Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

MISE-OPTIMAL GROUPING OF POINT-PROCESS DATA WITH A CONSTANT DISPERSION RATIO

Huifen Chen

Department of Industrial and Systems Engineering
Chung-Yuan University
Chung-Li, 320 Taoyuan, TAIWAN

Bruce Schmeiser

School of Industrial Engineering
Purdue University
West Lafayette, IN 47907, USA

ABSTRACT

Given a set of point-process event times with constant dispersion ratio, we are interested in estimating the rate function by grouping the event times into count data from equal-width time intervals. We group in order to smooth the resulting piecewise-constant rate function using one of our two existing methods: I-SMOOTH and MNO-PQRS. Using the mean integrated squared error (MISE) for piecewise-constant rate functions, we create two estimators; minimizing the estimated MISE function yields the chosen number of intervals. The MISE function provides insights into the optimal number of intervals as a function of the rate-function shape and expected number of events. Across several examples, our two number-of-intervals estimators perform well and similarly; nevertheless, one dominates in terms of realized MISE value.

1 INTRODUCTION

In modeling dynamic stochastic systems, a fundamental problem is estimating the unknown non-homogeneous rate function from event times. The events are often arrivals; in this paper we use the latter term. The arrival times are known either individually or as count data from known time intervals. Therefore, most relevant research papers consider one of these two contexts. The present paper considers the transition between the two contexts.

1.1 Problem Statement

Given a known interval of time containing observed arrival times $a_1, a_2, \ldots, a_n$, our problem is to determine an appropriate number of intervals k so that the associated count data is a good estimate of the unknown rate function λ^* over the known interval of time. The choice of k is often made intuitively by humans. Our goal is to automate the process. Most humans viewing the automated result should be satisfied; ideally the automated result is often better than what a human would do.

For convenience, scale the arrival times to the unit interval; that is, $0 < a_1, a_2, \ldots, a_n \leq 1$. Here n is the observed value of N, the random number of arrivals in $[0,1]$.

Assume a known constant dispersion ratio, d; that is, for any time interval, d is the ratio of the variance and expected value of the number of arrivals. (A non-homogeneous Poisson process (NHPP) has dispersion ratio $d = 1$.) Gerhardt and Nelson (2009) discuss estimating the value of d, which requires multiple data replications; we simply assume that the value of d is known.

The problem is to partition the unit interval into k equal-width intervals. Let $C_i(k)$ denote the count (i.e., number of arrivals) in the ith interval, $((i-1)/k, i/k]$, $i = 1, 2, \ldots, k$. Then $N = \sum_{i=1}^{k} C_i(k)$ and each $C_i(k)$ satisfies $\mathrm{var}[C_i(k)] = d\mathrm{E}[C_i(k)] = d\int_{(i-1)/k}^{i/k} \lambda^*(t)dt$. Maybe surprisingly, while the variance-mean ratio is fundamental to our analysis, the independence of $C_i(k)$ and $C_j(k)$ (when $i \neq j$) is not.

978-1-5386-6572-5/18/$31.00 ©2018 IEEE

We evaluate the quality of a rate estimator $\hat{\lambda}$ using the usual mean integrated squared-error (MISE) function

$$\mathrm{E}\Big\{\int_0^1 [\lambda^*(t) - \hat{\lambda}(t)]^2 dt\Big\}. \tag{1}$$

For a given rate function λ^* and estimated rate function $\hat{\lambda}$, the optimal number of intervals is the value that minimizes the MISE function.

1.2 Solution Approach

In general, the MISE criterion for estimating true rate λ^* with estimator $\hat{\lambda}$ is as shown in Equation (1). Because $\int_0^1 (\lambda^*(t))^2 dt$ is a constant, we consider only the gMISE criterion

$$g(k) = \mathrm{E}\Big\{\int_0^1 [\hat{\lambda}^2(t) - 2\lambda^*(t)\hat{\lambda}(t)]dt\Big\}. \tag{2}$$

The optimal number of intervals is $k^* \equiv \mathrm{argmin}_{k=1,2,\ldots} g(k)$.

The heart of our problem is to estimate $g(k)$ for $k = 1,2,\ldots$. For each estimator of $g(k)$, say $\hat{g}(k)$, the corresponding estimator of the optimal number of intervals is $\hat{k}^* \equiv \mathrm{argmin}_{k=1,2,\ldots} \hat{g}(k)$. The associated quality measure of an estimator $\widehat{K}$ of k^* is $\mathrm{E}[g(\widehat{K})]$. Always, of course, $\mathrm{E}[g(\widehat{K})] \geq g(k^*)$.

We choose the estimated rate to be a piecewise-constant function with k intervals:

$$\hat{\lambda}_0(k;t) \equiv kC_i(k) \tag{3}$$

for $t \in ((i-1)/k, i/k]$ and $i = 1,2,\ldots,k$. Then $g(k^*)$ is the optimal value against which estimators compete.

1.3 Motivation

Our interest began during our WSC presentation of Chen and Schmeiser (2015). Christos Alexopoulos asked whether count data in our example (New York State Department of Transportation, 2015) would be better used by merging the fifteen-minute time intervals. Intuitively the answer was yes, but we didn't have a clean answer. We answer his question in Section 4.5.

A more-general motivation arises in regard generalizing the use of our two rate-smoothing algorithms: I-SMOOTH in Chen and Schmeiser (2013) and MNO-PQRS in Chen and Schmeiser (2017). Given a piecewise-constant rate function with k equal-width intervals, these algorithms return a smoother version. (I-SMOOTH creates new piecewise-constant functions, doubling the number of intervals and halving the width with each iteration. MNO-PQRS creates a piecewise-quadratic function.) Following initial efforts in Schmeiser et al. (2003), both algorithms are deterministic, taking the k-interval piecewise-constant rate function as given; their smoothed rate functions are constrained to maintain the expected number of arrivals for each of the k intervals. (Zheng and Glynn, 2017, refer to this constraint as the Local Equal Area Property (EAP).) Being deterministic, I-SMOOTH and MNO-PQRS (and, similarly the piecewise-linear algorithms in Nicol and Leemis 2014a and 2014b) estimate nothing; the given piecewise-constant rate function could arise from infinitely many possible true rate functions.

A single question needs to be answered to use I-SMOOTH and MNO-PQRS (and Nicol and Leemis's algorithms) to estimate a true rate function: How to group given arrival-time data into k equal-width counts? Zheng and Glynn (2017) compare their maximum-likelihood algorithm to a Local EAP algorithm when arrival-time data are given, but with the number of intervals being a given algorithm parameter. Using an estimated optimal number of intervals would allow each algorithm to perform at its best.

Finally, a comment about what is not our motivation. We are not estimating the cumulative rate function from arrival-time data, which is well solved by Arkin and Leemis (2000) and Leemis (1991 and 2004). We consider here the piecewise-constant rate estimator, rather than the more complicated rate models as in Kuhl and Wilson (2000 and 2001). Gerhardt and Nelson (2009), Liu (2013), Oreshkin et al. (2016), and Zhang et al. (2014) model more-general point processes, while we have restricted ourself to non-homogeneous non-Poisson processes with a constant dispersion ratio. (We can hope, however, that our number-or-intervals applied to more general non-Poisson processes might work well.)

1.4 Literature Review

Shimazaki and Shinomoto (2007) use the MISE criterion to find the optimal interval width for NHPPs. Their results are the same as ours when the dispersion ratio $d = 1$; we provide here a simpler and more general analysis. Shimazaki and Shinomoto (2007) and Henderson (2003) show that the optimal interval width is proportional to $N^{-1/3}$ when the rate function is smooth.

Also related is the optimal-bin-size literature for density estimation (e.g., Wasserman 2004, Section 20.2). Conditioned on the number N of arrivals from a NHPP, the N arrival times are independently and identically distributed with density function proportional to the rate function; estimating the arrival-time density function and rate function are then equivalent. Our analysis is simpler without needing conditioning on N.

1.5 Organization

The present paper's organization is as follows. In Section 2 we discuss examples and implications for the value of k^* using gMISE. In Section 3 we derive three estimators for $g(k), k = 1, 2, ...$; the natural estimator $k[N - \sum_{i=1}^{k} C_i^2(k)]$, the unbiased estimator $k[2N - \sum_{i=1}^{k} C_i^2(k)]$, and the approximately unbiased estimator $k[N - \sum_{i=1}^{k} C_i^3(k)/(1 - C_i(k))]$. In Section 4 we argue with examples that the natural estimator works badly and that the unbiased estimator works best (and the approximately unbiased estimator is a close second). In Section 4.5 we illustrate merging intervals to create better count data. Section 5 is a summary.

2 THE gMISE CRITERION

When specialized to the piecewise-constant estimator $\{\hat{\lambda}_0(k;t), 0 \le t \le 1)\}$ of Equation (3), the gMISE criterion becomes a function of the number of intervals, k. Define $\overline{\lambda} \equiv \int_0^1 \lambda^*(t)dt$, the expected number of arrivals. (Recall that $C_i(k)$ is the number of arrivals in interval i when there are k intervals.)

Result 1 The gMISE criterion for the piecewise-constant estimator $\hat{\lambda}_0$ is

$$g(k) = k \sum_{i=1}^{k} \Big\{ \text{var}[C_i(k)] - \text{E}^2[(C_i(k)] \Big\}. \tag{4}$$

Result 1 yields Corollary 1 because of the constant dispersion ratio. Appendix A contains proofs for Result 1 and Corollary 1.

Corollary 1 Since the dispersion ratio $\text{var}[C_i(k)]/\text{E}[C_i(k)]$ equals d for all $i = 1, ..., k$,

$$g(k) = k\{d\overline{\lambda} - \sum_{i=1}^{k} \text{E}^2[C_i(k)]\}. \tag{5}$$

Corollary 1 implies the following two corollaries that concern using a single interval. Corollary 2 follows because, when there is only one interval, $\sum_{i=1}^{k} \text{E}^2[C_i(k)] = \text{E}^2[C_1(1)] = \text{E}^2(N) = \overline{\lambda}^2$.

Corollary 2 $g(1) = \overline{\lambda}(d - \overline{\lambda})$.

Corollary 3 says, regardless of the true rate function, λ^*, that a single interval is optimal when there are very few arrival times.

Corollary 3 If $0 < \overline{\lambda} \le d$, then $k^* = 1$.

The proof of Corollary 3 is less direct; it also is in Appendix A.

The optimal number of intervals, k^*, can be small despite the expected number of arrivals, $\overline{\lambda}$, being large. For example, let $\lambda^*(t) = 0$ for $0 < t \le 1/2$ and $\lambda^*(t) = 2\overline{\lambda}$ for $1/2 < t \le 1$. Because the true rate function has two intervals with constant rates, $k^* = 2$ seems natural. But Corollary 3 says that $k^* = 1$ when $\overline{\lambda} \leq d$; therefore k^* is a function of $\overline{\lambda}$. In fact, $k^* = 2$ for $\overline{\lambda} \geq d$. (Proof sketch: If k is even, $g(k) = \overline{\lambda}(kd - 2\overline{\lambda})$, so $g(2) = 2\overline{\lambda}(d - \overline{\lambda})$ is minimal for all even k and all values of $\overline{\lambda}$. If k is odd, $g(k) = \overline{\lambda}[kd + \overline{\lambda}(k^{-1} - 2)]$, which is minimized if k is $(\overline{\lambda}/d)^{1/2}$ (or an odd positive integer next to it); $g((\overline{\lambda}/d)^{1/2}) = 2\overline{\lambda}[(d\overline{\lambda})^{1/2} - \overline{\lambda}]$. For $\overline{\lambda} = d$, $g(1) = g(2) = 0$ and hence, k^* can be 1 or 2. For $\overline{\lambda} > d$, $g(2) < g((\overline{\lambda}/d)^{1/2})$ because $(d\overline{\lambda})^{1/2} > d$ and hence, $k^* = 2$.)

Result 2 shows that if the rate function is smooth, k^* increases at order $(\overline{\lambda}/d)^{1/3}$. The cube-root growth is consistent with the analysis in Henderson (2003) and Shimazaki and Shinomoto (2007) for NHPPs.

Result 2 If the rate function $\lambda^*(t)$ is continuous and its first derivative exists with $\int_0^1 [\frac{\partial}{\partial t}\lambda^*(t)]^2 dt < \infty$, then

$$\mathrm{E}\left\{\int_0^1 [\lambda^*(t) - \hat{\lambda}(t)]^2 dt\right\} \approx \overline{\lambda}\left\{kd + \overline{\lambda}(12k^2)^{-1}\int_0^1 \left[\frac{\frac{\partial}{\partial t}\lambda^*(t)}{\overline{\lambda}}\right]^2 dt\right\} \tag{6}$$

and hence,

$$k^* \approx \left\{\frac{\overline{\lambda}}{6d}\int_0^1 \left[\frac{\frac{\partial}{\partial t}\lambda^*(t)}{\overline{\lambda}}\right]^2 dt\right\}^{1/3}. \tag{7}$$

Equation (6) is based on the Taylor's approximation of $\lambda^*(t)$ up to the first order. The proof is similar to that in Wasserman (2004, Section 20.2). Our numerical results (not reported here) show that the cube-root growth still holds if $\lambda^*(t)$ is only differentiable almost everywhere.

As an illustration, consider the specific true rate function $\lambda^*(t) = 2\overline{\lambda}t$ for $0 < t \leq 1$. Then

$$g(k) = \overline{\lambda}[kd - \overline{\lambda}(4 - k^{-2})/3]. \tag{8}$$

For large values of $\overline{\lambda}$, minimizing $g(k)$ over $k = 1, 2, \ldots$ yields

$$k^* \approx [2\overline{\lambda}/(3d)]^{1/3}. \tag{9}$$

To help visualize, consider a NHPP ($d = 1$) with $\lambda^*(t) = \overline{\lambda}[1 + \sin(4\pi t)]$ for $t \in (0, 1]$; that is, the true rate function is a sine-wave with two periods in the unit interval. Figure 1 shows the rate function λ^* for $\overline{\lambda} = 1, 100, 1000$ in the top row. In the bottom row are the corresponding gMISE criteria. The optimal numbers of intervals are $k^* = 1, 11, 23$, the minimal point of $g(k), k = 1, 2, \ldots$. In Section 4.3, we return to this example (but only with $\overline{\lambda} = 1000$) to visualize estimation.

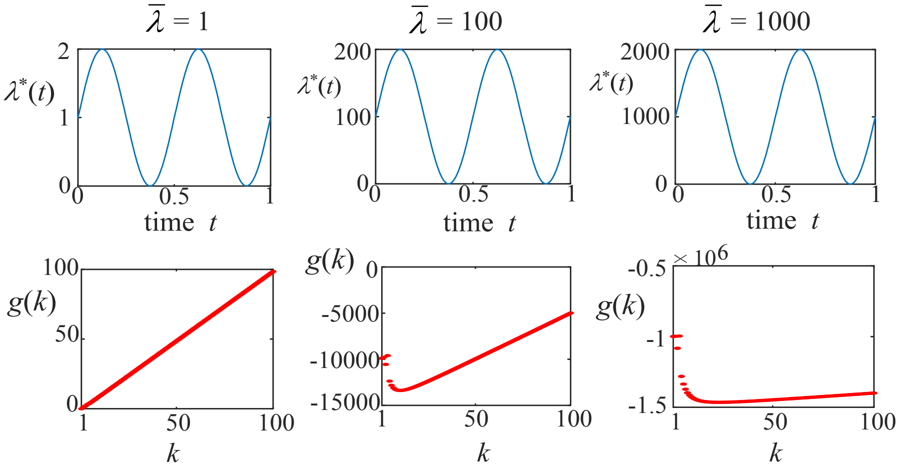


Figure 1: Sine rate function and three associated gMISE functions.

In summary, at least for these examples, the behavior of g and k^* is consistent with expectations for both small and large values of $\overline{\lambda}$.

3 ESTIMATING THE *g*MISE CRITERION

We now discuss estimating $g(k)$, for $k = 1, 2, \ldots$. Given an estimator, say $\hat{g}(k)$, for $k = 1, 2, \ldots$, our estimator $\hat{k}^*$ of k^* will be the minimizer of $\{\hat{g}(k), k = 1, 2, \ldots\}$. Our three estimators of $g(k)$ are developed in Sections 3.1, 3.2, and 3.3.

3.1 Natural Estimator

Our first estimator is the *natural* estimator of $g(k) = k\{d\overline{\lambda} - \sum_{i=1}^{k} \mathrm{E}^2[C_i(k)]\}$, obtained by substituting natural estimators for the unknowns $\overline{\lambda}, \mathrm{E}[C_1(k)], \mathrm{E}[C_2(k)], \ldots, \mathrm{E}[C_k(k)]$. The natural estimator is then

$$\hat{g}_0(k) \equiv k\Big[dN - \sum_{i=1}^{k} C_i^2(k)\Big]. \tag{10}$$

Estimating $\overline{\lambda}$ with N causes no bias, but estimating $\mathrm{E}^2[C_i(k)]$ with $C_i^2(k)$ is concerning, since $\mathrm{E}[C_i^2(k)] = \mathrm{var}[C_i(k)] + \mathrm{E}^2[C_i(k)]$. Taking the expected value of $\hat{g}_0(k)$ yields

$$\mathrm{E}[\hat{g}_0(k)] = k\Big[0 - \sum_{i=1}^{k} \mathrm{E}^2[C_i(k)]\Big]. \tag{11}$$

The bias is $\mathrm{E}[\hat{g}_0(k)] - g(k) = -kd\overline{\lambda}$. In Section 4, this natural estimator performs so badly as to be useless.

3.2 Unbiased Estimator

Our second estimator subtracts $-kdN$, an estimator of the bias in $\hat{g}_0(k)$. Therefore, define the unbiased estimator

$$\hat{g}_1(k) \equiv k\Big[2dN - \sum_{i=1}^{k} C_i^2(k)\Big]. \tag{12}$$

In Section 4.4, this estimator performs best. Appendix B contains a Fortran implementation `kuhat`, where this unbiased estimator is called simply `khat`. For $d = 1$, the estimator $\hat{g}_1(k)$ is the same as that in Shimazaki and Shinomoto (2007) and similar to the *cross-validation* bin-width estimator for optimizing histograms (Wasserman 2004, p. 310).

3.3 Approximately Unbiased Estimator

Our third estimator also adjusts $\hat{g}_0$ to reduce bias. Recall that $\mathrm{E}[C_i^2(k)] = \mathrm{var}[C_i(k)] + \mathrm{E}^2[C_i(k)]$. Then

$$\frac{\mathrm{E}^2[C_i(k)]}{\mathrm{E}[C_i^2(k)]} = \frac{\mathrm{E}^2[C_i(k)]}{\mathrm{var}[C_i(k)] + \mathrm{E}^2[C_i(k)]} = \frac{\mathrm{E}^2[C_i(k)]}{d\mathrm{E}[C_i(k)] + \mathrm{E}^2[C_i(k)]} = \frac{\mathrm{E}[C_i(k)]}{d + \mathrm{E}[C_i(k)]}. \tag{13}$$

The natural estimator of this fraction is $C_i(k)/[d + C_i(k)]$, so our third estimator is

$$\hat{g}_2(k) \equiv k\Big\{dN - \sum_{i=1}^{k} C_i^3(k)/[d + C_i(k)]\Big\}. \tag{14}$$

In Section 4.4, $\hat{g}_2(k)$ works well, but not as well as the unbiased estimator $\hat{g}_1(k)$.

3.4 Other Estimators

Although we compare to no other estimators, here we mention alternatives.

Whatever the estimator of $g(k)$, the search for the minimizer $\hat{k}^*$ is affected by the randomness of $\hat{g}(k)$. The random variables $\hat{g}(1), \hat{g}(2), \ldots$ are computed from the same data set, so they are positively correlated, making the search more reliable. Nevertheless, occasionally a particularly large value of k is selected. Therefore, alternatives to choosing the global minimum might be useful. In the analogous situation of minimizing the MSER statistics for the initial-transient problem, Pasupathy and Schmeiser (2010) discuss considering local minima.

We (the authors, between ourselves) have discussed, but not implemented, computationally intensive estimators, such as jackknifing. Although $\hat{g}_1(k)$ is an unbiased estimator of $g(k)$, the estimated gMISE optimal number of intervals, $\hat{k}_1^*$, is a biased estimator of k^*.

4 PERFORMANCE

Using Monte Carlo experiments of the common NHPPs ($d = 1$), we now compare and evaluate the performance of our three estimators(from Section 3) for k^*, the MISE-optimal number of intervals, for a given true rate function λ^* and expected number of arrivals, $\overline{\lambda}$.

4.1 Performance Measures

Performance of the estimators of k^* is not directly about the distribution of the $\hat{k}^*$s, at least not via the usual measures such as means, standard deviations, and probability that $\hat{k}^* = k^*$. Consider an example where the true rate function is a step function, with three steps and $k^* = 3$. For $\overline{\lambda} = 1000$, The corresponding (estimated via Monte Carlo) values of $g(k), k = 1,2,\ldots,15$ are $-985809.9, -1422752.0, -1642366.2, -1531895.3, -1552870.2, -1642027.7, -1577776.7, -1584700.7, -1638908.5, -1593894.4, -1597008.0, -1636926.1, -1602005.5, -1602739.3$, and -1633231.3. The g function is minimized at $g(3) = -1642366.2$, so $k^* = 3$. But $g(6) = -1642027.7$ is a close second best, followed by $g(9) = -1638908.5$, $g(12) = -1636926.1$, and $g(15) = -1633231.3$. How did $\hat{k}_1^*$ perform? Here are the histogram counts: $0,0,763,0,0,79,0,1,53,1,0,32,0,0,21,0,3,10,1,0,14,0,1,10,3,1,2,1,0,4$. For this example, the usual properties of $\hat{k}_1^*$ are much less relevant than the property that $\hat{k}_1^*$ frequents multiples of three. For smoother rate functions λ^*, of course, g is also smoother and $\hat{k}^*$ being close to k^* is meaningful.

Because gMISE is appropriate for all applications, we use it as our primary and only criterion. For presentation we use SE-gMISE, the scaled excess gMISE; specifically, $\{\mathrm{E}[g(\hat{k}^*)] - g(k^*)\}/\overline{\lambda}^2$. Subtracting $g(k^*)$ guarantees that the SE-gMISE is nonnegative; dividing by $\overline{\lambda}^2$ makes SE-gMISE unitless. SE-gMISE is still a function of $\overline{\lambda}$ (and the entire true rate function λ^*); in particular, $k^* = \mathrm{argmin}_{k=1,2,\ldots} g(k)$ depends upon $\overline{\lambda}$.

A *scenario* is defined by a true rate function, λ^*. From it, the associated expected number of arrivals, $\overline{\lambda}$, is implied. Also implied is the SE-gMISE function and its minimizer k^*, the MISE-optimal number of intervals. Given a sequence of scenarios created by scaling by $\overline{\lambda}$, the optimal number of intervals k^* becomes as function of $\overline{\lambda}$. Comparison to the cube-root growth in Result 2 is then possible.

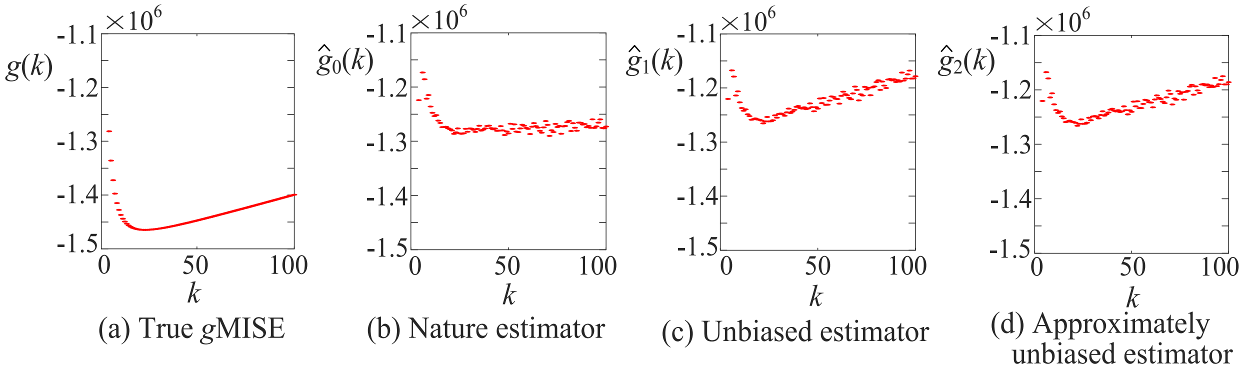
4.2 The Experiment

Fix a scenario. For each of the three estimators, we estimate the distribution of $\hat{k}^*$ over $\{1,2,\ldots\}$. We then can estimate the value of the SE-gMISE using

$$\frac{\hat{\mathrm{E}}[g(\hat{k}^*)] - g(k^*)}{\overline{\lambda}^2} = \frac{[\sum_{k=1}^{\infty} g(k)\hat{P}(\hat{k}^* = k)] - g(k^*)}{\overline{\lambda}^2}. \tag{15}$$

Specifically, our Monte Carlo experiment replicates `nreps` practitioners. For each replication, NHPP arrival times $a_1, a_2, \ldots, a_n$ are generated. For each estimator $i, i = 1,2,3$, $\hat{g}_i(k)$ is computed for $k = 1,2,\ldots$ and the associated minimizer $\hat{k}_i^*$ is found and stored in histogram i.

To allow arbitrary true rate functions λ^*, we estimate $g(k), k = 1,2,\ldots$ using the unbiased estimator $\hat{g}_1(k)$ for each of the `nreps` replications. After the experiment is completed, for each of the estimators, Equation (15) is used to estimate the SE-gMISE, treating the estimated g as the true function.



(a) True gMISE (b) Nature estimator (c) Unbiased estimator (d) Approximately unbiased estimator

Figure 2: The function g and its three estimators of Example 1.

Details. Monte Carlo results are correct to within one unit of the least significant digit shown, as discussed in Song and Schmeiser (2009). We use an implementation of the random-number generator in L'Ecuyer (1999).

4.3 Example 1: The Natural Estimator Works Badly

We show, with a single example, that the natural estimator $\hat{g}_0$ performs badly; so badly, that after this section we no longer consider $\hat{g}_0$.

Consider again (from Section 2) the sine-wave $\lambda^*(t) = \overline{\lambda}[1+\sin(4\pi t)]$ for $t \in [0,1]$. Figure 2 (a) shows the associated gMISE, which is minimized at $k^* = 23$. From a single replication with $\overline{\lambda} = 1000$, Parts (b), (c), and (d) show the gMISE estimates using $\hat{g}_0$, $\hat{g}_1$, and $\hat{g}_2$. The associated estimates of k^* are $\hat{k}^*_0 = 71$, $\hat{k}^*_1 = 22$, and $\hat{k}^*_2 = 22$.

Why is the natural estimator so bad? Part (b) shows that, without a bias correction, the natural estimators $\hat{g}_0(k)$ are almost flat after an initial descent for small values of k. The minimizer of $\hat{g}$ could lie almost anywhere from $k = 20$ and higher. The bias corrections used to create $\hat{g}_1$ and $\hat{g}_2$ have a first-order effect, giving them a well-defined area of minimal values in Parts (c) and (d).

A second point to be observed is that, for this realization, all three $\hat{g}$ estimates are noticeably higher than the true g function. For another realization, the estimated values could be lower. What allows minimization of $\hat{g}$ to work well is that the estimated function values are positively correlated, not that any particular $\hat{g}_i$ has a small variance.

4.4 Example 2: The Unbiased Estimator is Best

We now perform a Monte Carlo experiment to compare the unbiased estimator, $\hat{g}_1$, and the approximately unbiased estimator, $\hat{g}_2$. This example is consistent with every other example that we have tried. The two estimators are highly correlated with similar performance measures, but the unbiased estimator performs slightly better and is a bit easier to compute. Therefore, after this section, we consider only the unbiased estimator, $\hat{g}_1$.

Let $\lambda^*(t)$ be piecewise constant with rates $1000, 2000, 1500, 3000, 2500$ for the five intervals of length 0.2. Then $\overline{\lambda} = 2000$ and $k^* = 5$. We estimated performance with $10,000$ replications. The distributions of $\hat{k}^*_1$ and $\hat{k}^*_2$ are quite similar: the means are 6.8 and 6.9; the standard deviations are 5.1 and 5.2; the maximums are both 42; the probabilities of returning $\hat{k}^* = k^*$ are 0.83 and 0.83. In addition, their histograms are quite similar, with values falling on multiples of five. All of these measures slightly favor $\hat{g}_1$, but as discussed in Section 4.1, we trust the SE-gMISE criterion, where $\widehat{\text{SE-}g\text{MISE}}_1 = 0.0012$ and $\widehat{\text{SE-}g\text{MISE}}_2 = 0.0012$ but we observed no exception where the former is larger.

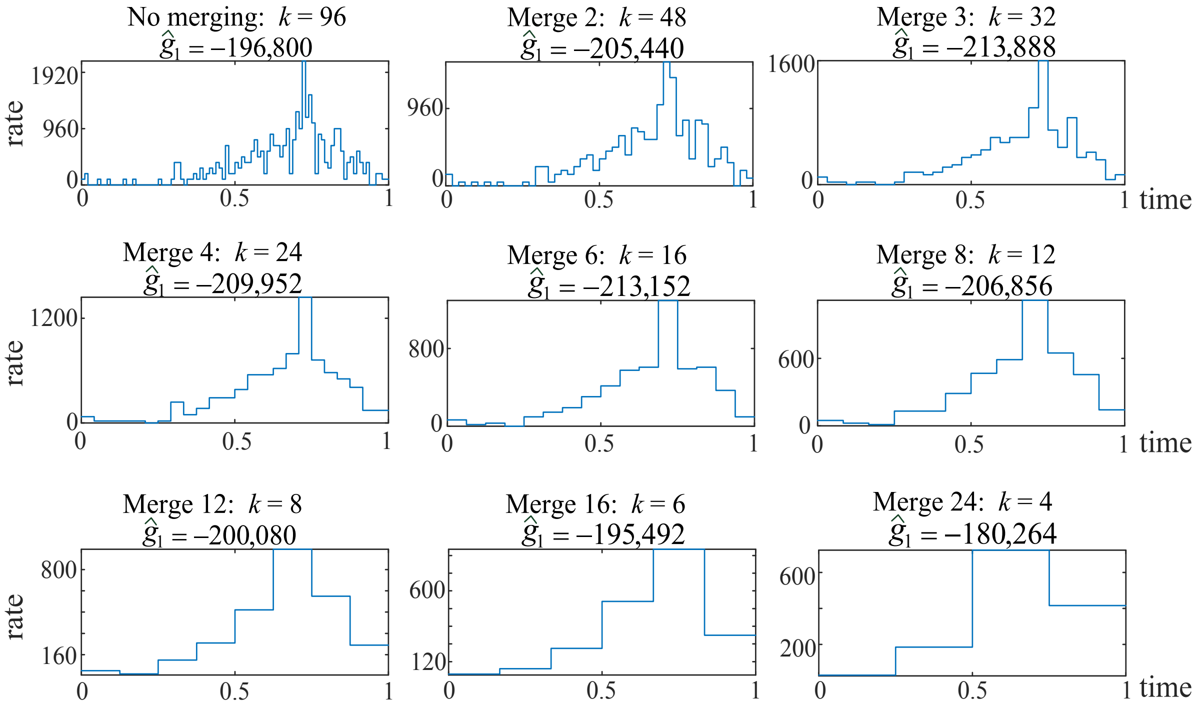


Figure 3: Nine options for merging the traffic-count data of Example 3. The time unit is one day.

4.5 Example 3: Merging Intervals

We revisit our (Chen and Schmeiser, 2015) example of New York traffic counts (New York State Department of Transportation, 2015) to answer Christos's question about merging intervals. The website provides counts for 96 fifteen-minute intervals over twenty-four hours. There are twelve possible mergers, with $k \in \{96, 48, 32, 24, 16, 12, 8, 6, 4, 3, 2, 1\}$. Figure 3 illustrates the piecewise-constant estimated rates for the first nine of the possible mergers. Most humans, we think, gravitate toward $k \in \{32, 24, 16\}$.

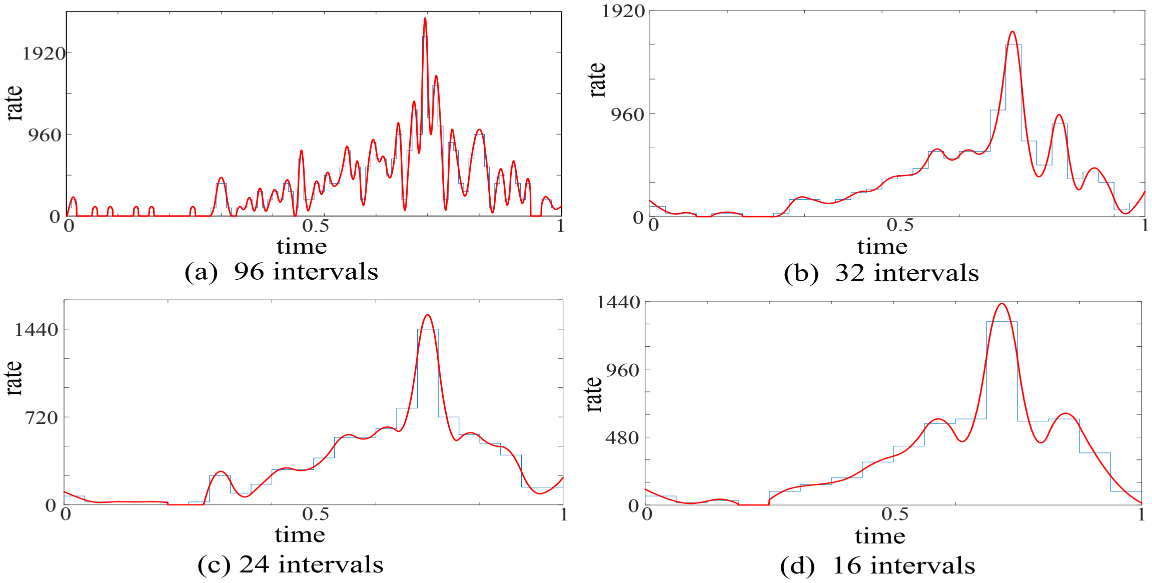
The estimated gMISE criterion values, assuming $d = 1$, are shown. (The value of SE-gMISE is not shown, since for real-world data we know neither the value of $g(k^*)$ nor the value of $\overline{\lambda}$. We assume that $d = 1$ because d can not be estimated from only one data realization.) The gMISE criterion selects $k = 32$ as best, with $k = 16$ a close second best, and $k = 24$ a close third best. To answer Christos's question, the original $k = 96$ is not best. If our purpose in 2015 had been to estimate the true rate function (rather than smoothing the rate function while maintaining the Local EAPs), merging intervals would have been wise.

Aside. Conveniently, $k = 96$ has many factors; what should be done when the original number of intervals has fewer factors (or, in the extreme, is prime). Another thought: For cyclic contexts, where the ending rate is known to equal the beginning rate, some type of overlapping analysis, analogous to that for batch means in Meketon and Schmeiser (1984), might be considered.

4.6 Example 3: Comparing MNO-PQRS Smoothing

Returning to our motivation, we continue with the New York traffic data of Example 3. We use MNO-PQRS to smooth the original $k = 96$ counts and compare the MNO-PQRS smoothing of the merged intervals. Figure 4 shows the MNO-PQRS fits for the original $k = 96$ and for three mergings that are close to being gMISE optimal: $k = 32, 24,$ and 16.

As with the piecewise-constant fits in Section 4.5, these merged-interval piecewise-quadratic fits appear (to the authors) to be an improvement compared to the fit using the original intervals.



(a) 96 intervals

(b) 32 intervals

(c) 24 intervals

(d) 16 intervals

Figure 4: MNO-PQRS fits from the original 96, and from 32, 24, and 16 merged intervals of Example 3.

5 SUMMARY

The underlying problem is to estimate, from given arrival times, the true rate function with a piecewise-constant rate function. The problem is to choose a good number of equal-width intervals. We choose to minimize the mean integrated squared-error (MISE) criterion under the fixed-dispersion-ratio assumption. From that choice, we develop three estimators, concluding that the unbiased estimator is preferable. We provide a Fortran implementation of the subroutine `kuhat`, which takes arrival-time data and returns an estimated optimal number of intervals. We illustrate how to merge count data when the original arrival times are not available.

ACKNOWLEDGMENTS

Taiwan's Ministry of Science and Technology provided financial support through Project MOST 105-2221-E-033-050-MY3. Pierre L'Ecuyer, Shane Henderson, Jeff Hong, Barry Nelson, Raghu Pasupathy, Lee Schruben, and Mike Taaffe provided helpful discussions.

A PROOFS

The proof of Result 1 is as follows.

$$
\begin{aligned}
g(k) &= \mathrm{E}\Big\{\int_0^1 [\hat{\lambda}_0^2(k;t) - 2\lambda^*(t)\hat{\lambda}_0(k;t)]dt\Big\} = \int_0^1 \{E[\hat{\lambda}_0^2(k;t)] - 2\lambda^*(t)\mathrm{E}[\hat{\lambda}_0(k;t)]\}dt \\
&= \sum_{i=1}^{k}\int_{(i-1)/k}^{i/k} \{\mathrm{E}[k^2C_i^2(k)] - 2\lambda^*(t)\mathrm{E}[kC_i(k)]\}dt \\
&= \sum_{i=1}^{k}\int_{(i-1)/k}^{i/k} \Big\{k^2\big[\mathrm{var}[C_i(k)] + \mathrm{E}^2[C_i(k)]\big] - 2\lambda^*(t)k\mathrm{E}[C_i(k)]\Big\}dt \\
&= \sum_{i=1}^{k}\Big\{k\,\mathrm{var}[C_i(k)] + k\mathrm{E}^2[(C_i(k)] - 2k\mathrm{E}[C_i(k)]\int_{(i-1)/k}^{i/k}\lambda^*(t)dt\Big\} = k\sum_{i=1}^{k}\Big\{\mathrm{var}[C_i(k)] - \mathrm{E}^2[(C_i(k)]\Big\}.\square
\end{aligned}
$$

The proof of Corollary 1 uses the property that var$[C_i(k)] = d\mathrm{E}[C_i(k)] = d\int_{(i-1)/k}^{i/k}\lambda^*(t)dt$. Since $\sum_{i=1}^{k}\int_{(i-1)/k}^{i/k}\lambda^*(t)dt = \overline{\lambda}$, we have $g(k) = k\Big\{d\overline{\lambda} - \sum_{i=1}^{k}\mathrm{E}^2[(C_i(k)]\Big\}$.

To prove Corollary 3, assume that $0 < \overline{\lambda} \le d$. For $i = 1,2,\ldots$ define $p_i(k) \equiv \mathrm{E}[C_i(k)]/\overline{\lambda}$, the fraction of expected arrivals in interval i. Then $\sum_{i=1}^{k} p_i^2(k) \le \sum_{i=1}^{k} p_i(k) = 1$. Rewrite $g(k)$ to obtain

$$g(k) = k\overline{\lambda}\Big[d - \overline{\lambda}\sum_{i=1}^{k} p_i^2(k)\Big]; \tag{16}$$

then $g(k) \ge k\overline{\lambda}(d-\overline{\lambda}) \ge g(1) \ge 0$. Both nonnegative factors in Equation (16) are minimized by $k = 1$; the first factor is immediate and the second follows from $\overline{\lambda} \le d$ implying that $0 \le d - \overline{\lambda} \le d - \overline{\lambda}\sum_{i=1}^{k} p_i^2(k)$. $\square$

B COMPUTER CODE

The Fortran subroutine `kuhat` is the logic needed to compute the unbiased estimator $\hat{k}_1^*$ from n arrival times, as discussed in Section 3.2. Here $\hat{k}^*$ becomes the variable `khat`. Unlike the analysis discussion, where the arrival-time data are assumed to lie in the unit interval, in `kuhat` the arrival times can lie in any interval, from `tbegin` to `tend`.

Other than the time intervals and the dispersion ratio d, there are no user-specified parameters. The variable `kmax`, however, requires discussion. As coded, `kmax = 2 * (n/d)**(0.4)` is the largest possible value of `khat`; that is, the enumeration over the number of intervals k is from one to $2(n/d)^{0.4}$. By design, `kmax` grows faster than the asymptotic optimal $(n/d)^{1/3}$ rate. Larger values of `kmax` could be used, with corresponding additional computation time.

To think of Fortran as pseudocode, some facts are needed. An exclamation point begins a comment. Variables beginning with i, j, k, l, m, and n are integers; all other variables are doubles. `1d100` is 10^{100} expressed as a double. Exponentiation is expressed with `**`. `do` loops are analogous to `for` loops. The line `dimension atime(1:n)` defines `atime` to be a one-dimensional array indexed one to n; the size `n` could be any value, since memory for `atime` is assigned by the calling program. All variables are local, with no reliance upon earlier calls to `kuhat`.

```
subroutine kuhat( tbegin, tend, n, atime, d, khat )
! Huifen Chen and Bruce Schmeiser. February 10, 2018.
! Practitioner version, for real-world use and publication.
! Purpose: Estimate the gMISE-optimal number of intervals for
!    using a piecewise-constant rate function to estimate the
!    rate function from which the arrival times are sampled.
! Reference: 2018 Winter Simulation Conference Proceedings
! Input:  tbegin = start time of data collection
!         tend   = end time of data collection
!         n      = number of arrival times
!         atime  = contains n arrival times in [tbegin, tend]
!         d      = dispersion ratio, V[C_i(k)] / E[C_i(k)]
! Output: khat   = estimated optimal number of intervals
implicit double precision (a-h,o-z)
implicit integer (i-n)
dimension atime(1:n)
khat = 1
```

```
if (n .gt. 1) then
   tspread = tend - tbegin
   gmin = 1d100
   kmax = 2 * (n/d)**(0.4)
   do k=1,kmax    ! Unbiased gMISE estimation for k intervals
      sum2 = 0
      width = tspread / k
      nevent = 1
      do i=1,k
         count = 0
         dowhile (nevent.le.n .and. atime(nevent).le.i*width)
            count  = count  + 1
            nevent = nevent + 1
         enddo
         sum2 = sum2 + count**2
      enddo
      ghat = k * (2*d*n - sum2)
      if (ghat .lt. gmin) then ! If new smallest ghat, update
         gmin = ghat
         khat = k
      endif
   enddo
endif
return
end
```

REFERENCES

Arkin, B.L. and L.M. Leemis. 2000. “Nonparametric Estimation of the Cumulative Intensity Function for a Nonhomogeneous Poisson Process from Overlapping Realizations”. *Management Science* 46(7):989–998.

Chen, H. and B.W. Schmeiser. 2013. “I-SMOOTH: Iteratively Smoothing Mean-Constrained and Nonnegative Piecewise-Constant Functions”. *INFORMS Journal on Computing* 25(3):432–445.

Chen, H. and B.W. Schmeiser. 2015. “The MNO–PQRS Poisson Point Process: Generating the Next Event Time”. In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz et al., 575–585. Piscataway, New Jersey: IEEE.

Chen, H. and B.W. Schmeiser. 2017. “MNO–PQRS: Max Nonnegativity Ordering—Piecewise-Quadratic Rate Smoothing”. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 27(3), Article 18:1–19.

Gerhardt, I. and B.L. Nelson. 2009. “Transforming Renewal Processes for Simulation of Nonstationary Arrival Processes”. *INFORMS Journal on Computing* 21(4):630–640.

Henderson, S.G. 2003. “Estimation for Nonhomogeneous Poisson Processes from Aggregated Data”. *Operations Research Letters* 31(5):375–382.

Kuhl, M.E. and J.R. Wilson. 2000. “Least Squares Estimation of Nonhomogeneous Poisson Processes”. *Journal of Statistical Computation and Simulation* 67:75–108.

Kuhl, M.E. and J.R. Wilson. 2001. “Modeling and Simulating Poisson Processes having Trends or Non-Trigonometric Cyclic Effects”. *European Journal of Operational Research* 133(3):566–582.

L’Ecuyer, P. 1999. “Good Parameters and Implementations for Combined Multiple Recursive Random Number Generators”. *Operations Research* 47:159–164.

Leemis, L.M. 1991. “Nonparametric Estimation of the Cumulative Intensity Function for a Nonhomogeneous Poisson Process”. *Management Science* 37(7):886–900.

Leemis, L.M. 2004. “Nonparametric Estimation and Variate Generation for a Nonhomogeneous Poisson Process from Event Count Data”. *IIE Transactions* 36(12):1155–1160.

Liu, R. 2013. *Modeling and Simulation of Nonstationary Non-Poisson Processes*. Ph.D. Dissertation. Raleigh, North Carolina: North Carolina State University.

Meketon, M.S. and B.W. Schmeiser. 1984. “Overlapping Batch Means: Something for Nothing?” In *Proceedings of the 1986 Winter Simulation Conference*, edited by S. Sheppard et al., 227–230. Piscataway, New Jersey: IEEE.

New York State Department of Transportation. 2015. Available via https://www.dot.ny.gov/divisions/engineering/technical-services/highway-data-services/hdsb/albany [most-recently accessed February 6, 2018].

Nicol, D.M. and L.M. Leemis. 2014a. *Continuous Piecewise-Linear Intensity Function Estimation for Nonhomogeneous Poisson Process Count Data.* Technical Report, Department of Mathematics, The College of William & Mary.

Nicol, D.M. and L.M. Leemis. 2014b. “A Continuous Piecewise-Linear NHPP Intensity Function Estimator”. In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk et al., 498–509. Piscataway, New Jersey: IEEE.

Oreshkin, B.N., N. Régnard, and P. L’Ecuyer. 2016. “Rate-Based Daily Arrival Process Models with Application to Call Centers”. *Operations Research* 64(2):510–525.

Pasupathy, R. and B.W. Schmeiser. 2010. “The Initial Transient in Steady-State Point Estimation: Contexts, a Bibliography, the MSE Criterion, and the MSER Statistic”. In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson et al., 184–197. Piscataway, New Jersey: IEEE.

Schmeiser, B.W., R. Rao, and N. Kumala. 2003. “Smoothing Piecewise-Constant Rate Functions”. In *Proceedings of the Industrial Engineering Research Conference*, Norcross, Georgia: Institute of Industrial Engineers, 6 pages (cd only).

Shimazaki, H. and S. Shinomoto. 2007. “A method for Selecting the Bin Size of a Time Histogram”. *Neural Computation* 19:1503–1527.

Song, W.-T. and B.W. Schmeiser. 2009. “Omitting Meaningless Digits in Point Estimates: The Probability Guarantee of Leading-Digit Rules”. *Operations Research* 57(1):109–117.

Wasserman, L. 2004. *All of Statistics*. New York: Springer.

Zhang X., L.J. Hong, and J. Zhang. 2014. “Scaling and Modeling of Call Center Arrivals”. In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk et al., 476–485. Piscataway, New Jersey: IEEE.

Zheng, Z. and P.W. Glynn. 2017. “Fitting Continuous Piecewise Linear Poisson Intensities via Maximum Likelihood and Least Squares”. In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan et al., 1740–1749. Piscataway, New Jersey: IEEE.

AUTHOR BIOGRAPHIES

HUIFEN CHEN is a Professor in the Department of Industrial and Systems Engineering at Chung-Yuan University, Taiwan. She completed her Ph.D. in Industrial Engineering at Purdue University in 1994. Her research interests center on Monte Carlo simulation. Her email address is huifen@cycu.edu.tw.

BRUCE SCHMEISER is a Professor Emeritus in the School of Industrial Engineering at Purdue University. His research interests center on developing methods for better simulation experiments. He is a fellow of INFORMS and IIE, as well as recipient of the Informs Simulation Society’s 2014 Lifetime Professional Achievement Award. A long-time participant in the WSC, he served as the 1983 Program Chair and the 1988–1990 President of the Board of Directors. His e-mail address is bruceschmeiser@gmail.com.