Nonparametric Estimation of the Cumulative Intensity Function for a Nonhomogeneous Poisson Process from Overlapping Realizations

Bradford L. Arkin · Lawrence M. Leemis

Reliable Software Technologies, Sterling, Virginia 20166

Department of Mathematics, The College of William & Mary,

Williamsburg, Virginia 23187

barkin@rstcorp.com · leemis@math.wm.edu

A nonparametric technique for estimating the cumulative intensity function of a nonhomogeneous Poisson process from one or more realizations on an interval is extended here to include realizations that overlap. This technique does not require any arbitrary parameters from the modeler, and the estimated cumulative intensity function can be used to generate a point process for simulation by inversion.

(*Input Modeling; Nonstationary Poisson Process; Repairable Systems; Simulation; Time-Dependent Arrivals; Variate Generation*)

---

1 Introduction

A nonhomogeneous Poisson process (NHPP) is often used as a model for systems whose rate varies with time. This paper describes a nonparametric technique for estimating the cumulative intensity function of a NHPP on the time interval $(0, S]$ from one or more overlapping realizations. If the NHPP is used as an input to a Monte Carlo or discrete-event simulation, it is possible to use inversion to generate event times. The estimation technique may be applied to any sequence of events occurring over time or space, such as arrival times to a queue, failure times of a repairable system, earthquake times, or pothole positions on a highway.

A NHPP is a generalization of a homogeneous Poisson process where events occur randomly over time at an average rate of λ events per unit time. The rate at which events occur in a NHPP varies with time as determined by the *intensity function,* $\lambda(t)$, which is an integrable function of time. The *cumulative intensity function* is defined by

$$\Lambda(t) = \int_0^t \lambda(\tau) d\tau, \quad t > 0,$$

and is interpreted as the expected number of events by time t. The probability of exactly n events occurring in the interval $(a, b]$ is given by

$$\frac{\left[\int_a^b \lambda(t)dt\right]^n e^{-\int_a^b \lambda(t)dt}}{n!}$$

for $n = 0, 1, \ldots$ (Cinlar 1975).

2 Estimation Procedure

The intensity function, $\lambda(t)$, for a NHPP is assumed to be nonnegative for all $t \in (0, S]$. The cumulative intensity function is to be estimated from realizations of the NHPP on any interval $(a, b]$, where $0 \leq a < b \leq S$, and S is a known constant. The interval $(0, S]$ may represent the time a system allows arrivals (e.g., 9 AM to 5 PM at an office) or one period of a cycle (e.g., one 24-hour period at a convenience store). The estimation procedure described in this section is nonparametric and does not require any arbitrary decisions (e.g., parameter values) from the modeler.

Let the interval $(0, S]$ be partitioned into the fewest possible number of regions r such that within each region $(s_j, s_{j+1}]$ the number of realizations k_{j+1} is constant throughout. Thus, for any time $t \in (s_j, s_{j+1}]$ the number of realizations at time t is equal to $k_{j+1}, j = 0, 1, \ldots, r-1$. Note that from this partitioning of $(0, S]$ into r regions it follows that $s_0 = 0$ and $s_r = S$. Let n_{j+1} be the total number of observed events (across all realizations) during the time span $(s_j, s_{j+1}]$. Define $n = \sum_{q=1}^{r} n_q$. It is assumed that the regions $(s_j, s_{j+1}]$ form a partition of $(0, S]$, and that $k_{j+1} > 0$ as well as $n_{j+1} \geq 0$ for $j = 0, 1, \ldots, r-1$.

Let $t_{(0)}, t_{(1)}, \ldots, t_{(n+r)}$ be the order statistics of the superposition of the realizations as well as the region boundaries $s_0, s_1, \ldots, s_r$. Note that it follows that $t_{(0)} = s_0 = 0$, and $t_{(n+r)} = s_r = S$. While the region boundaries $s_0, s_1, \ldots, s_r$ are included in the order statistics $t_{(0)}, t_{(1)}, \ldots, t_{(n+r)}$, they are not counted as events when constructing $\hat{\Lambda}(t)$; thus the values $n_1, n_2, \ldots, n_r$ reflect only the number of observations in each region.

Figure 1 shows how the data collected from overlapping realizations relates to the defined notation.

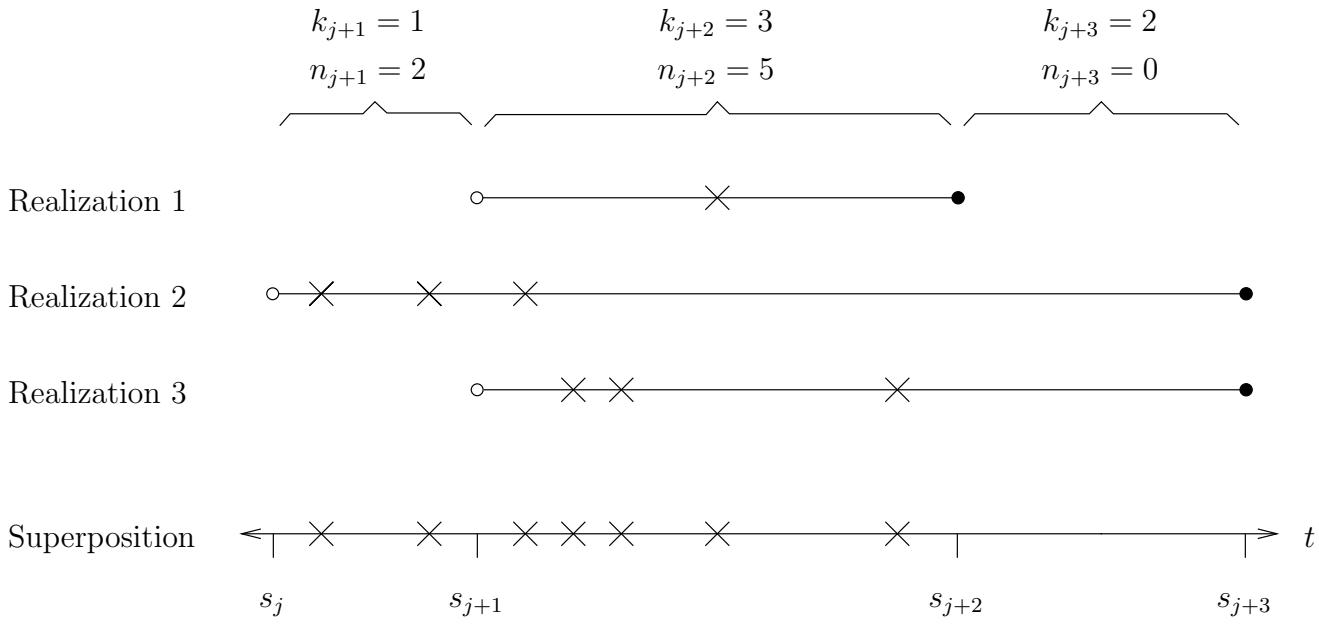

Figure 1: Three Sample Regions Associated with Three Realizations

Setting $\hat{\Lambda}(s_{j+1}) = \sum_{q=1}^{j+1} n_q/k_q$ yields a process where the estimated expected number of events by time s_{j+1} is the sum of the average number of events in the first $j+1$ regions, since $\Lambda(s_{j+1})$ is the expected number of events by time $s_{j+1}, j = 0, 1, 2, \ldots, r-1$. In a similar fashion to the approach in Leemis (1991), the piecewise-linear estimator of the cumulative intensity function between the time values in the superposition is

$$\hat{\Lambda}(t) = \sum_{q=1}^{j} \frac{n_q}{k_q} + \frac{\left(i - \sum_{q=1}^{j} (n_q + 1)\right) n_{j+1}}{(n_{j+1} + 1)\, k_{j+1}} + \left[\frac{n_{j+1}\left(t - t_{(i)}\right)}{(n_{j+1} + 1)\, k_{j+1}\left(t_{(i+1)} - t_{(i)}\right)}\right],$$

$$t_{(i)} < t \le t_{(i+1)}; \qquad i = 0, 1, 2, \ldots, n + r - 1,$$

$$s_j < t \le s_{j+1}; \qquad j = 0, 1, \ldots, r - 1,$$

where the subscript $j + 1$ determines the active region and the subscript $i + 1$ determines

the active ordered observations.

Figure 2 shows a single segment of $\hat{\Lambda}(t)$ between $t_{(i)}$ and $t_{(i+1)}$ in the $(j+1)$st region,

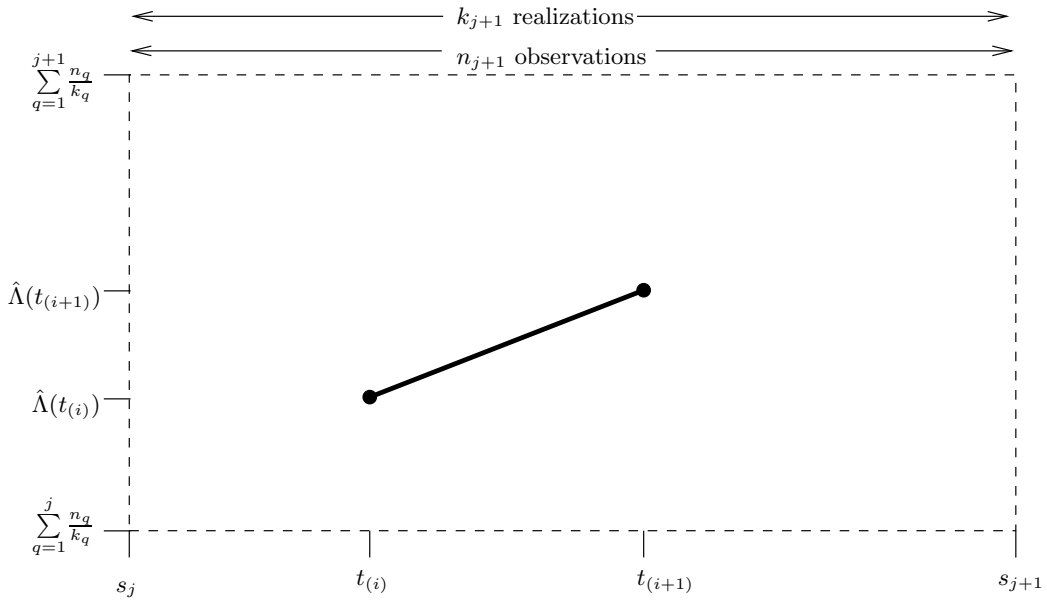

Figure 2: A single segment of $\hat{\Lambda}(t)$ in the $(j+1)$st region

where time runs horizontally and $\Lambda(t)$ runs vertically. Over the course of k_{j+1} realizations, there were n_{j+1} observations during the time interval $(s_j, s_{j+1}]$.

Some empirical evidence supporting the proposed extension of the estimator is provided in Figure 3, where the population cumulative intensity function (smooth curve in dashed line) and the proposed extended estimator (piecewise-linear function in solid line) are plotted for several realizations of customers arriving to a lunchwagon. The realizations are generated by thinning. The parent functions are the piecewise-linear intensity function

$$\lambda(t) = \begin{cases} 10t + 1, & 0 < t \leq 1.5, \\ 16, & 1.5 < t \leq 2.5, \\ -6t + 31, & 2.5 < t \leq 4.5, \end{cases}$$

and the cumulative intensity function

$$\Lambda(t) = \begin{cases} 5t^2 + t, & 0 < t \leq 1.5, \\ 16t - 11.25, & 1.5 < t \leq 2.5, \\ -3t^2 + 31t - 30, & 2.5 < t \leq 4.5, \end{cases}$$

from Klein and Roberts (1984). This intensity function models arrivals to a lunchwagon between 10:00 AM and 2:30 PM. The estimator plotted in Figure 3 was generated over the time span (0, 4.5], and consists of the $r = 3$ partitions (0, 1.5], (1.5, 3], and (3, 4.5]. The first and last regions, $(0, s_1]$ and $(s_2, s_3]$, utilize only $k_1 = k_3 = 1$ realization, while the middle region, $(s_1, s_2]$, utilizes $k_2 = 12$ realizations. Thus Figure 3 portrays a (1, 12, 1) *instance*. Note that the sampling variability on the middle time interval (1.5, 3] is smaller than on the other two intervals. These regions can be seen in Figure 3, where rectangles with upper

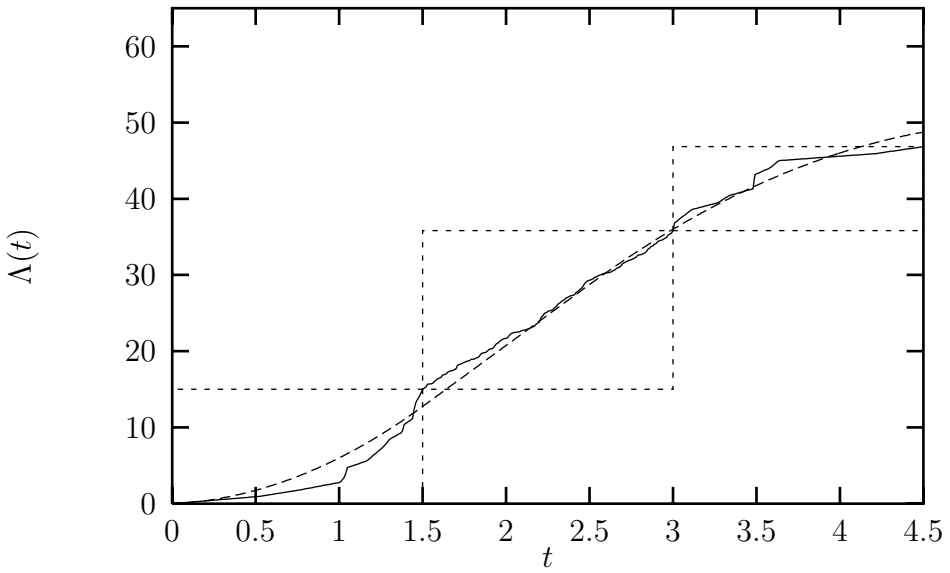

Figure 3: Depiction of the Three Regions

right-hand corner coordinates $(s_{j+1}, \hat{\Lambda}(s_{j+1})), j = 0, 1, 2$ have been added. Each rectangle in

Figure 3 represents a NHPP modeling solution equivalent to the one proposed by Leemis (1991).

An asymptotically exact $100(1-\alpha)\%$ confidence interval for $\Lambda(t)$ is

$$\left|\Lambda(t) - \hat{\Lambda}(t)\right| < z_{\alpha/2}\sqrt{\frac{\hat{\Lambda}(t) - \hat{\Lambda}(s_j)}{k_{j+1}} + \sum_{q=1}^{j} \frac{\hat{\Lambda}(s_q) - \hat{\Lambda}(s_{q-1})}{k_q}},$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ fractile of the standard normal distribution, and $t \in (s_j, s_{j+1}]$. This interval is a generalization of an interval derived in the appendix of Leemis (1991). The performance of this confidence interval is evaluated in Section 4. Figure 4 shows $\Lambda(t), \hat{\Lambda}(t)$

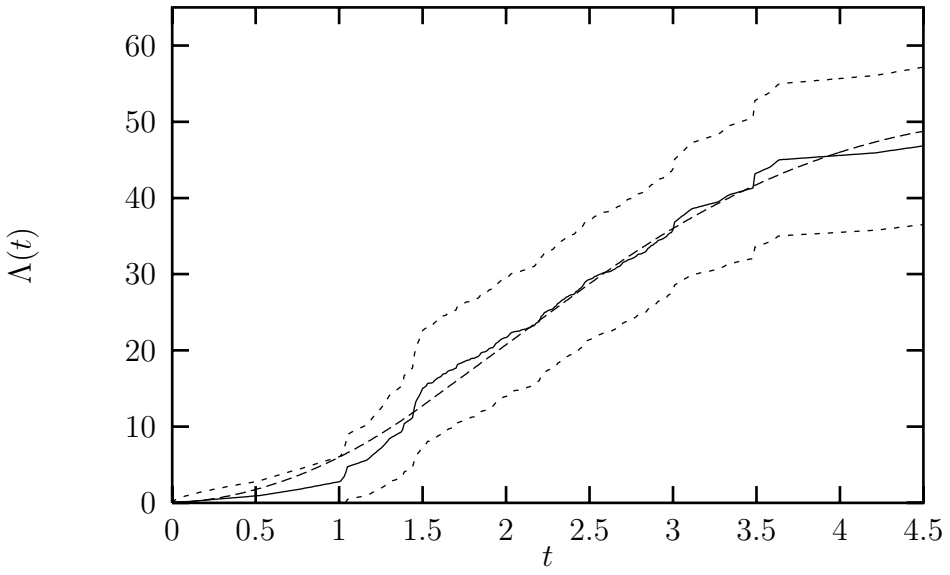

Figure 4: Parent Cumulative Intensity Function, Nonparametric Estimator, and 95% Confidence Bands

and 95% confidence bands for the lunchwagon example. Note that the bands spread more rapidly on the first and third regions, where only one realization has been observed.

3 Variate Generation

The cumulative intensity function for a NHPP is often estimated to generate variates for simulation. Using a time transformation (Cinlar 1975), the event times from a unit Poisson process, $E_1, E_2, \ldots$, can be transformed to the event times of a NHPP via $T_i = \Lambda^{-1}(E_i)$. For the NHPP estimate considered here, the events at times $T_1, T_2, \ldots$ on $(0, S]$ can be generated for Monte Carlo simulation by the algorithm below, given the superpositioned values $t_{(0)}, t_{(1)}, \ldots, t_{(n+r)}$, as well as $r, n_1, n_2, \ldots, n_r, k_1, k_2, \ldots, k_r$.

$i \leftarrow 1$ [initialize variate counter]

$j \leftarrow 0$ [initialize region counter]

MAX $\leftarrow \sum_{q=1}^{r} n_q/k_q$ [set MAX to $\hat{\Lambda}(S)$]

generate $U_i \sim U(0,1)$ [generate initial random number U_1]

$E_i \leftarrow -\log_e(1 - U_i)$ [generate initial unit exponential variate E_1]

while $E_i <$ MAX **do**

begin

while $E_i > \sum_{q=1}^{j+1} n_q/k_q$ **do** [update region counter if necessary]

begin

$j \leftarrow j + 1$ [increment region counter]

end

$m \leftarrow \left\lfloor \frac{(n_{j+1}+1)k_{j+1}\left(E_i - \sum_{q=1}^{j} n_q/k_q\right)}{n_{j+1}} \right\rfloor + \sum_{q=1}^{j} (n_q + 1)$

[set $m \ni \hat{\Lambda}(t_{(m)}) < E_i \leq \hat{\Lambda}(t_{(m+1)})$]

$T_i \leftarrow t_{(m)} + [t_{(m+1)} - t_{(m)}] \left(\frac{(n_{j+1}+1)k_{j+1}\left(E_i - \sum_{q=1}^{j} n_q/k_q\right)}{n_{j+1}} - \left(m - \sum_{q=1}^{j} (n_q+1)\right)\right)$

[generate event time]

$i \leftarrow i+1$ [increment variate counter]

generate $U_i \sim U(0,1)$ [generate next random number]

$E_i \leftarrow E_{i-1} - \log_e(1-U_i)$ [generate next HPP event time]

end

The algorithm above assumes the existence of a random number generator capable of producing the independent $U(0,1)$ variates $U_1, U_2, \ldots, U_i$. Thus, it is a straightforward procedure to obtain a realization of $i-1$ events on $(0, S]$ from the superpositioned process and $U(0,1)$ values $U_1, U_2, \ldots, U_i$. Inversion has been used to generate this NHPP, so certain variance reduction techniques, such as antithetic variates or common random numbers, may be applied. Replacing $1-U_i$ with U_i in the algorithm will save CPU time although the direction of the monotonicity is reversed. As $\sum_{q=1}^{r} n_q$ increases, the amount of memory required to store $t_{(0)}, t_{(1)}, \ldots, t_{(n+r)}$ increases, but the amount of CPU time required to generate a realization depends only on the sum of ratios $\sum_{q=1}^{r} n_q/k_q$, or equivalently, the sum of the average number of events per region. Thus, collecting more realizations (resulting in narrower confidence intervals) increases the amount of memory required, but does not impact the expected CPU time for generating a realization.

Due to a discrete measure of time or rounding, tied observations, i.e., $t_{(i)} = t_{(i+1)}$, occasionally occur in practice. If $t_{(m)} = t_{(m+1)}$ for some m, then a point of disconti-

nuity is introduced to $\hat{\Lambda}(t)$ at $t_{(m)}$. An example of this discontinuity is illustrated in Figure 5. These tied values do not pose a problem to the variate generation algorithm.

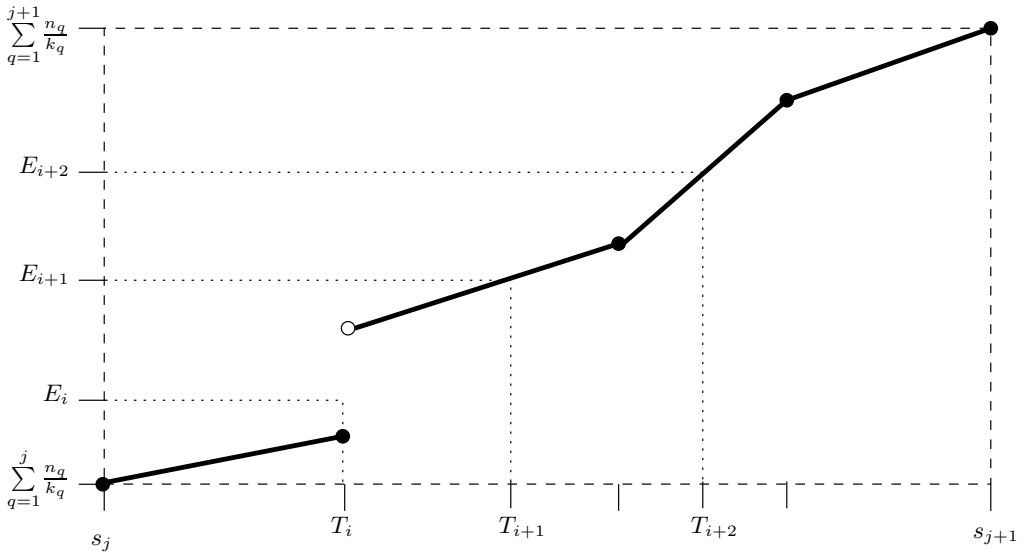

Figure 5: The Variate Generation Algorithm

If $t_{(m)} = t_{(m+1)}$ for some m, and a unit exponential variate E_i is generated such that $\hat{\Lambda}(t_{(m)}) = \hat{\Lambda}(t_{(m+1)}) \leq E_i \leq \lim_{t \downarrow t_{(m+1)}} \hat{\Lambda}(t)$, then the event time T_i corresponding to E_i is set equal to the time of the tied observation as follows:

$$\hat{\Lambda}^{-1}(E_i) = t_{(m)} = t_{(m+1)} = T_i,$$

as illustrated in Figure 5.

4 Examples

Three examples will be given in this section. The first revisits the lunchwagon model, the second contains failure times of a repairable system, and the third uses failure times from a nonrepairable system.

The functions $\lambda(t)$ and $\Lambda(t)$ given in Section 2 model arrivals to a lunchwagon during the midday rush hour time interval (10:00, 2:30]. These functions were used in a Monte Carlo experiment to test the accuracy of the confidence interval provided above. By creating 100,000 (1, 12, 1) instances with region boundaries {0, 1.5, 3, 4.5}, the same as the $\hat{\Lambda}(t)$ illustrated in Figures 3 and 4, we were able to measure the actual coverage of the following points at a nominal coverage of 0.95. Table 1 lists for each of the eight time values the actual coverage,

Time	Actual Coverage	Misses High	Misses Low
0.90	0.9501	0.0013	0.0487
1.35	0.9386	0.0048	0.0566
1.80	0.9505	0.0200	0.0296
2.25	0.9466	0.0196	0.0339
2.70	0.9498	0.0174	0.0329
3.15	0.9509	0.0295	0.0196
3.60	0.9498	0.0251	0.0251
4.05	0.9517	0.0167	0.0316

Table 1: Coverages in the Lunchwagon Example (Nominal Coverage 0.95)

and the number of high and low misses. This experiment indicates that the approximate confidence intervals for the cumulative intensity function estimate are fairly accurate for a large sample size n. This is not surprising since the Poisson distribution converges to a normal distribution as its mean increases. There appears to be some nonsymmetry in the proportion of intervals missing high versus those missing low for small time values.

The second example examines the data set consisting of failure times for a group of 20 copy machines (Zaino and Berke 1992). For these machines, time is measured by the number of *actuations*, i.e., the number of copies made, and the time at installation is defined to be 0. This data set (adjusted for staggered installation times) is displayed in Table 2, a plot of the failure times on (0, 75000] is given in Figure 6, and a plot of the individual cumulative

Machine #	First	Second	Third	Fourth	Fifth	Sixth	Seventh	Eighth
125	2774	6963	8954	9201	9507	10074	10278	10830
126	11070	14300	16160	20900	23029	27091	58472	72716
127	67827							
128	1440	10776	11016	15198	34392			
129	6237	6880	7463	11638				
130	1518	18872	40075	43543	54896	62364	65787	66149
131	50	3791	5000	7393	19252	22401	23214	25020
132	2793	6517	6982	13110	34389	43823	70675	
133	3962	7884	8187	10861				
134	415	15924	18616	19616	21235	23935	33709	39235
135	9111	10091	16649	16877	17628			
136	1215	1452	3676	40334	41354	48729	50573	54261
137	4199	5354	20931	26229	29081	30247	38942	41329
138	9449	10695	10895	54840	59661			
139	11286	26149	31149	59601	65184	75756	117510	124314
140	2009	33165	126646	165016	168508	235121	236969	
141	13676	152238	152644					
142	1532	2399	2938	7108	8094	16070	18178	42820
143	102	220	11720	18016	18143	27825	55399	56223
144	974	17664	20994	38118	43811	47320	49973	53654

Table 2: Copy Machine Failure Times (Actuations)

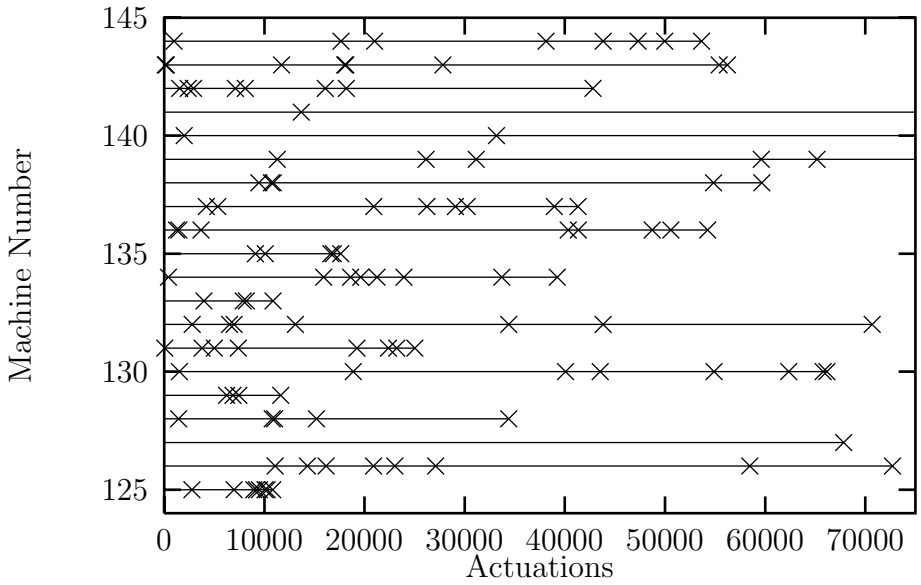

Figure 6: Copier Failure Times and Observation Periods

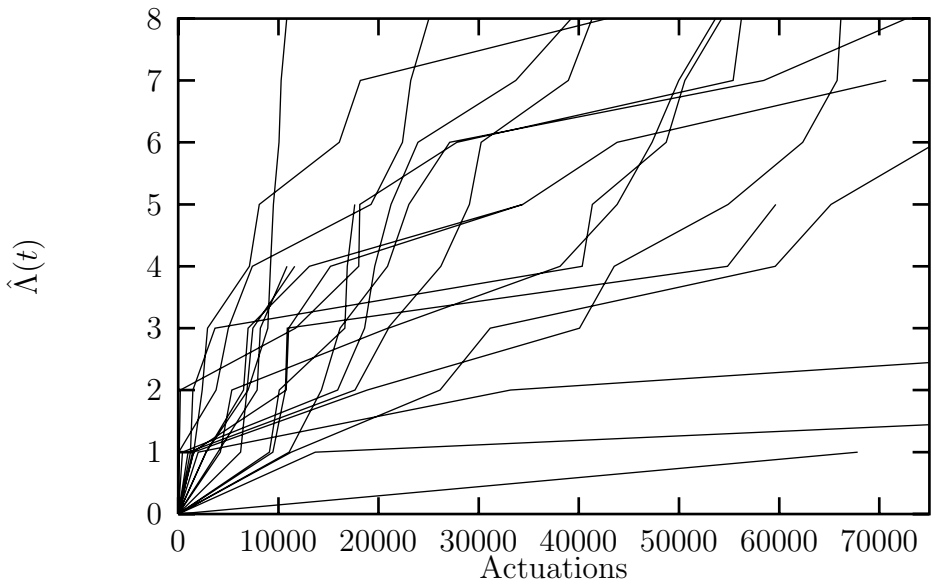

Figure 7: Individual Cumulative Intensity Function Estimators Copying Machines

intensity function estimators for each of the 20 machines is given in Figure 7. The copier failure times constitute a Type II right-censored sample since copiers are removed from the test upon their last observed failure. We consider the performance of the copiers over their first $S = 75,000$ copies here.

This data set consists of 20 realizations on the interval $(0, b]$, where $b \in (0, 75000]$. Since there are three machines (machine numbers 139, 140, 141) with $b = S$ (i.e., machines that are observed over the entire interval $(0, S]$, and the end value b is unique for every other machine), there are $r = 18$ regions such that $k_1 = 20, k_2 = 19, k_3 = 18, \ldots, k_{17} = 4, k_{18} = 3$, and $s_1 = 10,830, s_2 = 10,861, s_3 = 11,638, \ldots, s_{17} = 72,716, s_{18} = 75,000$. The total number of observed failure times n on (0, 75000] is equal to 119.

By using the above information, an estimator for the cumulative intensity function $\Lambda(t)$ of the copiers can be developed. This estimator is plotted along with 95% confidence bands in Figure 8. The fact that the confidence bands do not include the line between (0,0)

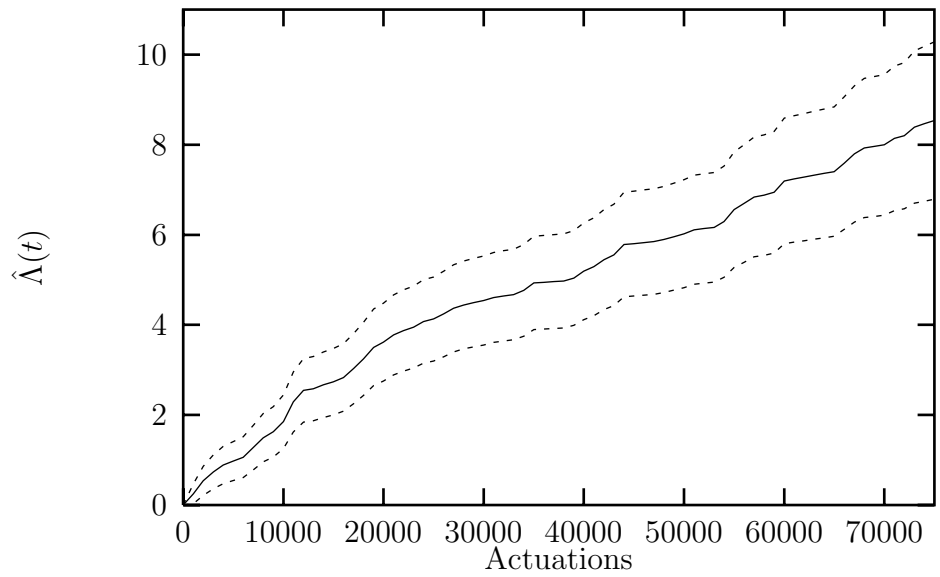

Figure 8: Estimated Cumulative Intensity Function for the Copier Failure Times

and $(S, \hat{\Lambda}(S))$ indicates that the failure rate $\lambda(t)$ does indeed vary, making a NHPP an attractive model. Also, this estimator supports the claim made by Zaino and Berke (1992) that an extended run-in period will decrease the number of early failures after installation. Using the third failure of machine number 143 at actuation 11,720 as the time value in the data set where the slope of the cumulative intensity function changes, the early and later failure rates can be estimated. Since $\hat{\Lambda}(11,720) \cong 2.53$ and $\hat{\Lambda}(75,000) \cong 8.54$, the early estimated failure rate on $0 < t \leq 11,720$ actuations is $\frac{2.53}{11,720} \cong 0.00022$ failures per actuation, and the subsequent estimated failure rate on $11,720 < t \leq 75,000$ actuations is $\frac{8.54-2.53}{75,000-11,720} = 0.000095$ failures per actuation.

The final example studies the data set consisting of failure times of heat pump compressors located in five separate buildings (Nelson 1990), each under repair contract for a time span $(a, b]$. At time a, each building has a set number of heat pump compressors, and no more are added during the time under contract (e.g., $a = 2.59$ and $b = 9.33$ for Building

B as shown in Table 3). The contract periods for the compressors in the five buildings are illustrated in Figure 9.

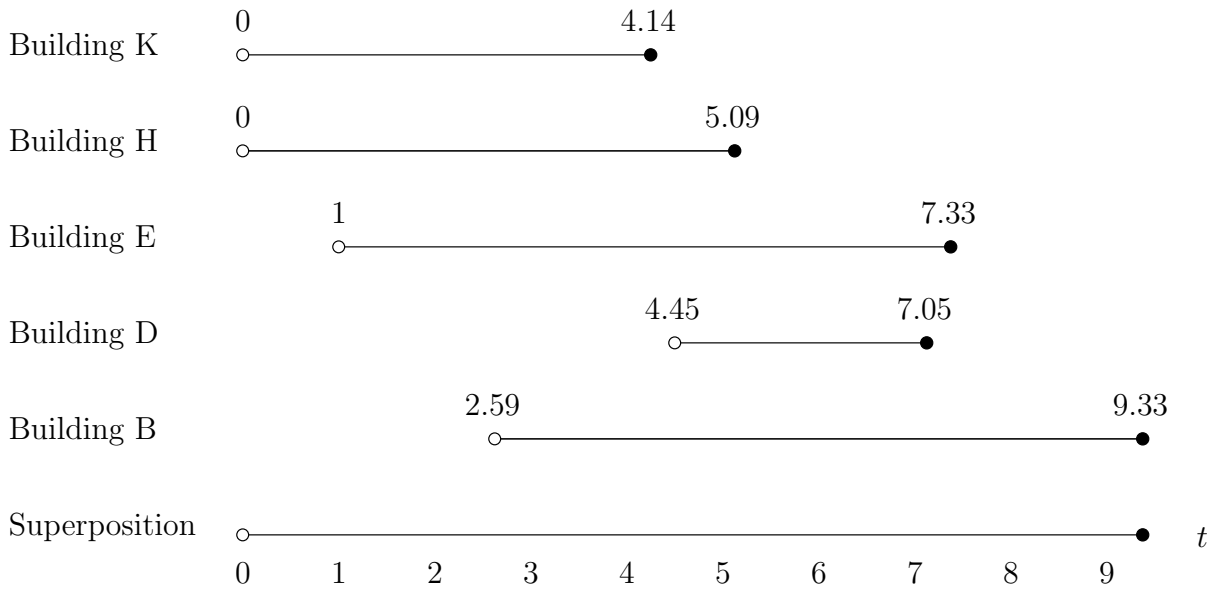

Figure 9: Observation Periods for the Heat Pumps in the Five Buildings

While the building is under repair contract, the failure of a heat pump compressor results in its removal from the building, decrementing the total number of heat pump compressors by one each time. The data set consists of $n = 28$ failure times, and yields $r = 29$ regions.

Building	Number of Compressors	Entry Time a	Compressor Failure Times	Exit Time b
B	164	2.59	3.30, 4.62, 4.62, 5.75, 5.75, 7.42, 7.42, 8.77, 9.27, 9.27	9.33
D	356	4.45	4.47, 4.47, 5.56, 5.57, 5.80, 6.13, 7.02	7.05
E	458	1.00	2.85, 4.65, 4.79, 5.85, 6.73	7.33
H	149	0.00	0.17, 0.17, 1.34	5.09
K	195	0.00	2.17, 3.65, 4.14	4.14

Table 3: Compressor Failure Times (Years)

Some interesting results occur when this data is modeled using the estimator presented in Section 2. Because a heat pump compressor is removed from the building upon its failure,

each region has either 0, 1, or 2 observations. In this example, a region has two observations only when there is a tie, which occurs six times in this data set. Another result of the removal upon failure nature of this data set is that every observation $t_{(m)}$ is equal to some right-hand region boundary $s_{j+1}, j = 0, 1, \ldots, r-1$. The number of realizations k_{j+1} for each region is equal to the number of compressors under contract during the time interval $(s_j, s_{j+1}]$. The values k_{j+1} are much larger than the number of realizations seen in the previous two examples, with values ranging from $k_{29} = 154$ to $k_{11} = 1,122$. The $k_{29} = 154$ value corresponds to 164 less the 10 failed compressors removed from Building B just prior to the end of the study. The $k_{11} = 1,122$ value corresponds to time 4.45^+ (just after the compressors are entered into service on Building D), when there are 149 + 458 + 164 + 356 = 1,127 less the 5 failed compressors at times 0.17, 0.17, 1.34, 2.85, and 3.30.

The estimator created from the data in Table 3 is graphed in Figure 10 along with 95%

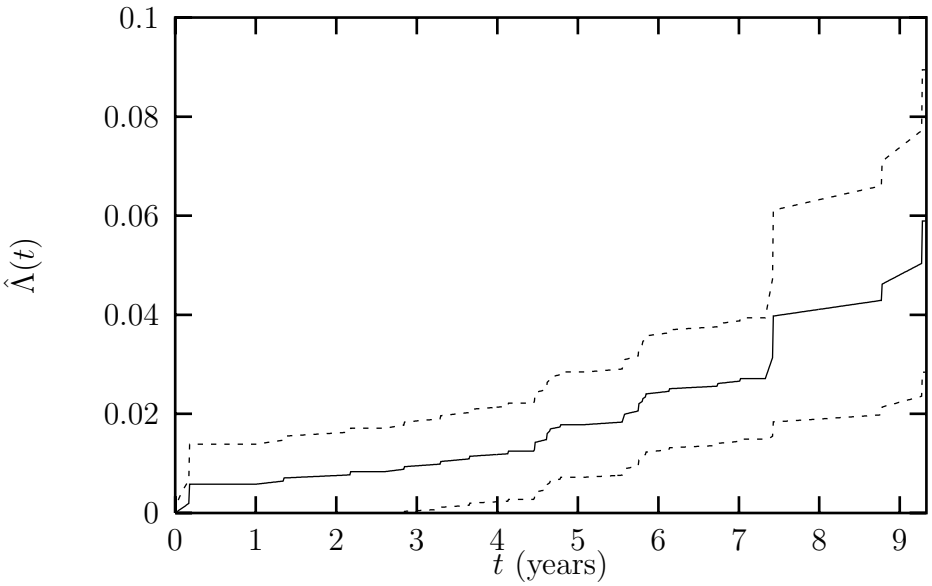

Figure 10: Estimated Cumulative Intensity Function for the Heat Pump Compressor Failure Times

confidence bands. The classic bathtub-shaped failure rate function might possibly be at work

here, as evidenced by the early failures (that suspiciously seem to be limited to Building H) and the late failures that occur after Year 7.

5 Summary

An extension of the estimator given by Leemis (1991) is presented that provides for a nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process using overlapping realizations. Simulation via inversion is straightforward.

As in classical statistics, an estimator derived from a data set containing a small number of observations or realizations should be approached cautiously as the results may be unrepresentative of the system on average, due to sampling variability. Data sets containing clustered observations, as in example three at time 7.42, will result in an estimator that produces simulations with similar characteristics. Increasing the number of realizations increases the precision of $\hat{\Lambda}(t)$. In a similar fashion to many confidence intervals from classical statistics, the confidence interval width is proportional to the inverse of the square root of the number of realizations (e.g., a quadrupling of realizations over all segments results in a halving of the width of the confidence interval).[1]

[1]The authors gratefully acknowledge support from the National Science Foundation for providing funding for an Educational Innovation Grant CDA9712718 "Undergraduate Modeling and Simulation Analysis." The authors also thank the editor and referees for their helpful suggestions.

References

Cinlar, E. 1975. *Introduction to Stochastic Processes.* Prentice–Hall, Englewood Cliffs, NJ.

Klein, R. W., S. D. Roberts. 1984. A time-varying Poisson arrival process generator. *Simulation* **43** (4) 193–195.

Leemis, L. M. 1991. Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process. *Management Sci* **37** (7) 886–900.

Nelson, W. 1990. Hazard plotting of left truncated life data. *J. Quality Tech.* **22** (3) 230–232.

Zaino Jr., N. A., T. M. Berke. 1992. Determining the effectiveness of run-in: a case study in the analysis of repairable-system data. *Proc. Ann. Reliability and Maintainability Sympos.*, IEEE, New York, 58–70.