Pergamon

Information Processing & Management, Vol. 33, No. 3, pp. 393–403, 1997

0306-4573/97 $17+0.00

PII: S0306-4573(96)00071-4

SPECTRAL METHODS FOR DETECTING PERIODICITY IN LIBRARY CIRCULATION DATA: A CASE STUDY

FRANCIS DECROOS, KRIS DIERCKENS, VINCENT POLLET, RONALD ROUSSEAU*, HUGO TASSIGNON and KOEN VERWEYEN

KHBO, Department of Industrial Sciences and Technology, Zeedijk 101, 8400, Oostende, Belgium

(Received 28 May 1996; accepted 18 September 1996)

Abstract—The purpose of this investigation is to show the feasibility of spectral methods in the field of information science, in particular for the analysis of library circulation data. Using the software package MATLAB® we applied the discrete Fourier transform to obtain frequency information about the noisy time series of circulation data. Over a time span of two academic years we could clearly detect a semestral and—less visibly—a weekly periodicity. Other periods were not so distinct and could be spurious. The normal loan period of four weeks could not be detected. Following McGrath (1996, *Journal of the American Society for Information Science, 47*, 136–145), and Naylor and Walsh (1994, *Library and Information Science Research, 16*, 299–314) we conclude that spectral methods show a lot of promise for analyzing all kinds of time series and other signals occurring in the field of information science. This approach certainly deserves more attention from practitioners. © 1997 Elsevier Science Ltd

1. INTRODUCTION

Spectral methods are well known in engineering sciences. They belong to the core curriculum of any engineering course. Indeed, in the introduction of their classical book on signals and systems Oppenheim *et al.* (1983) state that:

> The concepts of signals arise in an extremely wide variety of fields, and the ideas and techniques associated with these concepts play an important role in such diverse areas of science and technology as communications, aeronautics and astronautics, circuit design, acoustics, seismology, biomedical engineering, energy generation and distribution systems, chemical process control, and speech processing.

Moreover, among the first signals presented in their book they include the discrete-time signal of weekly Dow-Jones stock market indices and the species-abundance relation of an ecological community. All these fields are close enough to the information sciences to wonder why the analysis of signals and systems has not caught on much earlier. Moreover, one of the authors of this article (R.R.) has already on several occasions stressed the relations that exist between economics and ecology on the one hand, and the information sciences on the other (Rousseau, 1992, 1994; Rousseau & Van Hecke, 1993).

As far as we know, McGrath's (1996) paper, presented at the 4th International Conference on Bibliometrics, Informetrics and Scientometrics, Berlin, Germany in 1993, was the first to use spectral methods in the frequency domain to study library circulation data. Curiously enough, the few studies using spectral methods that preceded McGrath's work include several articles on the much more difficult topic of chaotic behavior (Kurtze *et al.*, 1992; Snyder & Kurtze, 1995; Tabah & Saber, 1990). Other scientists that use signal processing methods are Kunz (1987) in relation with time studies of patent data, and Hall (1990, 1992), who studied growth models. Recently, Naylor & Walsh (1994) have used spectral analysis as a step in deriving a best-fit time series model. As their aim was to obtain a complete model for internal library pick-up data,

* To whom all correspondence should be addressed.

detecting periodicity was only a part of their investigations. Consequently, their approach is more sophisticated than ours.

2. DATA—PURPOSE OF THE INVESTIGATION

2.1. Data

The original time series consists of loan data of the KIHWV/HTI library over two consecutive academic years. Data were collected on a daily basis by the library staff. Three series of data will be analyzed: the total number of loans per day, the number of book loans per day, and the number of thesis loans per day. The first series is the sum of the second and the third. As the library is closed during weekends and during the holiday periods, a week consists of five days, and a year is an academic year of 35 weeks (or 175 days), divided into two semesters. The first semester consists of 14 weeks of teaching, followed by a study and examination period of three weeks; the second one consists also of 14 weeks of teaching followed by a study and examination period of four weeks. Note that all holiday periods, i.e. the Christmas and Easter holidays are simply eliminated as the library is closed during these days. A description of the KIHWV/HTI library and a discussion of socio-cultural factors affecting the borrowing behavior of students can be found in Rousseau & Vandegehuchte (1995). The time series of all loans is shown in Fig. 1. Because of the decrease in loans during the study and examination periods the four semesters can easily be discerned. The complete data set is given in Appendix A, so that others can try to squeeze more information (than we did) out of it.

2.2. Purpose of the study

The aim of this investigation is to show the feasibility of spectral methods in the field of information sciences, in particular for the analysis of library circulation data. As such, we intend

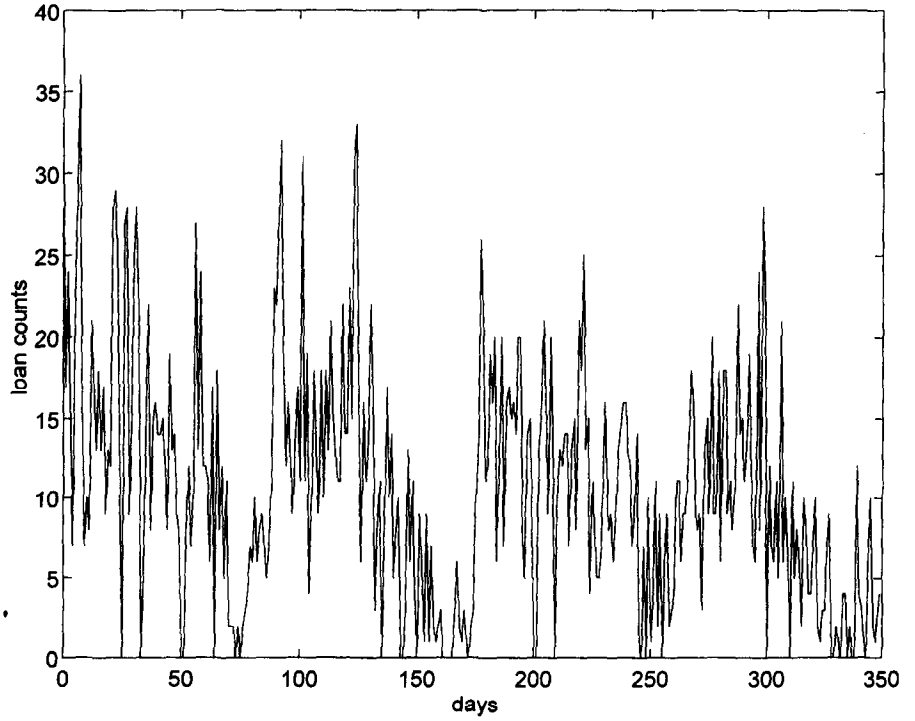

Fig. 1. Two-year loan data: time series.

to corroborate McGrath's approach (1996). In addition, we apply some transformations, e.g. detrending of the data, which were not performed by McGrath.

3. METHODS

3.1. Basics of signal processing

We will use elementary signal analysis methods to detect periodicity in the time series. In particular, we will transform the given signal (in time domain) into a series in frequency domain. The tool to perform this transformation is the discrete Fourier transform (DFT). This is a bijective linear transformation from the space of real or complex N-tuples (an N-dimensional vector space) onto itself. Here the first space is considered the time domain, consisting of all 350-tuples (we have 350 data points); the second space is the frequency domain, which in this concrete situation is also a 350 dimensional space. The exact form of the DFT is shown in Appendix B. Note that in time domain a difference of one unit denotes a difference of one day; in frequency space a value of k stands for a pure sinusoid signal with period $350/k$ days. The following example explains these notions.

Let $x=(x(n))_{n=0,\ldots,29}$ where

$$x(n)=1.0\cos(2\pi f_1 n)+2.0\cos(2\pi f_2 n)+3.0\cos(2\pi f_3 n) \tag{1}$$

with $f_1=1/5, f_2=1/3, f_3=1/10$. This signal is shown in Fig. 2. Clearly, knowing only the sequence $x(n)$, it is nearly impossible to decompose $x(n)$ into its three terms.

Applying the DFT (when N is an arbitrary integer, and a more efficient numerical realization, known as the FFT, i.e. the Fast Fourier Transform, when N is a power of two) yields Fig. 3. Now we have all the information we need to decompose the original signal x. The frequencies f_i, $i=1$, 2, 3 can be found by dividing the numbers associated with the peaks (here: 3, 6 and 10) by N (here 30). The amplitudes (here 1.0, 2.0 and 3.0) are obtained by dividing the values for $X(k)$ by

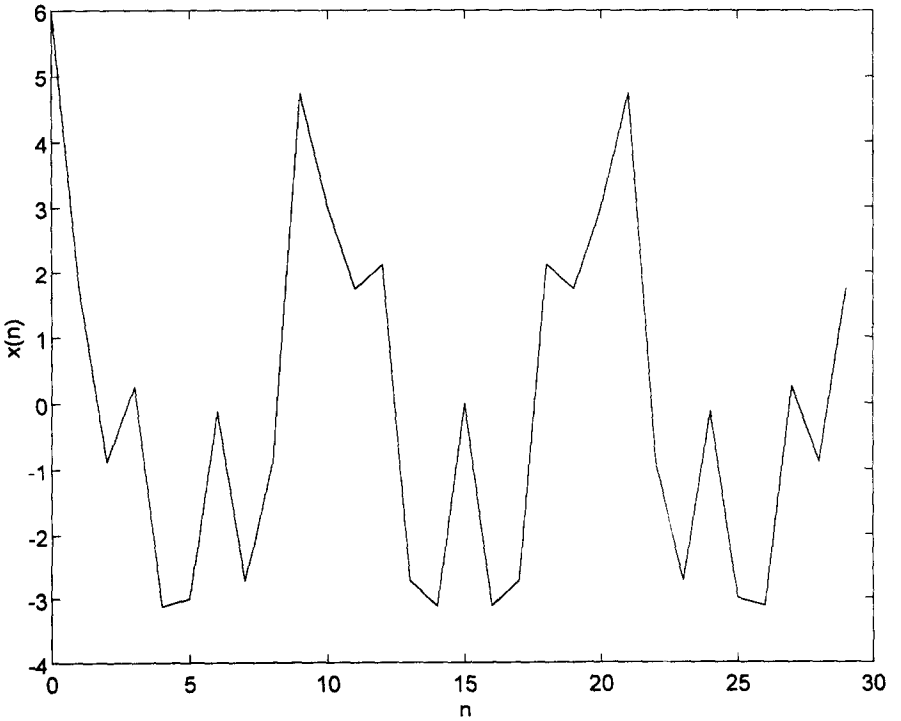

Fig. 2. Time–domain representation of equation (1).

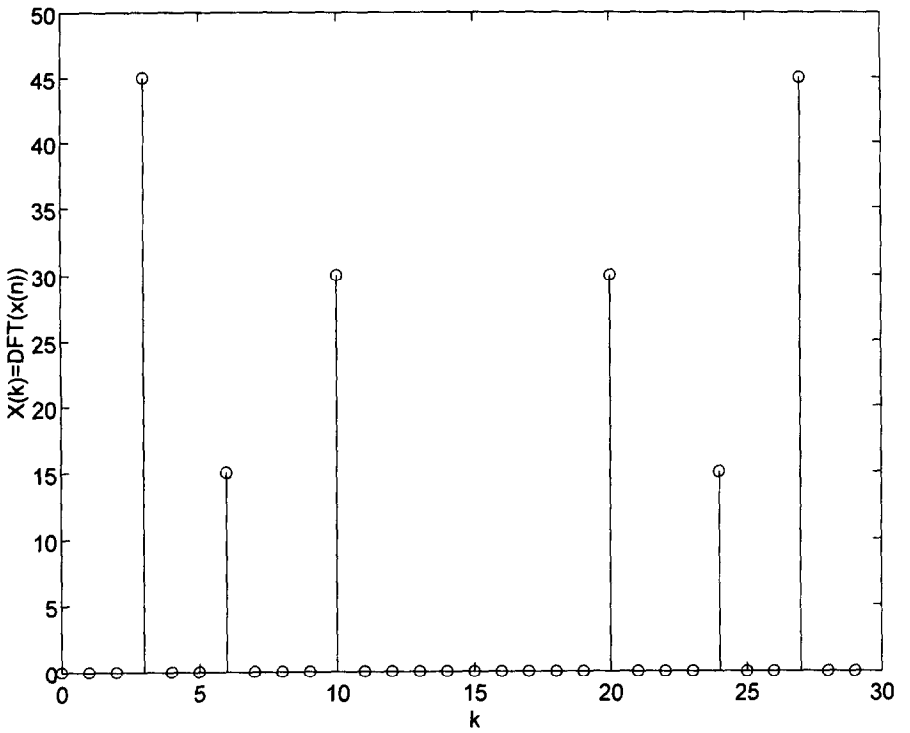

Fig. 3. Frequency–domain representation of equation (1).

$N/2$ (here 15). We further note that if the average of the signal is zero, the first component of the DFT (indexed by 0) is zero too.

We will further apply autocorrelation: this is a technique that suppresses random variations, and therefore enhances the real signal hidden in the noise. The mathematical formula used to calculate the autocorrelation function is given in Appendix B.

The power spectrum, denoted as PS, is the DFT of the autocorrelation function (Shiavi, 1991). It is always real and non-negative. Basically this provides us with the same information as the DFT of the original signal, but the advantage lies in the fact that now the unwanted influence of randomness (noise) is suppressed. Note, however, that to obtain the amplitudes of the original signal from the PS we first have to take the square root of the amplitude of the PS value and then divide the result by N/2.

How can we detect periodicity in a signal? It can be shown (see e.g. Jaffe, 1987b; Oppenheim *et al.*, 1983) that:

1. If the original signal is periodic, the autocorrelation function is also periodic with the same period;
2. If a signal is periodic then its DFT consists of a repetitive pattern of a non-zero component followed by a fixed number of zeros. The non-zero components are not necessarily of equal magnitude. This is illustrated in Fig. 4.

3.2. From theory to practice

Data manipulation was done by means of the student version of MATLAB®. Some functions (such as "detrend") were already available, other ones were not (such as the circular autocorrelation function) and were created by us as so-called m-files.

The first step we took was "detrending" the signal. This means that a best-fitting linear function was calculated and then the values corresponding to this linear function were subtracted from the original data. This resulted in a signal with zero average. The "detrending" operation is necessary to avoid scattering energy round zero. It is recommended in most books on signal

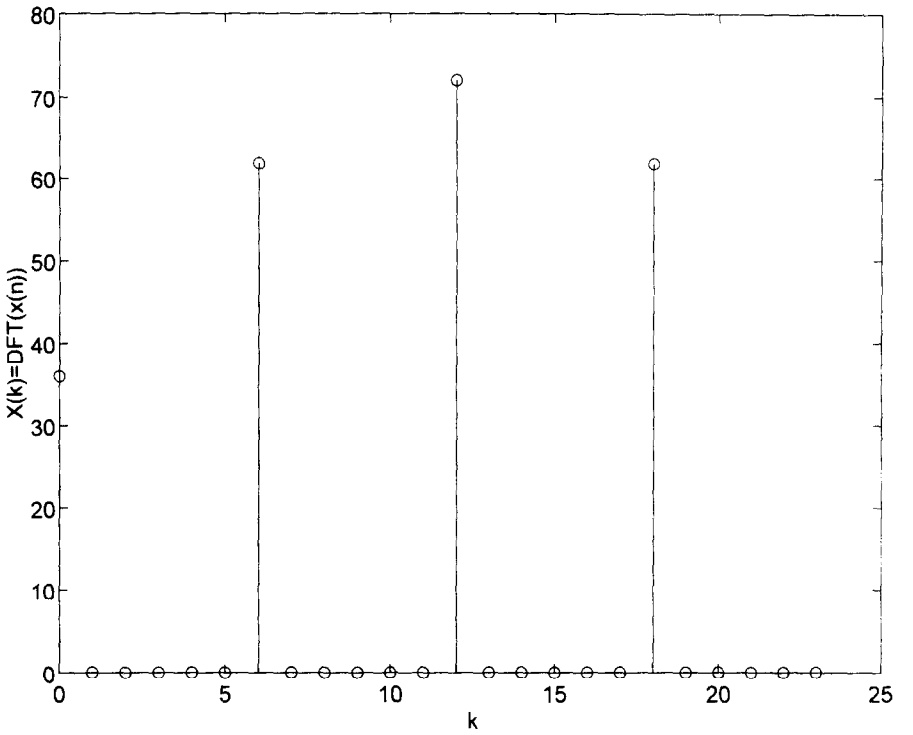

Fig. 4. DFT of the periodic pattern $x(n)=[7, -6, 2, 3]$ repeated six times.

processing, e.g. Shiavi (1991), but as far as we could see this was not applied by McGrath (1996).

Next, we applied the autocorrelation function, resulting in a roughly sinusoidal signal (see Fig. 5). Finally, we applied the DFT to this signal, giving us the power spectrum of the original signal.

4. RESULTS

The power spectra for all loans, book loans and thesis loans are shown in Figs 6, 7 and 8. We assumed that the input sequence $x(n)$ (the loan data) is a sum of cosines as in (2):

$$x(n)=\sum_{j=1}^{M} A_j \cos(2\pi f_j n) \tag{2}$$

The number M is unknown (in theory it could even be $+\infty$) but this does not matter, as we are interested only in the main contributing terms. Our first aim is to find periodic patterns in the data. Hence, we tried to identify the peaks in the frequency spectra. In the spectrum of all loans we observed a huge peak at $k=4$, and a smaller one at $k=12$; the peak at $k=46$ is clearly visible as well. Book and thesis loans separately yield similar results. For a period of one week ($=5$ days) we explicitly searched for a peak at $350/5=70$, where we detected a peak, although a fairly weak one. Table 1 "translates" frequency units to frequencies (in μHz) and periods (in library days), and gives a subjective indication of the relative magnitude of the effect.

A period of 87.5 days corresponds precisely to one average semester (17 weeks=85 days, 18 weeks=90 days). This result is not surprising and was already clearly visible in the original data. It constitutes, however, a corroboration of the spectral method. Although weak, we did find the expected five-day period. We have no explanation for the other two peaks: a period of 29 library days corresponds to 41 calendar days or about six weeks. It might be that this peak is just an harmonic of the "semester" peak. A partial explanation lies perhaps in the fact that students' lab reports are due approximately every three weeks. A period of seven library days corresponds

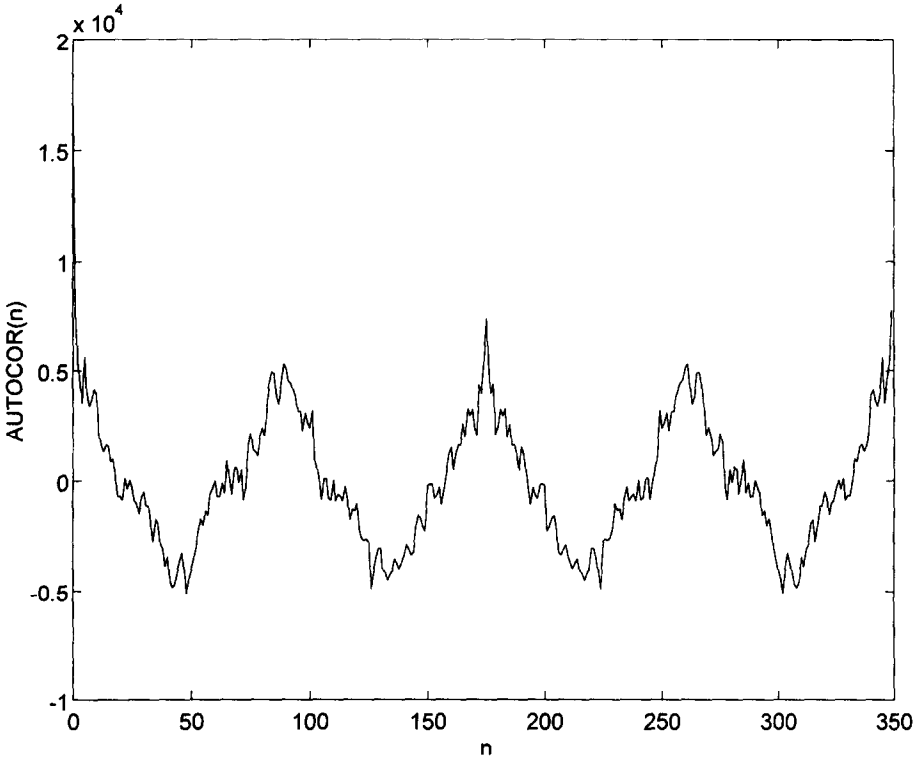

Fig. 5. Autocorrelation function applied to all loan data.

to one and a half weeks. Perhaps it could be that students tend to borrow a book roughly every 10 calendar days. Anecdotal evidence seems to suggest that some students borrow a book,

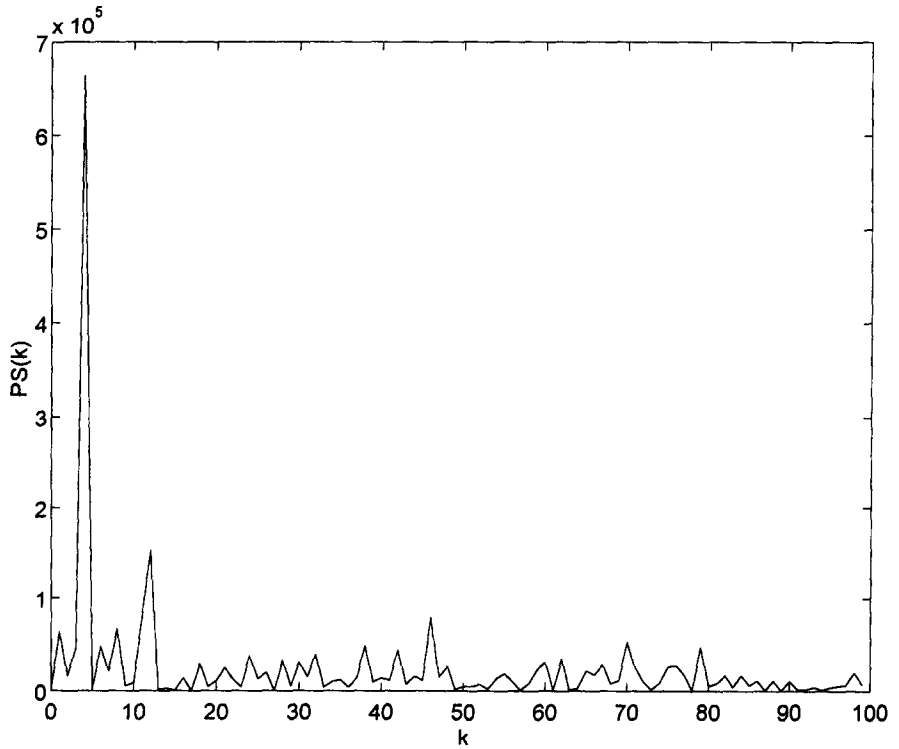

Fig. 6. Power spectrum of all loan data (truncated at $k=100$).

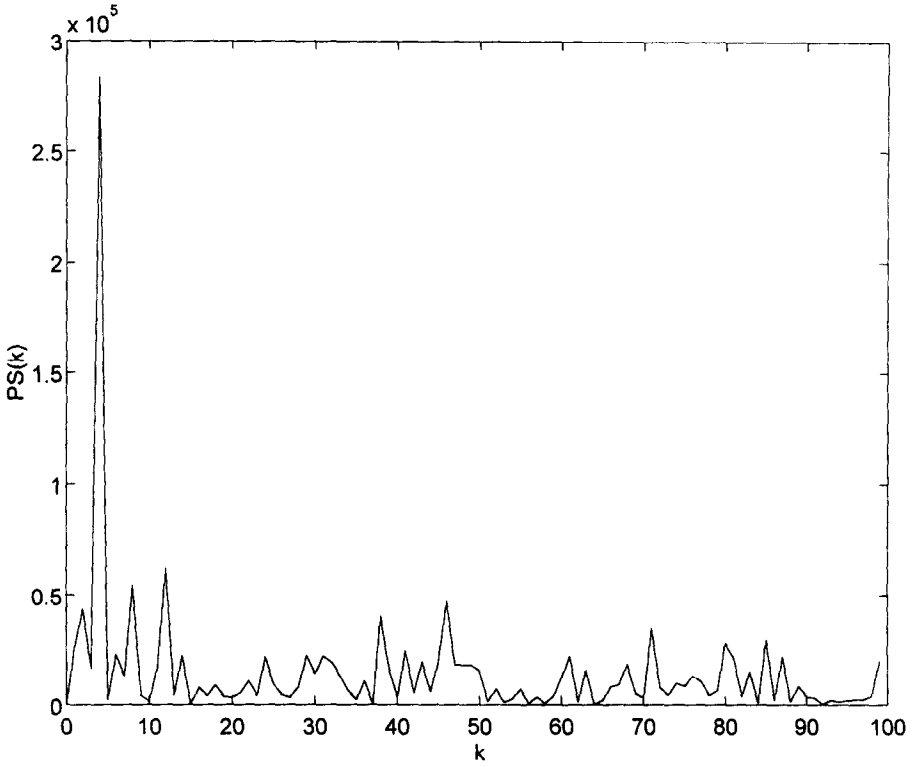

Fig. 7. Power spectrum of book loans (truncated at $k=100$).

glance through it, and either find what they were looking for or see that this book is not really what they want. After somewhat more than a week they return it to try another one. We realize

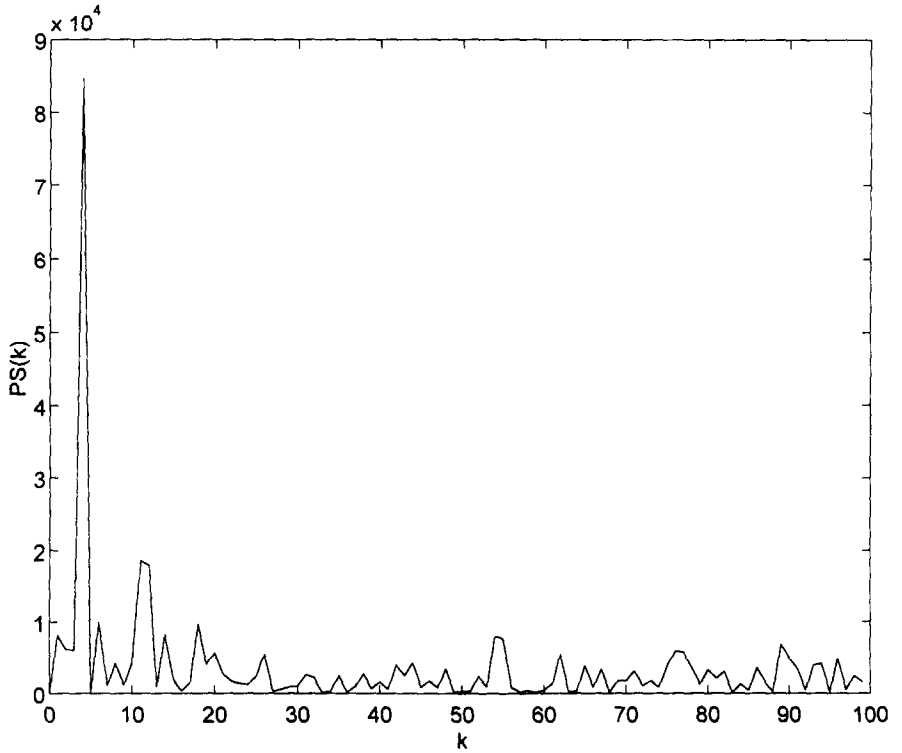

Fig. 8. Power spectrum of thesis loans (truncated at $k=100$).

Table 1. frequencies and approximate periods

Numbers	Frequency	Period	All loans	Books	Theses
4	0.13	87.5	strong	strong	strong
12	0.40	29	strong	strong	strong
46	1.54	7.5	moderate	strong	weak
70	2.31	55	weak	weak	none

this is a poor explanation, but after consulting the main librarian we could not find a better one. We further mention that the "maximum" loan period is four weeks. Yet, we could not detect a four-week period in the data. Perhaps, this could be explained by the fact that this "maximum" period only applies to students (not to staff) and theses are moreover often borrowed for one semester or even for a whole academic year (which is tolerated so long as no-one else requests this particular thesis). We note that there are several other "moderate" and "weak" peaks for which, however, we have no explanation.

The amplitudes associated with the main peaks are approximately 4.7 (all loans), 3 (books) and 1.7 (theses). Further, returning to Fig. 5 which shows the autocorrelation function of the input sequence $x(n)$, we observe that there is already a period of one semester (87.5 days). The amplitude of this sinusoidal wave is $2(5000)^{1/2}/350 \approx 5.3$ (cf. Appendix B for the method of calculation). The difference between 5.3 and 4.7 (theoretically they should be equal) can be explained by the fact that after applying the DFT to the autocorrelation function some energy is scattered in side lobes. We conclude that library loans are characterized by a long term wave with amplitude five (five loans up and five down).

We finally note that we tried several refinements, such as applying a window to the autocorrelation results (before computing the DFT) or adding corresponding days for the two years and then doing the analysis. However, this yielded at best a slight improvement. Therefore, we prefer to present the results of the more elementary approach. As an experiment, we extended our five-day week to a seven-day week by adding zeros. This resulted in a huge peak at the corresponding frequency point. However, this is not what we wanted to find. It is obvious that closing the library during weekends creates a regular pattern. What we were investigating was whether students' loan behavior also showed a weekly pattern.

5. CONCLUSIONS

The primary purpose of this investigation was to corroborate McGrath's approach concerning the feasibility of spectral methods for the analysis of circulation data. In particular we tried to find a periodicity in the loan data of books, theses and their union. The strongest evidence was found for a semestral periodicity. In the KIHWV/HTI library loans seem to be characterized by a long term wave of amplitude five. Further, a weekly period was detected as well. Other peaks in the frequency spectrum were observed but for these no clear explanation could be given.

Contrary to the approach taken by McGrath, we first applied a detrending of the signal (a procedure we strongly recommend), and applied autocorrelation to obtain a better separation between signal and noise. Nevertheless, it seems that our results are less clear than McGrath's. The reason for this is undoubtedly the different behavior of students and the different organization of the library. Yet, as shown by our experiment, it could be that the inclusion, by McGrath, of weekend data when there is less activity in the library, could explain the fact that his weekly period is more pronounced than ours. We are not sure whether it makes sense to try to detect chaos, in the scientific sense, in such a small signal (a procedure suggested by McGrath (1996) and elaborated in McGrath (1995), but we are convinced that the borrowing behavior of our students is chaotic in the everyday sense.

Finally, a more obvious next step would be to model the complete data set, following the Naylor–Walsh approach.

Acknowledgements—We thank P. Vandegehuchte, librarian of the KHBO library for the use of the data collected by him and his staff. We also thank Bill McGrath for his comments and advice. This research was done within the framework of the MSc course in Electronic System Design of the Leeds Metropolitan University

REFERENCES

Firth, J. M. (1992). *Discrete transforms*. London: Chapman & Hall.

Hall, D. H. (1990). Growth and citation selection rates in rapidly growing sciences from date stacking and bibliographic databases. In L. Egghe and R. Rousseau, (Eds) *Informetrics 89/90*, (pp. 163–192). Amsterdam: Elsevier.

Hall, D. H. (1992). The science–industry interface: correlation of time series of indicators and their spectra, and growth models in the nuclear fuels industry. *Scientometrics*, *24*, 237–280.

Jaffe, D. A. (1987a). Spectrum analysis tutorial. Part 1: the discrete Fourier transform. *Computer Music Journal*, *11*(2), 9–24.

Jaffe, D. A. (1987b). Spectrum analysis tutorial. Part 2: properties of the discrete Fourier transform. *Computer Music Journal*, *11*(3), 17–35.

Kunz, M. (1987). Time spectra of patent information. *Scientometrics*, *11*, 163–173.

Kurtze, D., Snyder, H., & Newby, G. B. (1992). An investigation of chaos in computer mediated communication. In I. K. Ravichandra Rao (Ed.) *Informetrics-91: Selected Papers from the Third International Conference on Informetrics*, (pp. 332–342). Bangalore: Sarada Ranganathan Endowment for Library Science.

McGrath, W. E. (1995). Dynamics of chaos in library circulation: preliminary analysis. In M. Koenig and A. Bookstein (Eds), *Proceedings of the Fifth Conference of the International Society for Scientometrics and Informetrics*, (pp. 283–292). Medford, NJ: Learned Information.

McGrath, W. E. (1996). Periodicity in academic library circulation: a spectral analysis. *Journal of the American Society for Information Science*, *47*, 136–145.

The Math Works Inc. (1995). *The student edition of MATLAB®, with tutorial by Duane Hanselman and Bruce Littlefield*. Englewood Cliffs, NJ: Prentice-Hall.

Naylor, M., & Walsh, K. (1994). A time-series model for academic library data using intervention analysis. *Library and Information Science Research*, *16*, 299–314.

Oppenheim, A. V., Willsky, A. S., & Young, I. T. (1983). *Signals and systems*. London: Prentice-Hall Int.

Rousseau, R. (1992). Concentration and diversity of availability and use in information systems: a positive reinforcement model. *Journal of The American Society for Information Science*, *43*(5), 391–395.

Rousseau, R. (1994). Similarities between informetrics and econometrics. *Scientometrics*, *30*(2-3), 385–387.

Rousseau, R., & Van Hecke, P. (1993). Introduction of a species does not necessarily increase diversity. *Coenoses*, *8*, 39–40.

Rousseau, R., & Vandegehuchte, P. (1995). Books and their users: socio-cultural and linguistic aspects in an engineering school. *Library Science with a Slant to Documentation and Information Studies*, *32*, 143–150.

Shiavi, R. (1991). *Introduction to applied statistical signal analysis*. Boston, MA: Irwin & Aksen.

Snyder, H., & Kurtze, D. (1995). How strongly deterministic is chaotic behavior in computer mediated network communication. *JISSI: The International Journal for Scientometrics and Informetrics*, *1*, 241–245.

Tabah, A. N., & Saber, A. J. (1990). Chaotic structures in informetrics. In L. Egghe and R. Rousseau (Eds), *Informetrics 89/90*, (pp. 281–289). Amsterdam: Elsevier.

APPENDIX A

Loan Data

Loan data: 1992–1994 (all loans), per week, read from left to right

Academic year: 1992–1993

32	17	24	15	7	22	28	36	16	7	10	8	21	17	13
18	13	17	9	13	12	28	29	25	11	0	27	28	9	15
25	28	21	0	5	15	22	8	15	16	14	14	15	12	8
19	13	14	9	8	0	0	8	12	7	10	27	13	24	12
12	11	6	17	0	18	8	12	5	11	2	2	2	0	2
0	2	3	4	7	6	10	6	8	9	7	5	7	11	23
22	27	32	19	12	16	12	9	14	17	11	31	11	19	4
10	18	12	9	18	10	18	13	21	15	13	11	11	22	14
14	23	15	31	33	15	6	16	11	13	22	17	3	8	11
0	8	17	10	14	5	8	10	0	0	3	13	6	11	4
0	9	6	1	9	1	7	2	1	2	3	0	0	0	0
0	3	6	2	1	3	1	0	2	3					

Academic year: 1993–1994

10	13	26	21	11	12	19	16	20	6	13	20	7	16	17
15	16	14	20	20	8	5	14	15	10	0	0	13	16	21
15	9	20	13	0	11	13	12	14	14	7	13	15	8	21
18	25	13	15	4	11	8	5	5	7	16	12	8	9	6
9	13	15	16	16	13	12	7	9	14	1	0	7	0	10
1	4	11	2	9	0	6	9	2	3	4	11	11	6	9
9	13	18	16	10	8	9	3	13	15	9	20	9	9	18
6	18	18	9	11	8	10	22	14	15	11	13	19	12	7
6	24	8	28	22	0	12	7	6	11	5	21	6	10	7
0	11	5	8	6	2	10	8	4	4	7	10	2	1	3
3	7	9	0	0	2	1	0	4	4	0	2	0	0	12
4	3	1	0	7	10	2	1	3	4					

Loan data: 1992–1994 (books), per week, read from left to right
Academic year: 1992–1993

16	12	18	4	4	16	24	24	15	4	8	6	14	12	12
16	9	12	6	6	12	18	18	21	7	0	10	18	5	11
17	22	19	0	5	9	20	8	11	8	13	12	15	9	7
16	10	13	9	8	0	0	7	11	6	7	24	9	15	9
10	11	5	17	0	16	8	12	5	11	2	2	1	0	2
0	2	2	4	7	6	10	6	6	7	7	3	6	11	19
14	22	27	17	12	12	5	7	11	13	11	29	9	16	1
9	15	8	7	15	8	15	11	18	14	12	6	7	17	11
8	14	10	19	20	9	4	8	7	13	9	16	3	8	10
0	7	15	8	12	5	7	8	0	0	3	9	6	4	4
0	7	6	0	9	1	7	2	1	2	3	0	0	0	0
0	3	6	2	1	3	1	0	2	3					

Academic year: 1993–1994

10	10	23	15	6	2	15	10	15	3	10	19	6	4	13
15	14	11	13	15	7	4	6	11	8	0	0	11	15	13
8	9	15	10	0	9	12	7	11	13	4	6	14	3	15
12	21	9	13	2	11	3	5	5	7	12	12	5	9	6
9	12	12	15	13	13	12	5	7	13	0	0	7	0	10
1	4	10	2	9	0	6	8	2	3	3	10	8	6	7
8	13	13	15	10	6	8	3	8	12	6	14	7	9	12
5	18	18	4	8	8	9	18	13	13	9	9	16	10	4
3	19	8	15	13	0	7	5	4	9	4	10	5	10	4
0	9	2	5	6	2	9	6	3	2	7	10	2	1	3
3	7	9	0	0	0	0	0	4	4	0	2	0	0	12
4	3	1	0	6	9	2	1	3	4					

Thesis loans can be obtained by subtracting book loans from all loans.

APPENDIX B

Definition and Properties of the DFT

[For further information we refer the reader to, e.g. Firth (1992); Jaffe (1987a, 1987b); Shiavi (1991)].

If $x=(x(n))_{n=0,\ldots,N-1}$ is a complex N-tuple, then the discrete Fourier transform (DFT) of x is defined as the mapping:

$$\mathbb{C}^N \rightarrow \mathbb{C}^N : x \rightarrow X$$

where $X=(X(k))_{k=0,\ldots,N-1}$ is the N-tuple given by:

$$X(k)=\sum_{n=0}^{N-1} x(n)e^{-i2nk\pi/N}$$

The n-th component of the autocorrelation of a complex N-tuple x is calculated as:

$$AUTOCOR_n(x) = \sum_{m=0}^{N-1} x(n+m)\bar{x}(m)$$

If $x(n)$, $n=0,1,\ldots, N\text{-}1$, is a sum of cosines as in equation (B1),

$$x(n) = \sum_{j=1}^{M} A_j \cos(2\pi f_j n) \tag{B1}$$

we can prove that $X(k)=0$ if $k \neq Nf_j$, $j=1,2,\ldots,M$ ($f_j < 1/2$) and that $X(k)=A_jN/2$ if $k=Nf_j$. Consequently, the frequencies f_j can be found by considering the index numbers k for which $X(k)\neq 0$ and putting $f_j=k/N$. The corresponding magnitude A_j is then equal to $2X(k)/N$. We further have the following relation:

$$PS(k) = DFT_k(AUTOCOR(x)) = |X(k)|^2$$

Hence the corresponding amplitudes A_j can also be obtained as $2\,(PS(k))^{1/2}/N$. It can finally be shown that if $x(n)$ is given by (3), then

$$AUTOCOR_n(x) = \sum_{j=1}^{M} B_j \cos(2\pi f_j n)$$

with $B_j = NA_j^2/2$.