that field [23]. While the power in the constant of his relation seems to differ depending on authors and their fields [24], Radhakrishnan and Kernizan argued that, in computer science at least, it may be 3 [25]. A comparison between various literatures may indeed confirm the variability in the power to x, and that power may turn out to be a fractal number (i.e. non-integer). Literatures with common fractal dimensions may exhibit common characteristics under a new classification scheme. Mandelbrot's findings that a graph of rankings within a linguistic structure exhibit fractal slopes may also find applications here [26].

3. Co-citation analysis :
The study of fractal dimensions can be extended to the study of co-citation maps. Maps with common fractal dimensions may be grouped together and may turn out to have comparable characteristics of growth and diffusion. We may also examine whether changes in the points on the maps are due to random walk or whether they are deterministic *and* non-linear.

REFERENCES

[1] Egghe, L., Rousseau, R., Informetrics 87/88. Select Proceedings of the first International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval. Amsterdam, Elsevier Science Publishers, 1988.

[2] Narin, F., Evaluative bibliometrics : the use of publication and citation analysis in the evaluation of scientific activity. Washington, D.C., National Science Foundation, 1976.

[3] Brookes, B.C., Ranking techniques and the empirical log law. Information Processing & Management, 20(1-2) : 37-46, 1984.

[4] Gleick, J., Chaos : making a new science. New York : Viking Press, 1987.

[5] Jensen, R.V., Classical chaos. American Scientist, 75(2) : 168-181, 1987.

[6] Bai-Lin, H., Chaos. Singapore : World Scientific Publishing Company, 1984.

[7] Degn, H. et al., editors. Chaos in biological systems. NATO Advanced Research Workshop on Chaos in Biological Systems. New York, Plenum Press, 1987.

[8] Glass, L., Mackey, M.C., From clocks to chaos : the rhythms of life. Princeton, NJ : Princeton University Press, 1988.

[9] Ruelle, D., Strange attractors. Mathematical Intelligencer, 2 : 126-137, 1980.

[10] May R.M., Simple mathematical models with very complicated dynamics. Nature, 261 : 459-467, 1976.

[11] Derrick, W.R., Grossman, S.I., Elementary differential equations with applications. Reading, MA : Addison-Wesley, 1976.

[12] Shaw, R., Strange attractors, chaotic behaviour, and information flow. Zeitschrift für Naturforschung, 36a : 80-112, 1981.

[13] Wolf, A., Swift, J., Progress in computing Lyapunov exponents from experimental data. In : Horton, C.W., Reichl, L.E., editors. Statistical physics and chaos in fusion plasmas. New York : Wiley, 1984, 111-126.

[14] Schaffer, W.M., Stretching and folding in lynx for returns : evidence for a strange attractor in nature. American Naturalist, 124(6) : 798-820, 1984.

[15] Schaffer, W.M., Chaos in ecology and epidemiology. In : Degn, H. et al., editors. Chaos in biological systems. NATO Advanced Research Workshop on Chaos in Biological Systems. New York : Plenum Press, 1987, 233-248.

[16] Moon, F.C., Chaotic vibrations. New York, Wiley, 1987.

[17] Beauchamp, K.G., Signal processing : using analog and digital techniques. London : George Allen & Unwin, 1973, 411.

[18] Shaw, R., The dripping faucet as a model chaotic system. Santa Cruz, CA : Aerial Press, 1984.

[19] Albano, A.M. et al., Data requirements for reliable estimation of correlation dimensions. In : Degn, H. et al., editors. Chaos in biological systems. NATO Advanced Research Workshop on Chaos in Biological Systems. New York : Plenum Press, 1987, 207-220.

[20] Brigham, E.O., The fast Fourier transform. Englewood Cliffs, NJ : Prentice-Hall, 1974.

[21] Goffman, W., Newill, V.A., Communication and epidemic processes. Proceedings of the Royal Society, 298A : 316-334, 1967.

[22] Aron, J.L., Schwartz, I.B., Seasonality and period doubling bifurcations in an epidemic model. Journal of Theoretical Biology, 110 : 665-679, 1984.

[23] Lotka, A.J., The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences, 16(12) : 317-323, 1926.

[24] Nicholls, P.T., Price's square root law : empirical validity and relation to Lotka's law. Information Processing & Management, 24(4) : 469-477, 1988.

[25] Radhakrishnan, T., Kernizan, R., Lotka's law and computer science literature. Journal of the American Society for Information Science, 30(1) : 51-54, 1979.

[26] Mandelbrot, B., An informational theory of the statistical structure of language. In : Jackson, W., editor. Symposium on Applications of Communication Theory. London : Butterworths, 1953, 486-502.