ELSEVIER

Earth and Planetary Science Letters 199 (2002) 1–6

EPSL

www.elsevier.com/locate/epsl

Estimating event rates in the presence of dating error with an application to lunar impacts

Andrew R. Solow *

Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA

Received 26 September 2001; received in revised form 12 February 2002; accepted 18 February 2002

Abstract

Radiometric ages of objects are often used to reconstruct historical variations in the rate function of geological events. Measurement error in such ages can lead to a bias in the estimated rate function. This paper describes a method for estimating the historical rate function that accounts for measurement error. The method is applied to the estimation of the rate of lunar impacts over the past 3.5 billion years from the argon–argon ages of 155 impact spherules. A simulation study of the performance of the method is also presented.

Keywords: absolute age errors; impacts; lunar samples; rates; simulation

1. Introduction

Radiometric ages of objects are commonly used to estimate variations in the historical rates of geological events. In a recent example considered in some detail below, Culler et al. [1] used the argon–argon ages of 155 impact spherules collected during the Apollo 14 mission to reconstruct variations in the rate of lunar impacts over the past $\sim$3.5 billion years. In doing so, Culler et al. used a method based on an ideogram to account for measurement error in the estimated ages of the spherules. This paper describes an alternative approach to analyzing data of this kind. Under this approach, the event times are modelled as arising from a point process [2]. A point process is a stochastic process giving rise to discrete events in continuous time. Such a process is characterized in part by a rate function that describes how the mean number of events in a unit time interval varies over time. The goal of the approach described here is to estimate this rate function.

Because there is often no a priori basis for specifying the form of the rate function, a nonparametric kernel estimator is used that assumes only that the rate function varies smoothly over time [3–5]. In using this (or any other) estimator, it is important to account for measurement error in the impact times. The effect of measurement error on the nonparametric estimation of the rate function of a point process does not appear to have been studied. However, results for the related problems of nonparametric density estimation and regression [6] suggest that measurement error

* Tel.: +1-50-82-89-27-46; Fax: +1-50-84-57-24-84.
E-mail address: asolow@whoi.edu (A.R. Solow).

0012-821X/02/$ – see front matter
PII: S0012-821X(02)00541-1

can lead to serious bias in this context. In this paper, a general purpose method called SIMEX [7] is used to correct for the effects of measurement error.

The remainder of the paper is organized in the following way. In Section 2, the basic approach is described and compared to the ideogram used in [1]. In Section 3, this approach is applied to the data from [1]. The results of a small simulation study of the performance of the proposed approach are presented in Section 4. Section 5 contains some brief concluding remarks.

2. Approach

Let $t_1,t_2,...,t_n$ be the unknown true ages of n events. These ages are assumed to have arisen from a point process with rate function $\lambda(t)$ operating over the known interval (0, T). The rate function $\lambda(t)$ can be interpreted as the mean number of events occurring in a unit time interval centered at t. The measured age Y_j of event j is assumed to follow the model:

$$Y_j = t_j + \varepsilon_j \tag{1}$$

where ε_j is a normal measurement error with mean 0 and known variance σ_j^2. The problem considered in this paper is the nonparametric estimation of the rate function $\lambda(t)$ from the measured ages $Y_1, Y_2,..., Y_n$.

To begin with, suppose that the ages of the events are measured without error, so that $Y_j = t_j$ for all j. Under the assumption that $\lambda(t)$ varies smoothly with t, Diggle [3] proposed the kernel estimator:

$$\tilde{\lambda}(t) = \frac{1}{h}\sum_{j=1}^{n} K\left(\frac{t-Y_j}{h}\right) \tag{2}$$

where the kernel function K is a probability density function symmetric about 0 and h is a bandwidth that controls the smoothness of $\lambda(t)$. For $t<h$, part of the kernel extends below the lower bound of the observation period at 0. This can lead to bias in estimating $\lambda(t)$ in the neighborhood of $t=0$. A simple way to avoid this is to reflect or fold at 0 the part of the kernel extending below 0, so that all of its mass lies within the observation period. The resulting estimator can be written:

$$\hat{\lambda}(t) = \frac{1}{h}\sum_{j=1}^{n} K\left(\frac{t-Y_j}{h}\right) + K\left(\frac{t+Y_j}{h}\right) \tag{3}$$

The same problem can arise and a similar correction can be made for $t>T-h$. In this paper, the so-called bisquare kernel:

$$K(u) = \begin{array}{ll} 0.9375\,(1-u^2)^2 & |u|\le 1 \\ 0 & |u|>1 \end{array} \tag{4}$$

will be used.

As with all estimators, the mean squared error of $\lambda(t)$ includes contributions from bias and variability. The bandwidth h controls the trade-off between these sources of error. If the bandwidth is small, then the contribution of bias to mean squared error will tend to be small, but the contribution of variability will tend to be large. As a result, $\lambda(t)$ will exhibit excessive variability over periods in which $\lambda(t)$ is smooth. Conversely, if the bandwidth is large, the contribution of bias to mean squared error will tend to be large, but the contribution of variability will tend to be small. In this case, $\lambda(t)$ will tend to smooth out local features in $\lambda(t)$. A number of automatic methods for bandwidth selection have been proposed that attempt to strike a reasonable balance between bias and variability [3–5,9]. Unfortunately, these methods are all affected by measurement error. Specifically, due to the attenuating effect of measurement discussed below, they will tend to select a bandwidth that is too large. Bandwidth selection in the presence of measurement error remains an open question. In the following section, a bandwidth corresponding to 10% of the length of the estimation period is used. This is toward the small end of the range of typical bandwidths (5–20% of the observation period) [7]. In qualitative terms, the results presented in the next section are not highly sensitive to bandwidths in this range.

The kernel estimator outlined above assumes that the ages of the events are observed without error. Here, a general purpose method called SI-

MEX [8] will be used to account for measurement error. This method has been used in the related problem of nonparametric regression [10]. The idea behind SIMEX (which, for reasons that are made clear below, stands for simulation–extrapolation) is to estimate $\lambda(t)$ by extrapolating the behavior of $\lambda(t)$ when *additional* measurement error is added to the observations to the case where measurement error is *reduced* to 0. For $\gamma > 0$, let:

$$Y_j(\gamma) = Y_j + \gamma^{1/2}\sigma_j\eta_j \tag{5}$$

where η_j is a normal error with mean 0 and variance 1. The mean and variance of $Y_j(\gamma)$ are:

$$E(Y_j(\gamma)) = t_j \tag{6}$$

and:

$$\mathrm{Var}\, Y_j(\gamma) = (1+\gamma)\sigma_j^2 \tag{7}$$

respectively, so that, notionally at least, $Y_j(-1) = t_j$. Let $\lambda(t;\gamma)$ be the mean value of $\lambda(t)$ when the observed ages follow Eq. 5. Note that $\lambda(t;0) = \lambda(t)$. It is straightforward to estimate $\lambda(t;\gamma)$ for selected positive values $\gamma_1,\gamma_2,...,\gamma_m$ of γ by averaging $\lambda(t)$ over repeated simulations from Eq. 5. This is the simulation step of SIMEX. In the extrapolation step, a parametric model $f_t(\gamma)$ of the dependence of $\lambda(t;\gamma)$ on γ is fit to the sequence $\lambda(t;0)$, $\lambda(t;\gamma_1)$,...,$\lambda(t;\gamma_m)$ and the final estimate $\lambda_s(t)$ is found by extrapolating the fitted function to $\gamma = -1$. This is illustrated in the next section.

As noted, Culler et al. [1] estimated the rate of lunar impacts from the ages of spherules using an ideogram. This ideogram can be written:

$$\hat{\lambda}_I(t) \propto \sum_{j=1}^{n} \phi\left(\frac{t - Y_j}{\sigma_j}\right) \tag{8}$$

where ϕ is the standard normal probability density function. Although the form of this ideogram is similar to that of the kernel estimator (Eq. 2) with a normal kernel, it is actually quite different. The difference arises from the use of the magnitude of the measurement error, as reflected in σ_j, in place of the bandwidth h. As a result, the degree of smoothing of the ideogram is directly related to the level of measurement error. The observed ages $Y_1, Y_2, ..., Y_n$ reflect two sources of variation: variation in the underlying point process and measurement error. While $\lambda_S(t)$ accounts for both, $\lambda_I(t)$ accounts only for measurement error. To see this, consider the case in which there is no measurement error. In that case, $\lambda_S(t)$ reduces to the ordinary kernel estimator, but $\lambda_I(t)$ simply reproduces the data as a set of spikes at $Y_1, Y_2, ..., Y_n$. Even in the absence of measurement error, it seems extreme to estimate the impact rate as 0 everywhere except at these spikes. This is particularly true as the data almost certainly represent an incomplete sample of impact events. Culler et al. [1] acknowledged this problem with the admonition to ignore spikes in the ideogram associated with single, well-dated spherules. On the other hand, when measurement error is present, there is no reason why the degree of smoothing should increase with the magnitude of this error, which is, after all, not a feature of the underlying rate function. To put it another way, the problem that measurement error commonly causes is the attenuation or flattening out of variations. This effect is illustrated in the next section. Statistical methods, including SIMEX, that are designed for estimation in the presence of measurement error are intended to reduce this attenuation. In contrast, by increasing the degree of smoothing when measurement error is large, the ideogram tends to exacerbate it.

3. Application

The methods described in the previous section were applied to the data of Culler et al. [1]. The goal was to estimate the rate of lunar impacts over the period (0, 3500) using the ages of 155 spherules. Throughout this section, time is measured in millions of years (Myr) and the impact rate is measured in impacts/Myr. The data consist of the measured age Y_j and measurement error variance σ_j^2 for each of $n = 155$ impact spherules. These data can be found at the website http:www.sciencemag.org/feature/data/1044416.shl. The average value of the error standard deviation δ_j is around 238 Myr, corresponding to approximately 7% of

the estimation period. A bandwidth of 350 Myr, corresponding to 10% of the estimation period, was used in the kernel estimation. The boundary-adjusted kernel estimate $\lambda(t)$ is shown in Fig. 1. For this estimate, the boundary adjustment was made in the neighborhood of $t=0$, but, because the data set contains spherules with $t>3500$, no adjustment was made in the neighborhood of $t=3500$. The SIMEX estimate $\lambda_S(t)$ is also shown in Fig. 1. This estimate was based on extrapolating to $\gamma=-1$ the quadratic extrapolant:

$$f_t(\gamma) = \beta_{0t} + \beta_{1t}\gamma + \beta_{2t}\gamma^2 \quad (9)$$

fitted to $\lambda(t;\gamma)$ for $\gamma=0$, 0.5, 1, 1.5, and 2 by ordinary least squares. For each value of t and each value of these values of γ, $\lambda(t;\gamma)$ was estimated from 200 simulations.

The extrapolation step has been described as the Achilles' heel of SIMEX estimation [8]. In Fig. 2, the fitted extrapolant is shown for $t=0$, 1500, and 3200. For $t=0$ and 3200, the quadratic extrapolant appears to fit the estimated values of $\lambda(t;\gamma)$ rather well. The fit is somewhat worse for $t=1500$, although the overall attenuating effect of measurement error is clear. This suggests that the estimated magnitude of the local minimum of

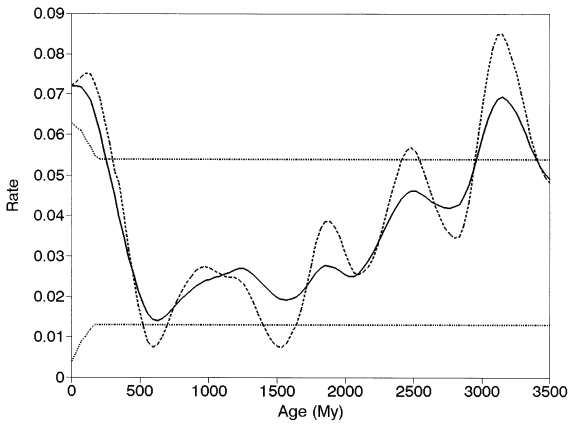

Fig. 1. The uncorrected kernel estimate $\lambda(t)$ (solid) and the SIMEX estimate $\lambda_S(t)$ (dashed) of the rate of lunar impacts over the past 3.5 billion years. Impact rate is measured in impacts/Myr. The dotted lines represent the upper and lower 0.05 quantiles of the distribution of $\lambda_S(t)$ when the true impact rate is constant.

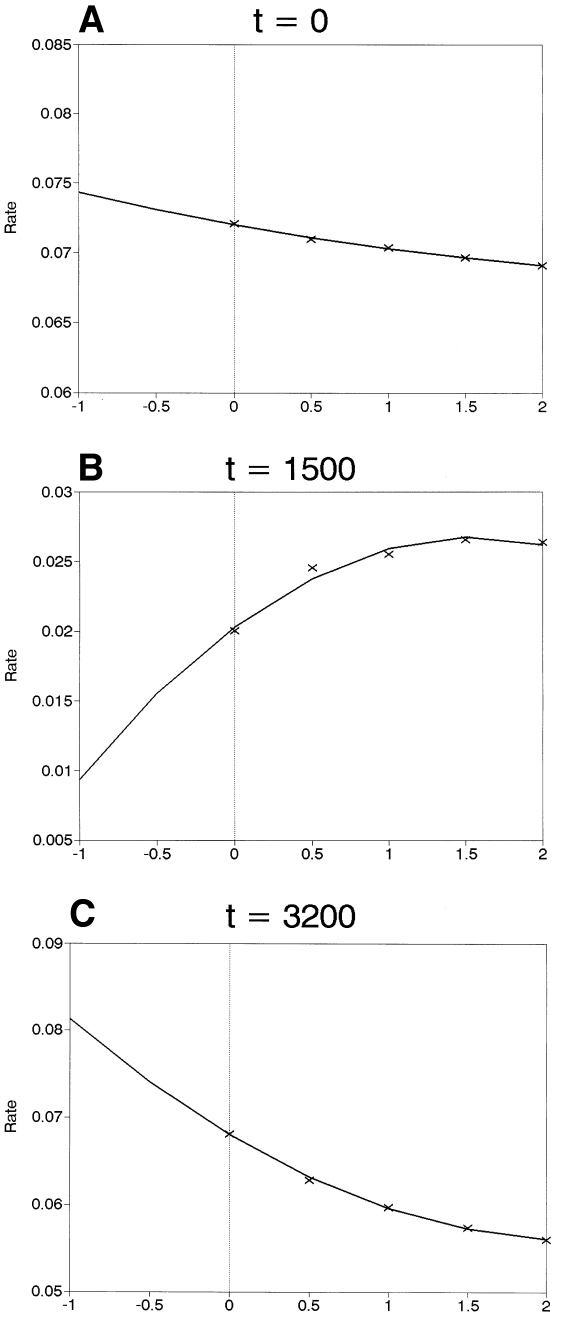

Fig. 2. The behavior of the quadratic extrapolant in estimating the impact rate at (A) $t=0$, (B) $t=1400$, and (C) $t=3200$. In each case, the crosses show the values of $\lambda(t;\gamma)$ or $\gamma=0$, 0.5, 1, 1.5, 2 and the solid curve shows the quadratic extrapolant fitted to these values by least squares and extrapolated to $\gamma=-1$.

$\lambda_S(t)$ near $t = 1400$ should be interpreted with caution.

To assist in interpreting $\lambda_S(t)$, significance bands were constructed under the assumption that the true rate function $\lambda(t)$ is constant by the following simulation procedure. A total of 155 true event times were distributed at random over the interval (0, 4700). A normal observation error was added to each of these times. The variance of this error was chosen at random without replacement from the 155 error variances reported in [1]. Under this model, the distribution of $\lambda_S(t)$ over repeated simulations is the same for all values of t larger than h. This distribution was estimated by finding the value of $\lambda_S(t)$ for a fixed value of t greater than the bandwidth of 350 Myr for each of 1000 data sets simulated as described above. The distribution of $\lambda_S(t)$ was found in the same way for selected values of t less than the bandwidth of 350 Myr. Significance bands given by the upper and lower 0.05 quantiles of these distributions are also shown in Fig. 1.

Turning to Fig. 1, the main effect of SIMEX estimation is to accentuate the quasi-periodic behavior of the estimated impact rate that is only weakly discernible in the uncorrected kernel estimate. As measurement error is known to attenuate variability, this result is in line with expectations. In overall terms, the estimate suggests that the Moon has experienced what could be called pulses of impact activity. The estimate exhibits several excursions outside the significance bands, indicating that this behavior is not due simply to random variations around a constant impact rate.

4. A simulation study

As the true rate of lunar impacts is unknown, it is not possible to determine from the results of the previous section how well the method of Section 2 performs. This section presents the results of a small simulation study of the performance of this method when the true rate function is known. Other results on the performance of kernel estimation and SIMEX estimation are covered in the references. The simulation proceeded in the following way. A realization of the non-stationary Poisson process with rate function:

$$\lambda(t) = 150 + 50\cos(4\pi t) \tag{10}$$

was simulated on the interval (0, 1.5) using the IMSL [11] FORTRAN subroutine RNNPP. The kernel estimator with bandwidth $h = 0.1$ was applied to the simulated data to estimate $\lambda(t)$ over the interval (0, 1), using the boundary correction near the lower boundary, but not at the upper boundary. Next, simulated normal measurement error with standard deviation $\sigma = 0.07$ was added to each simulated event time and the uncorrected kernel estimate $\lambda(t)$ and the corrected estimate $\lambda_S(t)$ were found for the data with measurement error. The procedure was repeated 200 times and the averages over these repeated simulations of $\lambda(t)$ and $\lambda_S(t)$ are shown in Fig. 3, along with the true rate function $\lambda(t)$ and the average of the kernel estimates for the data without measurement error. Note that the parameters used in this simulation were selected to correspond roughly to the data analyzed in the previous section: namely, a mean of 150 events, a bandwidth of 10% of the estimation interval, and measurement error with standard deviation 7% of the estimation interval. The attenuation in the uncorrected kernel esti-

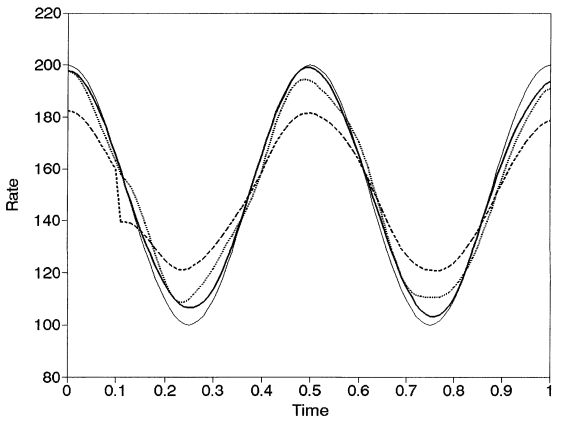

Fig. 3. True rate function $\lambda(t)$ (thin solid line) used in the simulation study along with the mean value of the kernel estimate applied to the true event times (thick solid line); the mean value of the uncorrected kernel estimate applied to the event times measured with error (dashed line); and the mean value of the corrected estimate (dotted line).

mate $\lambda(t)$ is clear in Fig. 3. It is also clear that, in this case, SIMEX works reasonably well at reducing this attenuation, in the sense that the mean of $\lambda_S(t)$ is close to the mean of the kernel estimate for the true event times. Although it is not shown in Fig. 3, in this case, the ideogram would exhibit even more attenuation than the uncorrected kernel estimate. This follows because, in this case, the ideogram is equivalent to the kernel estimate with a Gaussian kernel with bandwidth 0.07, which corresponds roughly to a bisquare kernel with bandwidth 0.14.

5. Discussion

Situations in which radiometric ages of objects are used to reconstruct the history of physical and other processes are pervasive in the earth sciences. It is widely recognized that such ages include measurement error and considerable effort is undertaken to assess the magnitude of this error. However, the effects of measurement error on estimation have been under-appreciated and methods for dealing with these effects have been under-utilized. The approach proposed here combines two modern statistical methods: nonparametric kernel estimation of the rate function of the point process of lunar impacts with SIMEX estimation to correct for the attenuating effect of measurement error. The simulation results presented here, although limited, confirm the usefulness of the method at reducing the attenuating effect of measurement error on the estimation of the rate function. The results of applying the method to the measured ages of impact spherules suggests that the lunar surface has undergone quasi-periodic pulses of impact activity.

Acknowledgements

The helpful comments of Raymond Carroll, Jennifer Grier, and Paul Renne are acknowledged with gratitude.***[BOYLE]***

References

[1] T.S. Culler, T.A. Becker, R.A. Muller, P.R. Renne, Lunar impact history from $^{40}Ar/^{39}Ar$ dating of glass spherules, Science 287 (2000) 1785–1788.
[2] D. Cox, V. Isham, Stochastic Point Processes, Chapman and Hall, London, 1980, 220 pp.
[3] P.J. Diggle, A kernel method for smoothing point process data, Appl. Stat. 34 (1985) 138–147.
[4] P.J. Diggle, J.S. Marron, Equivalence of smoothing parameter selectors in density and intensity estimation, J. Am. Stat. Assoc. 83 (1988) 793–800.
[5] A. Cowling, P. Hall, M.J. Phillips, Bootstrap confidence regions for the intensity of a Poisson point process, J. Am. Stat. Assoc. 91 (1996) 1516–1524.
[6] J. Fan, Y.K. Truong, Nonparametric regression with errors in variables, Ann. Stat. 21 (1993) 1900–1925.
[7] T. Hastie, R. Tibshirani, Generalized Additive Models, Chapman and Hall, London, 1990, 335 pp.
[8] J.R. Cook, L.A. Stefanski, Simulation-extrapolation estimation in parametric measurement error models, J. Am. Stat. Assoc. 89 (1994) 1314–1328.
[9] A.R. Solow, Model-checking in non-stationary Poisson processes, Appl. Stoch. Models Data Anal. (1992) 129–132.
[10] R.J. Carroll, J.D. Maca, D. Ruppert, Nonparametric regression in the presence of measurement error, Biometrika 86 (1999) 541–554.
[11] IMSL, FORTRAN Subroutines for Statistical Analysis, IMSL, Houston, TX, 1991, 1579 pp.