P. A. W. Lewis
G. S. Shedler

Statistical Analysis of Non-stationary Series of Events in a Data Base System

Abstract: Central problems in the performance evaluation of computer systems are the description of the behavior of the system and characterization of the workload. One approach to these problems comprises the interactive combination of data-analytic procedures with probability modeling. This paper describes methods, both old and new, for the statistical analysis of non-stationary univariate stochastic point processes and sequences of positive random variables. Such processes are frequently encountered in computer systems. As an illustration of the methodology an analysis is given of the stochastic point process of transactions initiated in a running data base system. On the basis of the statistical analysis, a non-homogeneous Poisson process model for the transaction initiation process is postulated for periods of high system activity and found to be an adequate characterization of the data. For periods of lower system activity, the transaction initiation process has a complex structure, with more clustering evident. Overall models of this type have application to the validation of proposed data base subsystem models.

Introduction

Description of the behavior of a running system and characterization of the workload are central problems in the performance evaluation of data base systems. These are systems in which there are many users who can access, via remote terminals, a (typically very large) data base managed by a computer. Such a system should respond to a query in a reasonably short time, given the number of users and the nature of the user environment. This must be accomplished as economically as possible, where the factors to be considered include direct customer (waiting) costs and computer system resource utilization. This is a typical operations research situation in which we are trying to allocate limited resources in an optimal way among competing demands. Because of the complexity of data base systems, detailed measurements of existing systems are needed in order to model and evaluate them; such measurements comprise just one aspect of performance evaluation, which in its entirety would encompass data collection, analysis, modeling, and interpretation. Ultimate goals of performance evaluation include tuning of existing systems and prediction of the performance of proposed systems.

This paper is concerned with methods for statistical analysis of series of events, which can be applied to obtain a graphical and mathematical description of the *behavior* of a running data base system. Such a description would be a useful starting point for studies aimed at workload characterization. The particular analysis of data given uses a combination of statistical data-analytic procedures and probability modeling (cf. Lewis and Shedler [1]). The specific results reported here for the analysis of a non-stationary univariate series of events occurring in an IMS data base system are intended neither to comprise in themselves a description of the running IMS system nor necessarily to be a sufficient basis for characterizing the workload of an IMS system. Rather, the results are to be considered illustrative of methods that may be useful in such studies.

In a data base system the workload may be taken to be a collection of data sequences identifiable at various levels of the system; workload characterization comprises the study of these data sequences (individually and jointly) along with the transformations among them. We are deliberately vague here about what is meant by data sequence; it could be a sequence of events occurring in time, i.e., a point process, or a sequence of observations of a stochastic process, i.e., a time series. For example, in an IMS data base system we can consider, at the user level, sequences of transactions and DL/I calls; at the logical level, sequences of target segments; at the segments searched level, sequences of path segments; at the paging level, sequences of path blocks, etc. Associated with these identified basic workload data sequences, there may be other data sequences of interest, e.g., the subsequence of path block exceptions. We may also be interested in external measurements related to the workload data sequences such as response times for users.

Given the complexity of data base systems and the resulting relative difficulty of carrying out meaningful

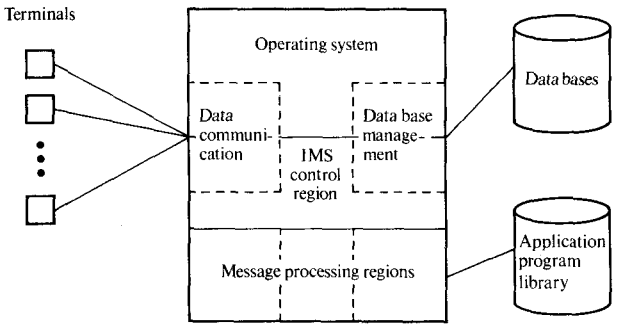

Figure 1 IMS system configuration. Conceptual diagram of a computer system running IMS.

performance evaluations and designs for such systems, the collection and analysis of measurement data from representative systems to identify and characterize significant performance phenomena seems appropriate. The availability of such measurements presents the possibility of obtaining thereby empirically valid, parameterized mathematical models for workload data sequences. However, the sheer volume of data that can be collected from a running data base system (e.g., tens of thousands of transactions per day, hundreds of thousands of DL/I calls per day, millions of path segments per day, etc.) is a source of some difficulty. Such a volume of data is not only costly to manipulate, it is difficult to comprehend. In practice it appears that if we wish to do a detailed analysis (and modeling) of any of the several workload data sequences mentioned above, it is necessary to select "representative" sequences observed during (relatively) short periods of time. If useful information is to be obtained from the data collection, analysis, and modeling (e.g., for the determination of pertinent system requirements), it is important to be able to describe the system context in which the transaction workload phenomena are observed and analyzed.

In addition to models of the workload, models of the system or subsystem structure are needed in performance evaluation. The authors feel that stochastic models of the type obtained in this study have application to the detailing of proposed system models, i.e., filling in the fine structure of parts of the model. A second application is to the "validation" of system models in the sense of establishing their predictive value. The methods used for the statistical analysis of data from the running system can also be used to analyze the output of simulations of proposed subsystem models. Consistency of a process predicted by the system model with the corresponding process observed in the running system would constitute evidence of the predictive value of the model. Thus, for example, the results of the statistical analysis of the transaction initiation process reported here could be used in attempting to validate a stochastic model of the IMS DL/I component such as the queueing model developed by Lavenberg and Shedler [2].

Description of the available data

The analysis given here, illustrating methods for the examination of non-stationary series of events, is of data obtained from an IMS data management system. The following is a brief outline of the structure of IMS [3], which is a processing program for the implementation of large data bases shared in common by several applications. The IMS program executes under the operating system of the computer system to extend the data communication and data base management capabilities of the operating system. In IMS, users can access the data base from remote terminals by entering messages called transactions. A particular transaction uses, and thus uniquely identifies, an application program which processes the message (or transaction) and accesses the data base. The data management facility of IMS is called Data Language/I (DL/I). The two interfaces of an application program with DL/I are a data base description and a program linkage which allows DL/I to process data base access requests that arise during execution of an application program. The execution of an application program thus gives rise to a sequence of calls to the DL/I component of IMS.

A conceptual diagram of a computer system running IMS is given in Fig. 1. As shown there, a portion of memory is devoted to the operating system. The IMS program occupies a portion of memory called the IMS control region. Application programs reside in secondary storage in an application program library. For execution an application program must be loaded into one of several (typically three or four) regions in memory called message processing regions. The data base resides in secondary storage, and data are transferred into memory for processing in response to transaction initiations.

Data on the processing of transactions have been obtained from a computer system running IMS for production control under the IBM operating system OS. Entry of data into the system is on-line and is governed by the occurrence of events on the production line. The epochs of time at which individual DL/I calls were completed (i.e., control was returned to the application program) have been recorded, along with information sufficient to identify the epochs of time at which individual transactions were initiated. From these time stamps the sequence of times between transaction initiations were derived. Most of the results presented in this paper are for a time period of high system activity referred to as time period H. These data consisted of 1999 transaction initiations over a period of time (in unspecified units) of $t_0 =$

11936.6066. Much of the statistical analysis was done using the experimental SASE-IV program (Lewis, Katcher, and Weis [4]) for analyzing series of events. SASE-IV has a maximum input of 1999 events; this accounts for the length of the period under study. This high system activity period was selected after an initial overall look at the several days of data on transaction initiations which were available. The analysis also used SASE-VI, an improved version of SASE-IV, APL implementations of parts of SASE-VI, and APL implementations of rate estimation procedures.

Preliminary analysis of transaction initiation process

• *Prior considerations and assumptions*

In analyzing the transaction initiation data, there were a number of prior assumptions that could be made about the data to serve as a starting point for the analysis. The purpose of the data analysis is to confirm these assumptions or to point to suitable modifications.

1. Since the data are taken over a whole day (in fact, six whole days), we expect a time-of-day effect as activity builds up through the working day and then declines during the evening. Thus, any kind of initial analysis based on an assumption of stationarity is inappropriate.
2. Since the data consist of times of transaction initiations, so that we are dealing with a point process or series of events, the usual null model (which is delineated in a subsequent section of this paper) is a nonhomogeneous Poisson process (NHPP). This could be appropriate here since the transaction initiation process is a superposition (Cox and Lewis [5], Ch. 8; Çinlar [6]) of inputs from a number of sources (users).
3. Because each user's activity is likely to consist of a (random) number of transactions after initial sign-on, some clustering in the data might be expected. An appropriate model here is the nonhomogeneous Poisson cluster process (Lewis [7]). In this process an initial primary (main) event generates a finite sequence of secondary (subsidiary) events; the complete process is then the superposition of the primary and secondary events, where the main events are assumed to be generated by a nonhomogeneous Poisson process. If enough initial events are generated (high-activity) so that the number of active secondary processes is large, this process is hard to distinguish from a Poisson process.

Starting from these assumptions, the analysis of the data proceeded as follows:

a. A very rough, model-free procedure was used to estimate the rate function for the transaction initiation process over the whole day, the rate function being the derivative of the expected number of transactions in a time period (0, t]. This rate would be constant for a stationary (homogeneous) process.
b. On the basis of this trend analysis, relatively homogeneous high- and low-activity periods were selected, and an attempt was made to verify the NHPP model or the clustering model, for the transaction initiation process.
c. Based on this local analysis and modeling of the transaction initiation process, more formal model-dependent estimation procedures were applied to the transaction rate function for the several days. In later sections it will be seen that the Poisson assumption is reasonably valid for high-activity periods, clustering becomes more evident at low-activity periods, and there is a surprising amount of local inhomogeneity of an almost oscillatory (cyclic) nature. It is this last phenomenon that is perhaps the most interesting aspect of the analysis.

• *Analysis of transaction initiation counting process*

Point processes can be analyzed either in terms of the intervals between events, which is a stochastic sequence (time series), or the counting process (the number of events in an interval (0, t]) which, as a function of t, is a continuous-parameter stochastic process. Here 0 is some convenient fixed origin, the number of events in $(0, t]$ is denoted by N_t, and the expected value of N_t is

$$M(t) = \mathrm{E}\{N_t\}. \tag{1}$$

Its derivative, often called the rate function or intensity function, is $m(t) = dM(t)/dt = \lambda(t)$, the notation $\lambda(t)$ being generally used for the rate function of a Poisson process. (See Cox and Lewis [5], Ch. 4, for further definitions of point processes.)

Note that although the times of the transaction events were available, for an initial analysis we used counts of events in successive unit time intervals, i.e., $\Delta = 1$. This constitutes a sampling of the data; if the data were from a NHPP, these counts would be independent variates with possibly different means (see Section 4). Denote these counts by $n_j, j = 1; \cdots, \mathrm{n}$, where $n_j = N_j - N_{j-1}$ and $N_0 = 0$. If these counts are summed to give counts in C contiguous intervals, they will still be Poisson distributed. Such a summation can be considered as

1. A crude smoothing of the data to obtain an estimate and picture of the rate function over the day. Thus, since $\Delta = 1$,

$$\Delta\tilde{m}\left(\frac{C}{2}\right) = \sum_{j=1}^{C} \frac{n_j}{C} = \tilde{m}\left(\frac{C}{2}\right);$$

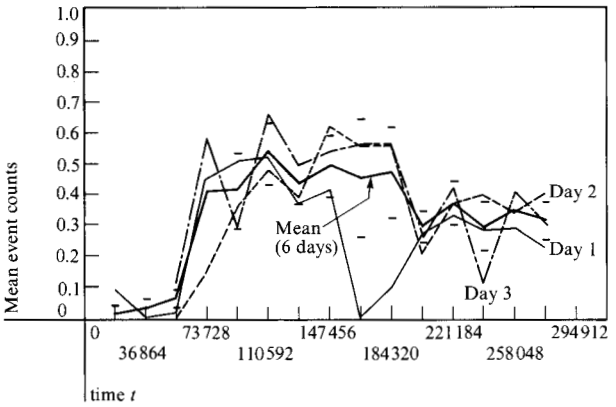

Figure 2 Estimated mean number of transactions initiated in a unit time interval for days 1, 2, and 3. Estimates obtained by averaging counts in 4800 adjacent unit time intervals. This very severe smoothing takes out local fluctuations but gives a picture of how the activity varies over a full day.

the weights in the smoothing all have value $1/C$. This constant smoothing function must be used with care; it can cause spurious effects if the rate is not changing linearly.

2. A coalescing of count data to test for homogeneity.

Plots of the smoothed counts using $C = 4800$ are shown in Fig. 2 for three of the six days, and for the average of the smoothed counts over all six days. Formal tests for homogeneity are available for Poisson variates (Cox and Lewis [5], Ch. 6), or else a one-way analysis of variance can be performed on the coalesced data after a square root transformation. The analysis of variance test is used because the counts are large enough to be considered to be normally distributed; the square root transformation is used because although Poisson counts with a large mean are approximately normally distributed (see Table 2.1 of [5], p. 21), the mean and the variance are the same, and this violates a basic assumption in the analysis of variance test. The square root of a Poisson variate N plus a fourth, $\sqrt{N+\frac{1}{4}}$, has mean approximately equal to $\sqrt{\mu}$, and variance $\frac{1}{4}$, where μ is the Poisson mean ([5], p. 44).

The analysis of selected time periods reported below is for periods chosen from day 2. In Table 1 we show in successive columns the number of counts (transaction initiations) in successive groups of forty 120-time-unit periods; the mean number of counts in one time unit (the rate function estimate plotted in Fig. 2) for day 2; $\bar{x}_i$ the average of forty quantities x_{ij} where $x_{ij} = \{$(number of counts in jth 120-time-unit period in group i) $+ \frac{1}{4}\}^{\frac{1}{2}}$; $\hat{\sigma}_i^2$ and $\hat{\sigma}_i$, the within group sample variance and standard deviation, respectively.

First, it can be seen that all of the variances $\hat{\sigma}_i^2$ are larger than the value $\frac{1}{4}$ postulated on the basis of a homogeneous Poisson count process; since $39 \times \hat{\sigma}_i^2/\frac{1}{4} = 156 \times \hat{\sigma}_i^2$ should, under the null hypothesis, have a χ^2_{39} distribution with upper 99 percent point of 62.281, all the $\hat{\sigma}_i^2$'s are significantly large (i.e., greater than $62.281/156 = 0.3992$) and either the Poisson or homogeneity (within group) assumptions are invalid.

Comparing the sum of the within group sample variances $\hat{\sigma}_i^2$, which is 42.1826, to the between-group variances (or sample variance of the $\bar{x}_i$), which has a value 1.7126 we get an F-ratio of 19.4878. The F-ratio, formally given by

$$F = m\hat{\sigma}^2_{x_i} \Big/ \sum_i \hat{\sigma}_i^2/k,$$

has an F-distribution with $\nu_1 = (m-1) \times k = 39 \times 12$, $\nu_2 = k - 1 = 11$ degrees of freedom, and the value 19.4878 in Table 1 is highly significant at a 5 percent level or at a 1 percent level. We conclude that the data are inhomogeneous, although departure from a Poisson assumption has not been ruled out.

The overall picture in Fig. 2 is of an initial buildup in transaction rate, a fairly constant transaction rate for a period of time, and then a drop to a lower level. This picture is consistent over days; the drop in day 1 (around $t = 165888$) was due to a period for which data were not available.

However, even in the two relatively stable periods, there is some evidence (large values of $\hat{\sigma}_i^2$ in Table 1 relative to $\frac{1}{4}$) of more microscopic inhomogeneity, and the analysis proceeded by examining sections of data in these high- and low-activity periods in more detail. The examination was of interest per se, but was also motivated by a need for more formal statistical rate estimation procedures.

Highly parametric global procedures for rate estimation are available at present only for NHPPs. Details of the procedure and the estimation are given in the next two sections. Application to the data for the high and low system activity periods and for the entire day is described in later sections.

In addition, non-parametric local smoothing procedures related to kernel-type density estimates (Rosenblatt [8]) are used. These are also described later. First we give properties of the NHPP.

Nonhomogeneous Poisson process model

The nonhomogeneous Poisson process model for a series of events N_t is discussed in a statistical context by Cox and Lewis [5], Ch. 3, Lewis [9], Cox [10], and Brown [11]. A very detailed mathematical account is given in Gnedenko and Kovalenko [12]; a recent treatment is by Çinlar [13]. Like the homogeneous Poisson process, the nonhomogeneous Poisson process arises as a limit of the superposition of a large number of nonstationary point processes (cf., Çinlar [6]). The

Table 1 One-way analysis of variance for counts. Transaction initiation process for day 2.

Group i	*Counts in* 4800 *unit time intervals*	*Mean counts in unit time interval*	$\bar{x}_i$	$\hat{\sigma}_i^2$	$\hat{\sigma}_i$
1	1034	0.2154	4.5638	5.6635	2.3798
2	1742	0.3629	6.5178	1.6084	1.2682
3	2455	0.5115	7.6421	3.5629	1.8876
4	1877	0.3910	6.6108	3.8181	1.9540
5	2841	0.5919	8.3752	1.4157	1.1898
6	2925	0.6094	8.5412	0.6898	0.8305
7	2446	0.5096	7.7840	1.0866	1.0424
8	1012	0.2108	4.3684	6.8893	2.6248
9	1910	0.3979	6.7616	2.5957	1.6111
10	1671	0.3483	5.9692	6.8401	2.6154
11	1988	0.4142	6.7364	4.9443	2.2236
12	1880	0.3917	6.6715	3.0682	1.7516
			$\Sigma\bar{x}_i = 80.5420$	$\Sigma\hat{\sigma}_i^2 = 42.1826$	
			$\Sigma\bar{x}_i/12 = 6.7118$	$\Sigma\hat{\sigma}_i^2/12 = 3.5152$	
			$\hat{\sigma}^2_{\bar{x}_i} = 1.7126$		
			$\hat{\sigma}_{\bar{x}_i} = 1.3086$		

assumptions underlying the nonhomogeneous or time-dependent Poisson process (NHPP) are the same as those for the ordinary Poisson process except that the rate parameter λ is now considered to be a continuous function of time $\lambda(t)$. One approach to the NHPP is via the incremental probabilities in small intervals. Thus, for $s, t \geq 0$, and denoting by $N(s;t)$ the number of events in the process in the interval $(t, t+s]$, the assumptions for a NHPP with rate function $\lambda(t)$ are that, as $s \to 0$,

$$\Pr\{N(s;t) = 0\} = 1 - \lambda(t)s + o(s),$$

$$\Pr\{N(s;t) = 1\} = \lambda(t)s + o(s), \qquad (2)$$

and that the random variable $N(s;t)$ is statistically independent of the number and position of events in $(0, t]$. As a consequence of Eq. (2), $\Pr\{N(s;t) \geq 2\} = o(s)$. The survivor function for the forward recurrence time in the process, the probability that there are no events in $(t, t+s]$, i.e., that $N(s;t) = 0$, is derived via first-order differential equations to be

$$R(s;t) = \exp\left[-\int_t^{t+s} \lambda(u)du\right].$$

A more general approach to defining the NHPP starts with the function $\Lambda(t)$, which is assumed to be monotone non-decreasing and continuous from the right; then the number of events occurring in any interval, say $(t, t+s]$, is assumed to have a Poisson distribution with parameter

$$\Lambda(t+s) - \Lambda(t) = \int_t^{t+s} \lambda(u)du,$$

i.e., for $k = 0, 1, 2, \cdots$

$$\Pr\{N(s;t) = k\} = \frac{\{\exp - [\Lambda(t+s) - \Lambda(t)]\}[\Lambda(t+s) - \Lambda(t)]^k}{k!}.$$

Consequently $\Lambda(t)$ is the expected value function $M(t)$, defined by Eq. (1). In addition, the number of events in any finite set of non-overlapping intervals are assumed to be independent random variables. There are other equivalent definitions, and also minimal definitions; see Gnedenko and Kovalenko [21] and Çinlar [13].

The following theorem (cf. Çinlar [13]) establishes that a homogeneous Poisson process of rate 1 can be obtained by transformation of the time scale of an NHPP, via the inverse of $\Lambda(t)$. This result, Theorem 1, and the following Theorem 2 are the bases for procedures described below for detrending the data and testing the goodness-of-fit of the NHPP model.

Theorem 1 Let $\Lambda(t)$ be a non-decreasing right-continuous function of $t \geq 0$. Then $T_1, T_2, \cdots$, are the times-to-events in an NHPP with $E\{N_t\} = \Lambda(t)$, if and only if $T_1' = \Lambda(T_1), T_2' = \Lambda(T_2), \cdots$ are the times-to-events in a homogeneous Poisson process with rate 1.

The next theorem establishes an important property of the NHPP which we use throughout the paper.

Theorem 2 Assume we have an NHPP observed for a fixed time $(0, t_0]$, in which $N_{t_0} = n$ events occur at times $T_1 < T_2 < \cdots < T_n < t_0$. Then conditional on having observed n(>0) events in the $(0, t_0]$, the T_i are distributed as the order statistics from a sample with distribution function

$$F(t) = \frac{\Lambda(t) - \Lambda(0)}{\Lambda(t_0) - \Lambda(0)}, \qquad 0 \le t \le t_0,$$

and when $\Lambda(t)$ is absolutely continuous, probability density function

$$f(t) = \frac{\lambda(t)}{\Lambda(t_0) - \Lambda(0)}, \qquad 0 \le t \le t_0.$$

Thus we see that (conditionally) the transformation of the time axis is exactly the same as the probability integral transform which is used to transform a random variable X with known distribution function $F(X)$ into a uniform random variable on (0, 1), i.e., $U = F(X)$ is uniform (0, 1). This transformation is the basis for nonparametric tests of distribution functions such as the Kolmogorov-Smirnov test. The analogy explains why tests for a homogeneous Poisson process (HPP) are similar to tests for completely specified distributions obtained from independent, identically distributed samples; the primary difference in the two procedures lies in the alternative hypotheses that arise (see Cox and Lewis [5], Ch. 6). Specifically, if we test that a random sample $X_1, X_2, \cdots, X_n$ with unknown distribution function $F(X)$ is from a given distribution function $F_0(X)$, then if $F_0(X) \neq F(X)$, the variables $U_1 = F_0(X_1), \cdots, U_n = F_0(X_n)$ are i.i.d., but not uniformly distributed. However, if we test (conditionally) that n observed times-to-events $T_1, \cdots, T_n$ are from a NHPP with given integrated rate function $\Lambda_0(t)$, then

1. if the process is NHPP but $\Lambda_0(t)$ is not equal to the true integrated rate function $\Lambda(t)$, then $T_1' = \Lambda_0(T_1), \cdots, T_n' = \Lambda_0(T_n)$ are i.i.d., but not uniform $(0, t_0]$,

and

2. if the process is not NHPP, then even if $\Lambda_0(t)$ is equal to $\Lambda(t)$, the T_i', $i = 1, \cdots, n$ are not conditionally a random sample.

The above leads to very different considerations in the power of tests for NHPPs and completely specified distributions, even though the test statistics are the same (see Lewis [14], for greater detail). It is difficult in testing for NHPPs with procedures based on the above theorems, to separate out the effects of departures from Poisson assumptions and departures from assumptions as to the form of $\Lambda(t)$. However, since both HPPs and NHPPs have independent count increments, tests for the global Poisson assumption are based on this property. In particular, the spectrum of counts (Cox and Lewis [5], Ch. 5) should be flat after detrending.

In the following section we discuss estimation of the NHPP rate function using parametric models, both to describe in a global way the rate function (as opposed to the local smoothing in Fig. 2) and to detrend the data so as to examine the global Poisson assumption. Nonparametric rate estimation is also briefly discussed.

Estimation of the NHPP rate function

• *Parametric model and rate estimation*

Following Cox and Lewis [5], Ch. 3, and Cox [10], an exponential polynomial rate function has been assumed for the NHPP, i.e., $\lambda(t)$ of the form

$$\lambda(t) = \exp\left(\sum_{m=0}^{r} \alpha_m t^m\right), \qquad -\infty < \alpha_0, \alpha_1, \cdots, \alpha_r < +\infty \quad (3)$$

This assumption is convenient and constitutes no real restriction because any continuous rate function can be approximated arbitrarily closely by an exponential polynomial. The result follows from results on ordinary polynomials by taking logarithms; note that $\lambda(t) \geq 0$ for any values of $\alpha_0, \alpha_1, \cdots \alpha_r$. We now describe statistical procedures based on this model. Formal tests for the degree r of an exponential polynomial rate function are discussed in the following section. Here a procedure is outlined for the maximum likelihood estimation of the coefficients $\{\alpha_m\}$ of an exponential polynomial of fixed degree r.

The times-to-events $T_1 < T_2 \cdots < T_n$ in a fixed time period and the random variable $N(t_0) = n$ have a joint density function (Cox and Lewis [5], Ch. 3)

$$f(t_1, \cdots, t_n; n) = \left[\exp - \int_0^{t_0} \lambda(u)\,du\right] \prod_{i=1}^{n} \lambda(t_i),$$

which, on substituting the rate function (4), becomes

$$f(t_1, \cdots, t_n; n) = \exp\left[-\sum_{m=0}^{r} \alpha_m s_m - \int_0^{t_0} \exp\left(\sum_{l=0}^{r} \alpha_l t^l\right) dt\right], \quad (4)$$

where

$s_m = t_1^m + \cdots + t_n^m$, $m = 0, \cdots, r$.

Thus the log-likelihood function, $\log L$, the logarithm of the density at the observed values of the random variables considered as a function of the $r + 1$ parameters, is

$$\log L(\alpha_0, \alpha_1, \cdots, \alpha_r) = \sum_{m=0}^{r} \alpha_m s_m - \int_0^{t_0} \exp\left(\sum_{l=0}^{r} \alpha_l t^l\right) dt.$$

It follows that the derivatives, known as the scores, are

$$\frac{\partial \log L}{\partial \alpha_k} = s_k - \int_0^{t_0} t^k \exp\left(\sum_{m=0}^{r} \alpha_m t^m\right) dt, \qquad k = 0, 1, \cdots, r. \quad (5)$$

The solutions $\{\hat{\alpha}_m\}$ to the system of Eqs. (9), the score vector when set to zero, are the maximum likelihood estimators of $\{\alpha_m\}$, and can be determined numerically

by Newton-Raphson iteration. The numerical procedure works well provided that an initial vector sufficiently near the solution is known. A two-step method for obtaining such an initial value has been proposed by MacLean [15]. His procedure consists of finding an ordinary polynomial representation of the same degree as $\lambda(t)$ having the observed sums of powers $\{s_m\}$ for its "moments". An exponential polynomial approximation to this polynomial, obtained by taking logarithms and again fitting moments, serves as the initial value for the Newton-Raphson iteration. This MacLean procedure has been implemented in APL and used to estimate coefficients $\{\hat{\alpha}_m\}$. The procedure appears to work well for polynomials up to degree 8. Estimates of the covariance matrix of the maximum likelihood estimates $\{\hat{\alpha}_m\}$ are obtained from the second order partial derivatives of the log-likelihood equation when evaluated at the estimated parameter values.

Once the appropriate degree of the polynomial is obtained by the methods of the next section, the rate function with the maximum likelihood estimates for the α's can be plotted to obtain a picture of the rate function. The procedures are clearly sensitive to the NHPP model; for this reason, we next discuss nonparametric kernel-type estimates.

• *Non-parametric kernel-type rate estimates*

Theorem 2, which relates (conditionally) the rate function $\lambda(t)$ in an NHPP to a density function in $(0, t_0]$,

$$f(t) = \frac{\lambda(t)}{\Lambda(t_0) - \Lambda(0)}, \qquad 0 \le t \le t_0,$$

suggests we could use nonparametric probability density function estimates to estimate rate functions, at least in NHPPs. The procedure chosen is the nonparametric kernel-type density estimate introduced by Rosenblatt [8]. Briefly, the procedure to estimate $f(t)$ from a random sample $T_1, T_2, \cdots, T_n$ is as follows:
Define

$$\hat{f}_n(t) = \frac{1}{nb(n)} \sum_{j=1}^{n} W\left(\frac{t - T_j}{b(n)}\right),$$

where $W(u)$ is a bounded non-negative integrable weight function with $\int_{-\infty}^{\infty} W(u)du = 1$, and $b(n)$ is a positive bandwidth function which tends to zero as $n \to \infty$, but is such that $o(b(n)) = 1/n$. Thus, we might have $b(n) \sim n^{-\frac{1}{2}}$, for example.

Note that for a given set of observations all estimates of this form are themselves density functions, i.e.,

$$\hat{f}_n(t) \ge 0, \int_{-\infty}^{\infty} \hat{f}_n(u)du = 1,$$

and since the T_j are random variables, $\hat{f}_n(t)$ is a continuous-parameter stochastic process, but clearly non-stationary. Although this type of density estimate does not require parametric assumptions to be made about $f(t)$, the bandwidth function and kernel $W(u)$ must be chosen. In this paper we have already chosen $W(u)$ to be a triangular function and $b(n)$ to be $1.25/n^{\frac{1}{2}}$.

The conditional structure of the NHPP makes the estimation of the rate function $\lambda(t)$ similar to the nonparametric estimation of the density function, but with two differences. First, care must be taken with normalization of the rate function estimate. This is because the procedure above estimates the rate normalized by dividing by $\Lambda(t_0) - \Lambda(0)$ and $\Lambda(t_0) - \Lambda(0)$ is unknown. For an NHPP this is the mean of a Poisson variable which is estimated by n, the number of events in $(0, t_0]$. We then get, as a rate function estimate,

$$\hat{\lambda}(t; n, t_0) = n\hat{f}_n(t) = \frac{1}{b(n)} \sum_{j=1}^{n} W\left(\frac{t - T_j}{b(n)}\right).$$

This will be modal about the usual estimate of the rate λ in a homogeneous Poisson process, which is estimated by $\hat{\lambda} = n/t_0$. The second difference is that when the density function estimation technique is applied to rate function estimation, there is no asymptotic justification for the procedure.

Tests for the degree of the exponential polynomial rate function

• *Theory*

The analysis of trends in an NHPP, based on the assumption of an exponential polynomial rate function, is discussed in Cox and Lewis [5] (Ch. 3), and Lewis [9]. In the latter paper, formal tests for the linear and quadratic terms in the exponential polynomial are derived. We use here a direct extension of these methods to yield tests for higher degree terms.

There are a number of possible hypotheses which can be tested when considering the exponential polynomial rate function

$$\lambda(t) = \exp\left(\sum_{m=0}^{r} \alpha_m t^m\right). \tag{10}$$

1. Some given subsets of the $r + 1$ parameters are zero. Asymptotic tests for this hypothesis are based on standard maximum likelihood arguments; see Cox [10] and MacLean [15] for details. Essentially the maximum values of the likelihood functions under the two hypotheses are compared; the difference has (asymptotically) a χ^2 distribution under the null hypothesis with known degrees of freedom. The problem with this test is phenomenological; one seldom knows a priori which subset to test.
2. It is possible to ask which subset of the $r + 1$ parameters gives the best (most parsimonious) fit to the

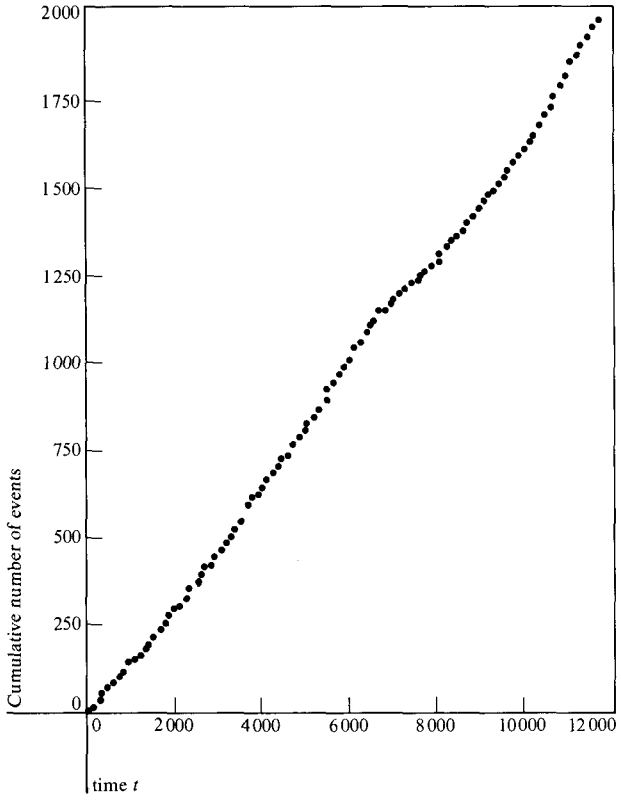

Figure 3 Cumulative number of transactions initiated for time period H (high-activity). There is enough departure from linearity to suggest inhomogeneity in the data. The test for a homogeneous Poisson process using the Kolmogorov-Smirnov statistic confirms the departure; the value 2.389 of the Kolmogorov-Smirnov statistic is highly significant.

data. This has been worked out for ordinary, normal theory linear polynomial regression (Daniel and Wood [16], but not for the NHPP case.

3. An alternative is to test for successive inclusion of higher order polynomial terms. This is reasonable if the exponential polynomial is being used in a purely descriptive way, and the statistical theory is known. Strictly, we test that, for some $k \geq 1$, $\alpha_0 \neq 0$, $\alpha_1 \neq 0$, $\cdots$, $\alpha_k \neq 0$, $\alpha_{k+1} = 0$, $\alpha_{k+2} = 0, \cdots$. (The analogous normal time series case is considered in great detail in Anderson [17], Ch. 2.) A possible drawback would occur where there is a cyclic effect, e.g..

$$\lambda(t) = \exp[\alpha_0 + k \sin(\omega_0 t + \theta)].$$

The series expansion of $\sin(\omega_0 t + \theta)$ gives a polynomial with alternating zero and nonzero coefficients for powers of t if the phase angle is appropriate. This, in turn, is tied into the starting point of observations.

We develop the procedure now for case 3; we have used it in an ad hoc manner by testing until two or more successive zero coefficients occur. For an NHPP with exponential polynomial rate function

$$\lambda(t) = \exp\left(\sum_{m=0}^{r} \alpha_m t^m\right),$$

the likelihood of n events in the period $(0, t_0]$ at times $t_1 < t_2 < \cdots < t_n$ is, from Eq. (4),

$$L(\alpha_0, \alpha_1, \cdots, \alpha_r) = \exp\left\{\sum_{m=0}^{r} \alpha_m s_m - \int_0^{t_0} \exp\left(\sum_{l=0}^{r} \alpha_l t^l\right) dt\right\} \tag{6}$$

where

$$s_m = \sum_{i=1}^{n} t_i^m, \qquad m = 0, 1, \cdots, r.$$

The observations $\{t_i\}$ enter Eq. (6) only through $(n, \Sigma t_i, \Sigma t_i^2, \cdots, \Sigma t_i^r)$, and it can be shown from the exponential form of Eq. (6) that these are a set of sufficient statistics for the set of parameters $\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_r$. There is, however, even more structure and a formal test for the rth degree term in the exponential polynomial rate function can be based on the idea that for any given r and α_r, $(n, \Sigma t_i, \cdots, \Sigma t_i^{r-1})$ are a set of sufficient statistics for $\alpha_0, \alpha_1, \cdots, \alpha_{r-1}$, i.e., the distribution of Σt_i^r, given n, $\Sigma t_i, \Sigma t_i^{r-1}$, is independent of $\alpha_0, \cdots, \alpha_{r-1}$ for all values of α_r. This is convenient since we want to test $\alpha_r = 0$ against $\alpha_r \neq 0$ regardless of the values of $\alpha_0, \cdots, \alpha_{r-1}$, i.e., they are nuisance parameters.

Denoting t_i/t_0 by u_i and $\Sigma u_i^l/n$ by c_l, a test for α_r is then based on the statistic c_r and its null hypothesis conditional distribution, given n, $c_1, \cdots, c_{r-1}$. This distribution is not known for small t_0 (equivalently small n). However, asymptotically $c_1, c_2, \cdots$, and c_r will be jointly normally distributed with mean value and variance that can be obtained from properties of the uniform distribution. We assume a uniform $(0, t_0]$ distribution for the t_i since $(n, \Sigma t_i, \cdots, \Sigma t_i^{r-1})$ are a set of sufficient statistics for $\alpha_0, \alpha_1, \cdots, \alpha_{r-1}$, so that assuming these parameters to have value zero does not affect the final result but does simplify computations. Then, also asymptotically, the conditional distribution of c_r, given $n, c_1, \cdots, c_{r-1}$ is normally distributed with mean $\mu_r = \mathrm{E}(c_r | c_{r-1}, c_{r-2}, \cdots, c_1, n)$ and variance $\sigma_r^2 = \mathrm{Var}(c_r | c_{r-1}, \cdots, c_1, n)$ obtainable from normal theory.

The normal theory results are that to test the null hypothesis $H_0: \alpha_r = 0$, $\alpha_{r+1} = 0, \cdots$, but $\alpha_0, \cdots, \alpha_{r-1}$ have any value, compute the statistic

$$U_r = (c_r - \mu_r)/\sigma_r$$

and test as a mean 0, variance 1 normal deviate, i.e., accept H_0 at, say, a 5% level if $|U_r| \leq 1.96$. Expressions for μ_r and σ_r have been derived by techniques of sym-

bolic mathematics and the matrix operations above. Details of the derivation will be reported elsewhere. The case $r = 1$ is discussed in detail in Cox and Lewis [5], Ch. 3.

• *Applications to high activity (H) data*

We discuss now the application of the parametric rate function testing scheme of the previous subsection and the rate estimation procedures of the preceding section to a more microscopic examination of the transaction initiation process during a period of high system activity for day 2. This high-activity period is, in Fig. 2, from approximately $t = 73728$ to $t = 85661$. We also use the kernel-type density estimate described in the second part of the preceding section. We do this most particularly because the NHPP assumption has, at this point, not been validated. Overall characteristics of the sample are shown in Table 2. (The sample moments given there should be used only as a guide; they are meaningless if the data are inhomogeneous.)

The first question to be addressed is whether the data can, in this relatively short high-activity period, be considered to be approximately homogeneous or stationary.

Figure 3 shows the cumulative number of transactions initiated during this time period. The departure from linearity is fairly gross; assuming a homogeneous Poisson process, the Kolmogorov-Smirnov measure of the departure from linearity is

$$D_n = \sqrt{n} \sup_{0 \le u \le 1} |F_n(u) - u|, \tag{7}$$

where

$$F_n(u) = \frac{\text{number of } t_i \le ut_0}{n}, \qquad 0 \le u \le 1. \tag{8}$$

This is the uniform conditional test in Cox and Lewis [5], Ch. 6; conditional on the observed value $N_{t_0} = 1999$ of events in $(0, t_0]$ it has the usual Kolmogorov-Smirnov statistic distribution with upper 1% point 1.628; the observed value is 2.389, which is an event of very small probability *under the Poisson assumption.*

These probabilities could be grossly in error if the data were more dispersed than under the Poisson assumption, where by dispersion we mean either that the standard deviation of the intervals between events or the counts of events in long intervals is larger than would be expected under a Poisson assumption. (The two are not independent.) These dispersions are usually measured by first normalizing to give the random variable Z mean one; for intervals, the result is the coefficient of variation, i.e.,

$$C(Z) = \frac{\text{S.D. } (Z)}{\text{E}(Z)} = \frac{\sigma(Z)}{\text{E}(Z)} = \sigma\left(\frac{Z}{\text{E}(Z)}\right).$$

Table 2 Sample characteristics of times-between-events. Transaction initiation process for time period H.

n	number of transactions initiated	1999
t_0	period of observation	11936.6066
$\overline{X}$	estimated mean time between transaction initiations	5.9698
$\hat{C}(X)$	estimated coefficient of variation of times between transaction initiations	1.0533
$\hat{\gamma}_1(X)$	estimated coefficient of skewness of times between transaction initiations	6.7399
$\hat{\gamma}_2(X)$	estimated coefficient of kurtosis of times between transaction initiations	107.7282
$X_{\max}$	maximum time between transaction initiations	133.6488
$X_{\min}$	minimum time between transaction initiations	0.0152

Table 3 Sample characteristics of times-between-events. Transaction initiation process for ten sections of time period H.

Section	*Sample mean* $\overline{X}$	*S.D. of mean* $\hat{\sigma}(\overline{X})$	*Coeff. of variation* $\hat{C}(X)$	*Coeff. of skewness* $\hat{\gamma}_1(X)$	*Coeff. of kurtosis* $\hat{\gamma}_2(X)$
1	7.4645	0.5561	0.8430	2.3096	11.4548
2	6.0584	0.3577	0.8328	2.2653	12.1494
3	5.4878	0.5414	1.3916	7.7614	84.6585
4	6.1348	0.3822	0.8789	1.1901	3.9449
5	5.0611	0.2854	0.7954	2.9991	18.6264
6	6.8133	0.4689	0.9708	2.2651	9.6977
7	7.5992	0.7779	1.4440	8.0075	89.8598
8	6.2456	0.3831	0.8652	1.8087	7.1174
9	4.2847	0.2425	0.7984	1.6654	6.6807
10	4.5566	0.2513	0.7750	1.7533	8.1512
Mean	5.9706	0.4137	0.9598	3.2026	25.2341
S.D. mean	0.3591	0.0506	0.0783	0.7952	10.4197

To examine the dispersion of the intervals in the data without confounding it with the apparent inhomogeneity, the 1999 intervals were divided into ten non-overlapping sections. The sample characteristics for each interval are shown in Table 3. The means within each group could be used to test for inhomogeneity, but more importantly the coefficients of variation, skewness and kurtosis, which for exponentially distributed intervals have values 1, 2, and 9, respectively, give us rough measures of departure that are sufficient to validate the tests for trend.

Table 3 gives no indication that the sample characteristics of the intervals of the process depart from an exponential distribution (although there may be correlation between intervals). Values for the sample coefficients of variation are all around unity, as is the sample co-

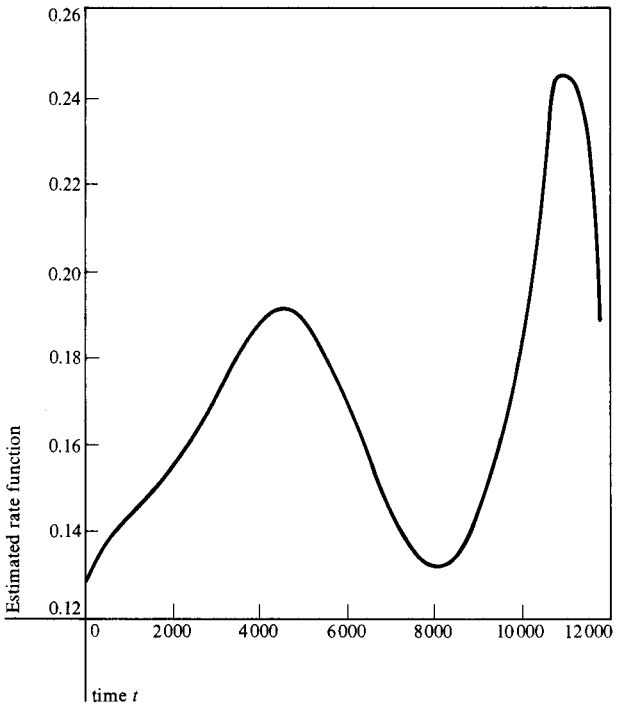

Figure 4 Estimate $\hat{\lambda}(t; \underline{\hat{\alpha}})$ of NHPP rate function using exponential polynomial (degree 6) of transaction initiation process for time period H (high-activity).

Table 4 Values of maximum log-likelihood and test statistic in NHPP exponential polynomial rate function for times between transaction initiations for time period H.

Degree of polynomial (r)	*Maximum log-likelihood* (max log. L)	*Absolute difference* (δ)	*Test statistic* (U_r)
1	−5563.8		3.8387
2	−5562.8	1.0	1.5727
3	−5549.4	13.4	5.3138
4	−5548.9	0.5	−0.4437
5	−5539.9	10.0	−4.2081
6	−5537.0	2.9	−2.6188
7	−5536.9	0.1	0.0188
8	−5536.8	0.1	0.1211
9	−5536.8	0.0	0.2038

Table 5 Estimated values of the coefficients $\{\hat{\alpha}_m\}$ in NHPP exponential polynomial rate function (degree $r = 6$) for times between transaction initiations for time period H.

m	α_m	$\hat{\alpha}_m t_0^m$
0	−2.1381	−2.1381
1	3.1832×10^{-4}	3.7996
2	-2.2607×10^{-7}	−32.2109
3	1.0211×10^{-10}	173.6660
4	-2.1286×10^{-14}	−432.1270
5	1.9331×10^{-18}	468.4494
6	-6.2664×10^{-23}	−181.2609

efficient of variation for the whole set of data as given in Table 2. We therefore proceed to use techniques based on the NHPP model to examine the trend in more detail; further tests of the Poisson assumption for this section of data are given in the next section.

Table 4 gives successive test statistic values for the tests for null parameters in the exponential polynomial model of Eq. (3),

$$\lambda(t) = \exp\left(\sum_{m=0}^{r} \alpha_m t^m\right).$$

This procedure has been described earlier, and as remarked there, is used fairly informally. A formal application would suggest stopping at $r = 2$ and accepting a log-linear model

$$\lambda(t) = \exp(\alpha_0 + \alpha_1 t),$$

but the test statistic for α_3, $U_3 = 5.3138$ is significantly large, and the tests have been continued up to $r = 9$. For $r = 7, 8, 9$, the test statistics are all small, well within the 5% limits of ± 1.96.

Table 4 also gives the values of the log-likelihood function evaluated at the maximum likelihood estimates. The log-likelihood must increase as more parameters are added; the difference, when suitably normalized, is used to test (asymptotically) for inclusion or exclusion of parameters (see MacLean [15] or Cox [10]), and is known asymptotically to have a χ^2 distribution. The absolute differences δ, given in column three of Table 4, are clearly correlated with values of the test statistic U_r, e.g., the large jump of 13.4 when including α_3 in the likelihood goes with a large value of U_3.

The results of both the U_r statistic and the likelihood function values suggest that an exponential polynomial of degree 6 will fit the data very well. The maximum likelihood estimates of the parameters and normalized values are given in Table 5. In computing these estimates in an APL program using MacLean's starting procedure, it is necessary to use normalized time $t/t_0 = u$ and normalized parameters $\alpha'_m = \alpha_m t_0^m$ to avoid scale problems.

The resulting estimated rate function $\hat{\lambda}(t; \underline{\hat{\alpha}})$ is plotted for the high-activity period in Fig. 4. The data give an intimation of a growth plus cyclic effect of fairly long period. A model for this could be

$$\lambda(t) = \exp[\alpha_0 + \alpha_1 t + \alpha_2 \sin(\omega_0 t)];$$

this equation is linear in the parameters if ω_0 is fixed and known (e.g., time of day effect). Moreover if the Taylor series expansion for the sine function is used, one has an exponential polynomial with even index parameters (beyond zero) equal to zero, i.e., $\alpha_2 = \alpha_4 = \alpha_6 = \cdots = 0$. This is the reason why the test for the order of the exponential polynomial indicated that we should have

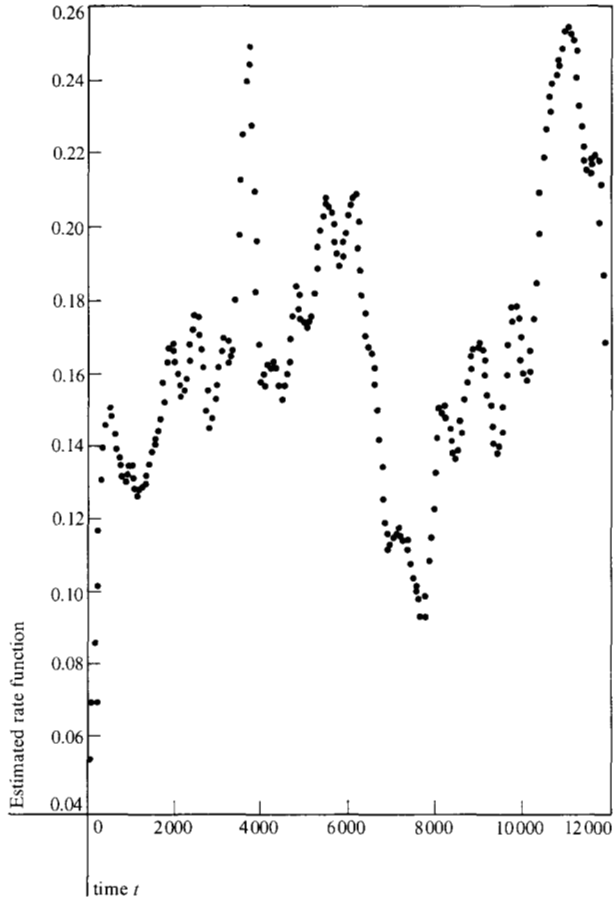

Figure 5 Estimate $\hat{\lambda}(t; n, t_0)$ of the rate function of the transaction initiation process for time period H (high-activity) using a kernel-type density estimator. Sample size $n = 1999$, bandwidth $b(n) = 1.25/n^{\frac{1}{2}}$. Triangular window.

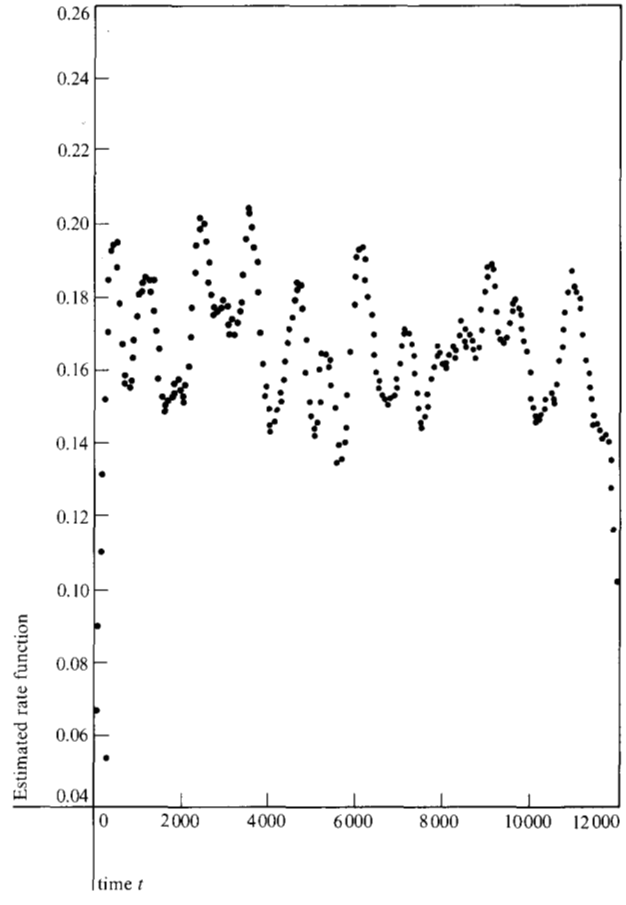

Figure 6 Estimate of the rate function of a homogeneous Poisson process of rate n/t_0. The estimator is the same as in Fig. 5. The greater dispersion in Fig. 5 is due to departures from the Poisson process, most probably inhomogeneity.

stopped at $r = 2$, and then gave an indication that α_3 was non-zero. Cyclic effects are more easily handled via spectral methods; we return to this in the final section.

Another way to examine the trend is to use the kernel-type local smoothing techniques. Although these have broader applicability than the particular global fitting under an NHPP assumption, they suffer as in all non-parametric density estimation (spectra, rate functions, probability density functions, intensity functions) from the need to choose a suitable kernel and bandwidth. In practice, it is usually reasonable to take a few different bandwidths and, by eye, judge when a balance between small variability and small bias is achieved.

A kernel-type rate function estimate $\hat{\lambda}(t; n, t_0) = n\hat{f}_n(t)$ with bandwidth $b_n = 1.25/n^{\frac{1}{2}}$ (chosen in the above way) is shown in Fig. 5. It again shows possible oscillatory behavior in the data, or greater dispersion that we would expect under an HPP assumption. Confidence bands for this type of estimate are available (Bickel and Rosenblatt [18], Lewis et al. [19]), but we have preferred to give, in Fig. 6, an identical smoothing of a simulated homogeneous Poisson process of rate $\hat{\lambda} = n/t_0$. Comparison of Figs. 5 and 6 graphically illustrates that the data are not HPP. The lack of gross departures from Poisson-type characteristics for the interval structure was discussed above; over-dispersion, rather than a trend, could give the large fluctuations in the rate estimate.

In Fig. 5 there is a large peak at about $t = 3000$; we have examined the data for any obvious anomalies at this point (e.g., very regular intervals) but have found none. In Fig. 7 we have overlaid the estimated integrated rate function $\hat{\lambda}(t; \underline{\hat{\alpha}})$ (exponential polynomial degree 6) on the empirical estimate of the integrated rate function which is just the cumulative number of events in $(0, t]$ as a function of t.

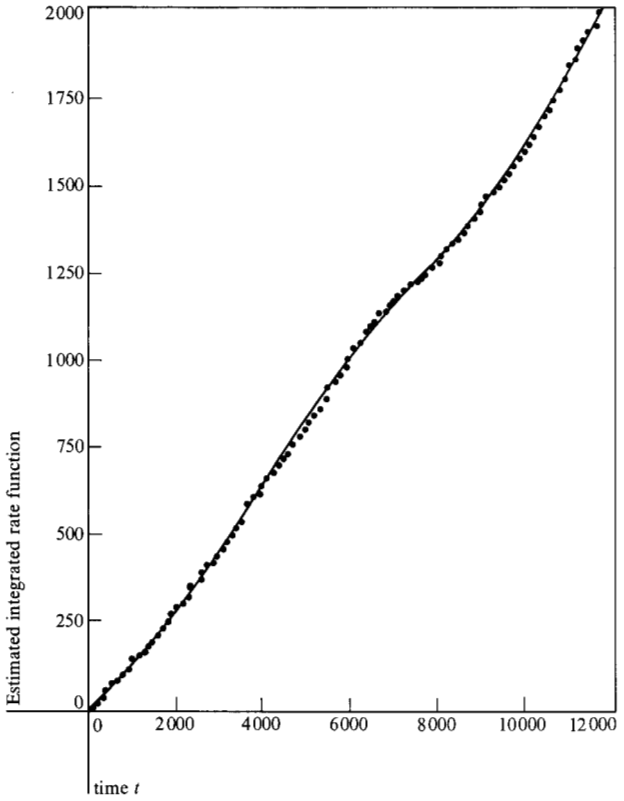

Figure 7 Parametric and empirical estimates of the integrated rate function for the period H (high-activity). Solid curve in the NHPP estimate $\hat{\Lambda}(t; \underline{\hat{\alpha}})$ using exponential polynomial (degree 6). Dotted curve is the cumulative number of events for time period H.

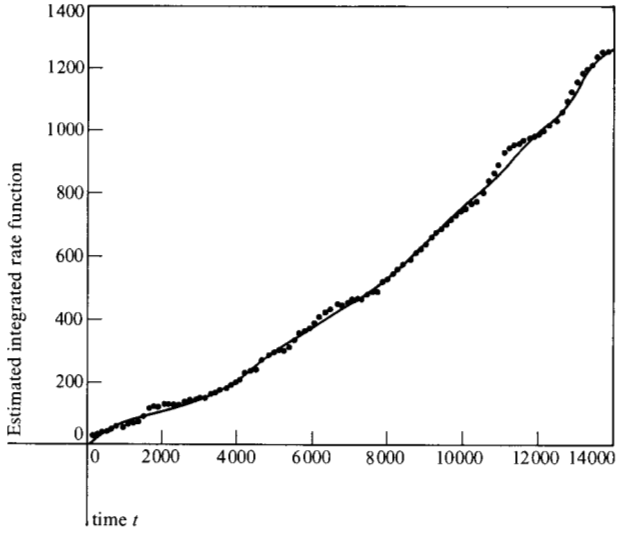

Figure 8 Parametric and empirical estimates of the integrated rate function for time period L (low activity). Solid curve is the NHPP estimate $\hat{\Lambda}(t; \underline{\hat{\alpha}})$ using exponential polynomial (degree 8). Dotted curve is the cumulative number of events for time period L.

Table 6 Sample characteristics of times-between-events. transaction initiation process for time period L.

n	number of transactions initiated	1258
t_0	period of observation	13819.5193
$\overline{X}$	estimated mean time between transaction initiations	10.9809
$\hat{C}(X)$	estimated coefficient of variation of times between transaction initiations	1.6563
$\hat{\gamma}_1(X)$	estimated coefficient of skewness of times between transaction initiations	3.7524
$\hat{\gamma}_2(X)$	estimated coefficient of kurtosis of times between transaction initiations	18.9686
$X_{\max}$	maximum time between transaction initiations	145.4241
$X_{\min}$	minimum time between transaction initiations	0.0263

• *Applications to low-activity (L) data*

We now give, in abbreviated form, an analysis of low-activity (L) data, which is similar to that given for high-activity (H) data in the preceding section. The low-activity data are for the period beyond $t = 145152$ in Fig. 2; the data are for a time period of approximately 1.15 times as long as for the high-activity (H) data, and only 1258 events (transaction initiations) occur. Overall characteristics of the sample are shown in Table 6.

An immediate observation from Table 6 is that the coefficient of variation of the intervals is high relative to the value one for an exponentially distributed random variable. To examine this further, five sections of the data were taken and the interval characteristics which were computed are given in Table 7. Each section of data contained 251 observations. It is fairly apparent that the means are decreasing (rate is increasing) over the five sections, the successive differences, on the basis of the estimated standard deviations of the mean estimates, being about three standard deviations. However, all the coefficients of variation, coefficients of skewness, and kurtosis are larger than the corresponding values for a Poisson process.

The first conclusion from the above analysis is that parametric detrending for these low-activity data must be done with care; we return in the next section to consider details of the structure of the low-activity process, but because the intervals are more dispersed than for a Poisson process, there is consistency with a cluster process hypothesis (Lewis [7], Vere-Jones [20]). Note, too, that a cluster process will look more and more like a Poisson process as activity increases and this is consistent with the finding that the high-activity data are approximately Poisson.

Table 7 Sample characteristics of times-between-events. Transaction initiation process for five sections of time period L.

Section	*Sample mean* $\overline{X}$	*S.D. of mean* $\hat{\sigma}(\overline{X})$	*Coeff. of variation* $\hat{C}(X)$	*Coeff. of skewness* $\hat{\gamma}_1(X)$	*Coeff. of kurtosis* $\hat{\gamma}_2(X)$
1	18.4683	1.6760	1.4378	2.2573	7.9515
2	12.5333	1.2289	1.5534	3.5112	16.6713
3	9.2178	0.9318	1.6015	5.0123	32.5160
4	8.2978	0.9430	1.8005	4.2290	23.2152
5	6.2124	0.3806	0.9706	3.7494	23.7669
Mean	10.9459	1.0321	1.4728	3.7519	20.8242
S.D. mean	2.1390	0.2116	0.1386	0.4532	4.0867

Table 8 Values of maximum log-likelihood and test statistic in NHPP exponential polynomial rate function for times between transaction initiations for time period L.

Degree of polynomial (r)	*Maximum log-likelihood* (*max log L*)	*Absolute difference* (δ)	*Test statistic* (U_r)
1	−4203.7		11.6960
2	−4203.6	0.1	1.2031
3	−4200.4	3.2	−2.4203
4	−4199.2	1.2	0.8175
5	−4191.0	8.2	−3.6703
6	−4190.2	0.8	0.4564
7	−4187.4	2.8	−2.3417
8	−4174.4	13.0	−5.0208
9	–	–	−5.9505
10	–	–	2.7145

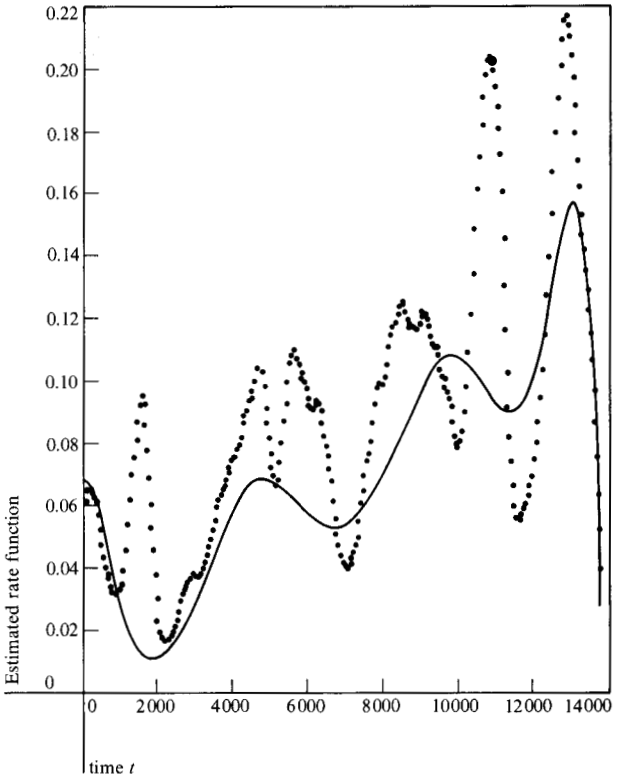

Figure 9 Estimates of the rate function for time period L (low-activity). Solid curve is the NHPP estimate $\lambda(t; \underline{\hat{\alpha}})$ using exponential polynomial (degree 8). Dotted curve is the estimate $\hat{\lambda}(t; n, t_0)$ using a kernel-type density estimator. Sample size $n = 1258$, band-width $b(n) = 1.25/n^2$. Triangular window.

Returning to the trend analysis, we show in Fig. 8 the cumulative number of events in $(0, t]$ as a function of t, which is a nonparametric estimate of the integrated rate function (dotted curve). It is by no means linear, and the Kolmogorov-Smirnov test statistic (see Eqs. (7) and (8) has value 6.048. This, we surmise, is significantly large even if the Poisson hypothesis were not true.

In Table 8 we give the successive test statistics U_r for successively more complicated exponential polynomial rate functions. There is a very definite overall increase in the rate, as measured by $U_1 = 11.696$, and again a phenomenon where U_2, U_4 and U_6 are not significant. However, it can also be seen that the tests are significant out to $r = 10$; it is not possible, even if it were desirable, to carry out the computations any further. The maximum log-likelihoods are also given in Table 8. Since the data are non-Poisson, the likelihoods must be interpreted very carefully. It is conceivable that using a likelihood based on a Poisson process would force the rate estimation procedure to fit the irregularity due to overdispersion by added local wrinkles in the rate function. It is, in fact, always difficult to discriminate between inhomogeneity and over-dispersion, but it is almost certain that it is the over-dispersion which gives rise to the high degree of the fitted polynomial for these data.

With the above qualifiers in mind, we have fitted an exponential polynomial of degree 8 to the data. Degree 8 was chosen because of computational limitations. The integrated rate function $\hat{\Lambda}(t; \underline{\hat{\alpha}})$ is shown overlaid on the non-parametric estimate in Fig. 8; the eighth degree exponential polynomial rate function $\hat{\lambda}(t; \underline{\hat{\alpha}})$ with estimated parameters is shown in Fig. 9 (solid curve). Again the outstanding feature is the cyclic nature of the rate, superposed on a generally increasing rate.

The kernel-type estimator $\hat{\lambda}(t; n, t_0)$ of the rate function is also shown in Fig. 9 (dotted curve); it is clear in comparing it to the exponential polynomial rate function estimate that the procedure using the NHPP assumption works well despite the apparent departures from a Poisson process; if anything, there is a fairly clear validation of the results in Table 8 that an exponential polynomial rate function of degree higher than 8 is needed.

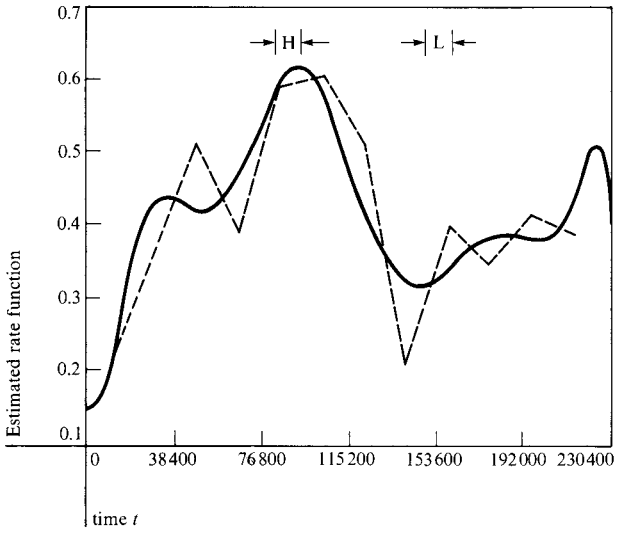

Figure 10 Estimates of the rate function of transaction initiation process for day 2. Solid curve is a global estimate based on an exponential polynomial of degree 9. Dotted curve is local estimate obtained as in Fig. 2. The high-activity (H) and low-activity (L) time periods are marked on the figure.

Table 9 Estimated values of the coefficients $\{\hat{\alpha}_m\}$ in NHPP exponential polynomial rate function (degree $r = 8$) for times between transaction initiations for time period L.

m	$\hat{\alpha}_m$	$\hat{\alpha}_m t_0^m$
0	-2.4784	-2.4784
1	5.7575×10^{-4}	7.9566
2	-2.4040×10^{-6}	-459.1064
3	1.6908×10^{-9}	4462.4093
4	-5.1474×10^{-13}	-18774.5509
5	8.2145×10^{-17}	41404.4332
6	7.1769×10^{-21}	-49991.0659
7	3.2524×10^{-25}	31307.4980
8	-5.9823×10^{-30}	-7958.1157

It is also of interest to note that the estimated parameters $\hat{\alpha}_r$ with even index r are negative (Table 9), a pattern similar to that for the high-activity data shown in Table 5, where $\hat{\alpha}_0$, $\hat{\alpha}_2$, $\hat{\alpha}_4$ and $\hat{\alpha}_6$ are negative, the remaining estimated $\hat{\alpha}_m$ being positive. This is again illustrative of the cyclic effect in the data. It is difficult to compare the magnitude of the estimates in the two periods since, if there were a cycle in the data, the relative phase at the beginning of the period of observations would influence the parameter values.

• *Applications to complete days data*

Recall that a very rough smoothing produced the smoothed estimate of the rate of transaction initiations given in Fig. 2. It is of interest to apply the global smoothing based on an NHPP assumption and an exponential polynomial rate function to the complete days data, even though they are not Poisson at low-activity, so as to have a formal, easily implemented procedure for this type of data that does not involve a choice of smoothing functions and bandwidths.

Over the whole day, 25,076 transaction initiations were observed; details of the testing for the degree of the exponential polynomial, and the values of the estimated parameters, are not tabulated here. Briefly, the tests up to $r = 10$, except for $r = 2$, indicate that the parameters are non-zero. Computation of the moments for the U_r only up to $r = 10$ imposes a limitation on the fit; more importantly, estimation of parameters in an exponential polynomial for an entire day's data is not feasible for degree greater than 9. Thus in Fig. 10 we have overlaid on the rate estimate for day 2 data (given in Fig. 2) an exponential polynomial of degree 9. The agreement between the two estimates is good.

We would expect that as the degree of the polynomial went up, the local fluctuations for the high- and low-activity sections would appear. The computational problems, however, are horrendous; it would be simpler to connect up polynomial rate function estimates within smaller, contiguous sections. This has not been pursued; in particular, it is not clear that the polynomials would connect smoothly.

The overall conclusion of this section is that the data are grossly nonhomogeneous; possible reasons will be discussed in a subsequent section.

Tests of fit of the NHPP

In the earlier discussion, it was noted that by transforming the observations in an NHPP with known rate function so that the times-to-events become $T_1' = \Lambda(T_1)$, $T_2' = \Lambda(T_2), \cdots$, the transformed process is a homogeneous Poisson process with unit rate function. Moreover, by conditioning on the number of events in $(0, t_0]$ or $(0, \Lambda(t_0)]$, the problem of testing for an NHPP can be reduced to testing, for some alternatives, that the times-to-events are order statistics from a uniform distribution. Other tests are given in Cox and Lewis [5], Ch. 6. The transformation is shown in Fig. 11.

Testing for an NHPP with unknown rate function is more difficult. The analogous problem in regression analysis is to test the usual assumption that the residuals ε_i in an additive model

$$Y_i = g(i; \underline{\beta}) + \varepsilon_i$$

are independent normal random variables with mean zero and constant variance σ^2. The problem is that after estimating the parametric mean value function, the residuals $\varepsilon_i = Y_i - g(i; \underline{\hat{\beta}})$ are no longer independent and normally distributed. (e.g., see Daniel and Wood [16]).

An analogous procedure suggested by Lewis [21], using Theorem 1, is to estimate the parameters in the parametric rate function $\Lambda(t; \underline{\alpha})$, which we denote by $\hat{\Lambda}(t; \hat{\underline{\alpha}})$ or $\hat{\Lambda}(t)$, via maximum likelihood and then to detrend the process by transforming the process to obtain $T_1' = \hat{\Lambda}(T_1; \hat{\underline{\alpha}})$, $T_2' = \hat{\Lambda}(T_2; \hat{\underline{\alpha}}), \cdots$ We would expect the departures from a homogeneous process to be small if the number of observations is large and the number of parameters small, and, of course, if the completely specified NHPP is correct.

Very little is known about this procedure. Note, however, that if the uniform conditional test is used with (conditional) Kolmogorov-Smirnov statistics, the problem is that of Kolmogorov-Smirnov tests of fit after parameter estimation. Lilliefors [22, 23] has investigated this for exponential and normal random variables; as expected, the estimated distribution function (integrated rate function) is, on average, closer to the empirical distribution function (empirical integrated rate function) than without parameter estimation. More recent work on Kolmogorov-Smirnov tests with estimated parameters is not yet developed for our purposes. Tests for a homogeneous Poisson process based on spectra (Cox and Lewis [5], Ch. 6) should be less sensitive to parameter estimation.

We now apply these methods to the low- and high-activity periods in an informal manner, relying more on properties of the intervals and the count spectra than on the rate function.

• *High-activity data: test for NHPP*

The following discussion of the validity of, or departures from, the NHPP model for the high-activity data is based, after transformation of the data, on the methodology in Cox and Lewis [5], which is implemented in the SASE-IV program. It is highly technical; our discussion is abbreviated and can be skipped by the reader interested primarily in the results of the data analysis. Briefly, the NHPP is found using the detrending technique to be approximately correct. Deviations occur because of an apparent inhibition effect that results in fewer very short intervals than would occur under the NHPP assumption.

To proceed with the analysis of the detrended high-activity data, in Table 10 we given results of several tests for dependence of intervals in the process. The normalized, estimated first serial correlation coefficient $(n-1)^{\frac{1}{2}}\hat{\rho}_1$ has a value -2.5532, higher than the 1% level of the normal distribution, while the tests for independence based on the cumulated periodogram (raw interval spectral density estimate) using the Kolmogorov-Smirnov statistic $D_{n/2}$ and the Anderson-Darling statistic $W^2_{n/2}$ (Cox and Lewis [5], Ch. 6) are just significant at a 1% level.

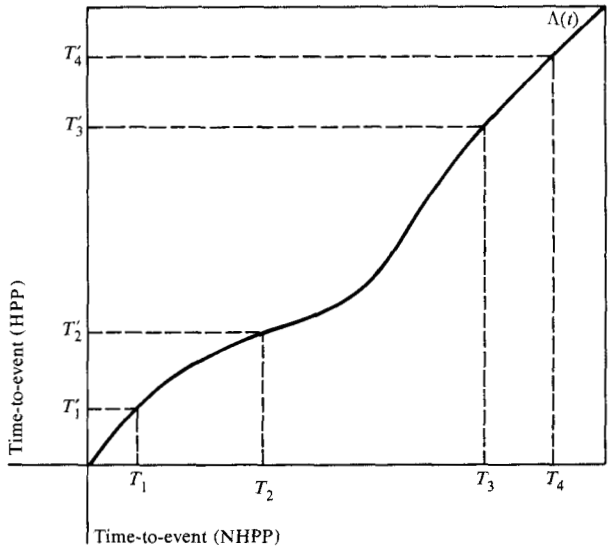

Figure 11 Transformation of the time scale for NHPP having integrated rate function $\Lambda(t)$. The T_i are the times-to-events in the NHPP; the T_i' are the times-to-events in a homogeneous Poisson process of rate 1.

Table 10 Tests for dependence on serial number and dependence between intervals. Detrended (NHPP exponential polynomial rate function of degree 6) transaction initiation process for time period H.

n	number of transactions initiated	1999
$\hat{\rho}_1$	estimated serial correlation coefficient of lag 1 for times between transaction initiations	−0.05762
$(n-1)^{\frac{1}{2}}\hat{\rho}_1$		−2.5532
	Tests for serial independence based on cumulated periodogram	
$D_{n/2}$	Kolmogorov-Smirnov statistic*	1.4897
$W^2_{n/2}$	Anderson-Darling statistic†	3.9941

*Upper 1% point is 1.518
†Upper 1% point is 3.857

We note that the smoothed interval spectral density, as computed in the SASE-VI program, shows no characteristic departure from flatness, and serial correlations beyond the first are small. Thus there appears to be only a residual dependence in the intervals, possibly due to the detrending or a residual trend.

Similarly, the estimated spectrum of counts (Cox and Lewis [5], Ch. 5; Lewis [21]) has no significant departure from flatness, showing that a Poisson process is a tenable hypothesis for the detrended data and consequently a NHPP hypothesis for the original data.

However, some very subtle departures from exponentiality appear when we look at the interval properties of the detrended process. These are given in Table 11. In the first place, the estimated coefficient of variation

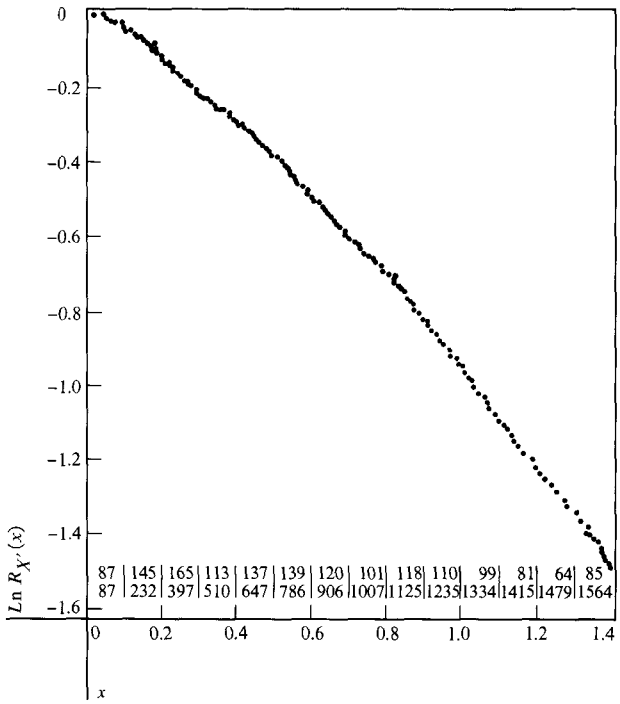

Figure 12 Empirical log survivor function of the detrended times X' between transaction initiations for time period H. Sample size $n = 1999$. Note that the figure shows the log survivor function out only to 1564 events.

of times between events, $\hat{C}(X')$ is smaller than 1. Estimated from five sections of the data, it has value $\hat{C}(X') = 0.9673$, with estimated standard deviation 0.0775, which is too large to give conclusive evidence of departure from the value $C(X') = 1$ for a Poisson process. The empirical log survivor function of the detrended times X' between transaction initiations is shown in Fig. 12.

This artifact of the data shows up clearly in an estimate of the intensity function, $m_f(t)$. There is a definite notch at zero in the estimate $\tilde{m}_f(\Delta t)$ (Cox and Lewis [5], Ch. 5). Thus there are only 720 observations within Δ of the origin, and subsequently the estimate is essentially flat, never deviating in any interval Δ from the modal value of 1,000 by more than 50.

Checking of the transaction initiation process showed that there was in fact a minimum time between transaction initiations imposed by the system. A simple model of a Poisson process with blocking (Type I counter) is sufficient to account for the deviations from a Poisson process.

Another artifact in the data appears in the fact that the estimated coefficients of skewness and kurtosis, $\hat{\gamma}_1(X')$ and $\hat{\gamma}_2(X')$ for the data (5.2363 and 68.3916 in Table 11) are large compared to the Poisson process values $\gamma_1(X) = 2$, $\gamma_2(X) = 9$. These are due to occasional very large times between transaction initiations; these seem to occur in very short periods of high variability of times between transaction initiations. This shows up in Fig. 5 as the spike at about $t = 3000$.

No explanation has been found for this departure from the NHPP; it could be due to special procedures in the use of the system but in any event is too minor to affect practical use of the NHPP model in evaluating such a system.

• *Low-activity data: test for NHPP*

The low-activity data, after detrending with an estimated rate function $\hat{\Lambda}(t; \underline{\hat{\alpha}})$ which is the integral of an exponential polynomial of degree 8, to give $T_1' = \hat{\Lambda}(T_1)$, $T_2' = \hat{\Lambda}(T_2), \cdots$, show a very definite indication of departure from a Poisson process. For $\hat{C}(X')$, $\hat{\gamma}_1(X')$, $\hat{\gamma}_2(X')$, we obtain values 1.475, 4.1233, 21.716, respectively, and these are too large to be consistent with a Poisson hypothesis after detrending.

The data also show considerable interval correlation. A detailed analysis will not be given here, especially since the detrending process is not completely valid. However, as remarked earlier, the low-activity data after detrending is consistent with a cluster process hypothesis. We emphasize that "consistent" here refers only to matching of gross characteristics of the observed and theoretical processes; there is no known formal way of verifying a non-homogeneous cluster process hypothesis.

Discussion

The outstanding feature of these data is the oscillatory nature of the rate function in both the high and low activity periods. Such oscillatory behavior is usually investigated by spectral analysis, but this of course is applicable only to stationary data. The data show a gross time-of-day effect superimposed on the oscillations, and it is not simple to filter this out, most particularly because the period of the oscillation is long, i.e., low-frequency. It is therefore likely to become mixed up in a spectral analysis with long term evolutionary (time-of-day) trends.

Nevertheless, an attempt was made to examine the cyclic effect in time periods H and L by

1. detrending after fitting an exponential polynomial of degree 1; and
2. computing the count spectrum of the detrended data using SASE-VI.

The results of these spectral analyses show generally flat spectra, with peaks at a low frequency corresponding to a rough guess at the frequency of the cycle, which was obtained from Figs. 4 and 9. There seems to be no evidence of a fixed frequency cycle; this would show up as a sharp peak in the spectrum.

The cycles observed in this exploratory analysis of a single series of events in the system bring up some interesting, difficult, and as yet, unresolved methodological and phenomenological questions.

1. The global techniques for rate function estimation need to be extended to larger sections of data as the best overall way of looking at these data. The most practical way of doing this would appear to be to apply the technique to non-overlapping or overlapping sections of the data. The problem of joining sections might lead to (exponential) spline function techniques; new problems of testing then arise.
2. The question arises as to what causes the oscillatory or cyclic effect. In the Introduction we pointed out that the transaction initiation process is an output or response process so that it is presumably driven by other processes associated with the system (e.g., message arrivals). The implications of this from a methodological point of view are twofold:
   a. The deterministic rate function estimated in previous sections might be considered, at least in the micro-aspects, to be purely descriptive. There is a possibility that what we are seeing is the effect of congestion in the system (e.g., DL/I component), and the data may perhaps be best described by something like a self-exciting process (Hawkes [24]), which is the point process analog of an autoregressive system. This would not be inconsistent with our findings, since (linear) self-exciting processes are special types of cluster processes (Hawkes and Oakes [25]). One problem with the above interpretation of the cyclic effect is that we would expect more oscillatory effect during high activity periods than during low activity periods. However, just the opposite is true.
   b. Since the observed transaction initiation process is driven by other processes associated with the system, a full description of the behavior of the system would involve an attempt to correlate the transaction initiation process studied in this paper with processes at other points of the system. In particular, it would be of interest to correlate the transaction initiation process with the process of message arrivals from terminals. It would also be desirable to correlate the transaction initiation process with the successive response times experienced by users of the system.

There are many methodological problems in analyzing very non-stationary systems, in particular the problem of estimating correlation and/or coherence. For the present case the fact that the high-activity data are close to Poisson, although nonhomogeneous, should make development of the necessary methodology simpler. The work of Cox and Lewis [5], and particularly Cox [10], should be useful.

Table 11 Sample characteristics of times-between-events. Detrended (NHPP exponential polynomial rate function of degree 6). Transaction initiation process for time period H.

n	number of transactions initiated	1999
t_0	period of observation	1999.02
$\overline{X'}$	estimated mean time between transaction initiations	0.9998
$\hat{C}(X')$	estimated coefficient of variation of times between transaction initiations	0.9784
$\hat{\gamma}_1(X')$	estimated coefficient of skewness of times between transaction initiations	5.2363
$\hat{\gamma}_2(X')$	estimated coefficient of kurtosis of times between transaction initiations	68.3916
$X'_{\max}$	maximum time between transaction initiations	17.4752
$X'_{\min}$	minimum time between transaction initiations	0.0031

References

1. P. A. W. Lewis and G. S. Shedler, "Empirically Derived Micromodels for Sequences of Page Exceptions", *IBM J. Res. Develop.* **17**, 86-100 (1973).
2. S. S. Lavenberg and G. S. Shedler, "Stochastic Modeling of Processor Scheduling with Application to Data Base Management Systems," *IBM J. Res. Develop.* **20**, 437 (1976, this issue).
3. "Information Management System/360, Version 2. General Information Manual," GH 20-0765, IBM Corporation, Armonk, New York, 1973.
4. P. A. W. Lewis, A. M. Katcher, and A. H. Weis, "SASE-IV: An Improved Program for the Statistical Analysis of Series of Events," *Research Report RC-2365*, IBM Thomas J. Watson Research Center, Yorktown Heights, New York, 1969.
5. D. R. Cox and P. A. W. Lewis, *The Statistical Analysis of Series of Events*, Methuen, London and John Wiley and Sons, Inc., New York, 1966.
6. E. Çinlar, "Superposition of Point Processes," in *Stochastic Point Processes*, edited by P. A. W. Lewis, John Wiley and Sons, Inc., New York, 1972, pp. 549-606.
7. P. A. W. Lewis, "Non-homogeneous Branching Poisson Processes," *J. Royal Statist. Soc. B*, **29**, 343-354 (1967).
8. M. Rosenblatt, "Remarks on Some Non-Parametric Estimates of a Density Function", *Ann. Math. Stat.* **27**, 3, 832-837 (1956).
9. P. A. W. Lewis, "Recent Results in the Statistical Analysis of Unvariate Point Processes," in *Stochastic Point Processes*, edited by P. A. W. Lewis, John Wiley and Sons, Inc., New York, 1972, pp. 1-54.
10. D. R. Cox, "The Statistical Analysis of Dependencies in Point Processes," in *Stochastic Point Processes*, edited by P. A. W. Lewis, John Wiley and Sons, Inc., New York 1972, pp. 55-66.
11. M. Brown, "Statistical Analysis of Non-Homogeneous Poisson Processes," in *Stochastic Point Processes*, edited by P. A. W. Lewis, John Wiley and Sons, Inc., New York, 1972, pp. 67-89.

12. B. V. Gnedenko and I. Kovalenko, *Introduction to Queuing Theory*, tr. by D. Louvish, Daniel Davey and Co., Hartford, Conn., 1969.
13. E. Çinlar, *Introduction to Stochastic Processes*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1975.
14. P. A. W. Lewis, "Some Results on Tests for Poisson Processes," *Biometrika* **52,** 67-77 (1965).
15. C. J. MacLean, "Estimation and Testing of an Exponential Polynomial Rate Function Within the Non-Stationary Poisson Process," *Biometrika* **61,** 81-86 (1974).
16. C. Daniel and F. S. Wood, *Fitting Equations to Data: Computer Analysis of Multifactor Data for Scientists and Engineers*, Wiley-Interscience Publishers, Inc., New York, 1971.
17. T. W. Anderson, *Statistical Analysis of Time Series*, John Wiley and Sons, Inc., New York, 1971.
18. P. J. Bickel and M. Rosenblatt, "On Some Global Measures of the Deviations of Density Function Estimates," *Ann. Math. Stat.* **44,** 1071-1075 (1973).
19. P. A. W. Lewis, L. H. Liu, D. W. Robinson, and M. Rosenblatt "Empirical Sampling Study of a Goodness of Fit Statistic for Density Function Estimation," *Naval Postgraduate School Report NPS55Lw75031*, Monterey, California, 1975.
20. D. Vere-Jones, "Stochastic Models for Earthquake Occurrence," *J. Royal Statist. Soc. B*, **32,** 1-62 (1970).
21. P. A. W. Lewis, "Remarks on the Theory, Computation and Application of the Spectral Analysis of Series of Events," *J. Sound Vib.* **12,** 353-75 (1970).
22. H. W. Lilliefors, "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown," *J. Amer. Statist. Assoc.* **62,** 399-402 (1967).
23. H. W. Lilliefors, "On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown," *J. Amer. Statist. Assoc.* **64,** 387-389 (1969).
24. A. G. Hawkes, "Mutually Exciting Point Processes," in *Stochastic Point Processes*, edited by P. A. W. Lewis, John Wiley and Sons, Inc., New York, 1972, 261-271.
25. A. G. Hawkes and D. Oakes, "A Cluster Process Representation of a Self-Exciting Process," *J. Appl. Prob.* **11,** 493-504 (1974).

Received January 19, 1976

Dr. Lewis is located at the Naval Postgraduate School, Monterey, CA 93940. Dr. Shedler is located at the IBM Research Laboratory, 5600 Cottle Road, San Jose, CA 95193.