Curve Height at Event Time Analysis

T. L. Graves
Statistical Sciences Group
MS F600, Los Alamos National Laboratory
Loa Alamos, NM 87545

J. S. Marron
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260

A. Mockus
Software Production Research Department
Bell Laboratories, Lucent Technologies
263 Shuman Boulevard, Room 2F-319
Naperville , IL 60566-7055

A. F. Glazner
Department of Geological Sciences
University of North Carolina
Chapel Hill, NC 27599-3315

Abstract

Curve Height at Event Time Analysis is proposed to assess the statistical significance of the relationship between event times of a point process and a given curve. The method is illustrated with examples from software engineering and from geology. In the first example, the events are the finding of software faults, and their rate is seen to be connected to a curve which measures the effect of past perfective maintenance. In the second example, the events are volcanic activity, and this is shown to be connected to periods of low glaciation.

1 Introduction

This paper proposes Curve Height at Event Time Analysis (CHETA), for evaluating the statistical significance of connections between the height of a curve that changes over time, and the times of occurrence of events. The general problem is motivated by questions from two diverse areas of science, which are illustrated in Figures 1 and 2. The CHETA approach to solving this problem

is developed in Section 2. Details of the applications, and positive conclusions for the motivating problems, are presented in Sections 3 and 4.

The data in Figure 1 are from software engineering, in particular from the change management history of part of a large software system. See Eick et. al. (1999) for discussion of many interesting aspects of these data. The problem addressed here is the effectiveness of "perfective maintenance". This type of change to the system was considered in Swanson (1976). We define a perfective change to be one which neither adds new functionality to the system nor fixes a known fault. Examples of such changes include adjustment to meet changing code standards, and re-organization with the goal of making the code easier to understand/maintain or to facilitate anticipated future changes. The value of such changes is an important question in software engineering, because they consume development resources, without producing a revenue generating product. Here the impact of perfective maintenance on the rate of faults is investigated.

The amount of perfective maintenance varies over time, because of changing system demands and also because of changing management policy. The (continually changing) proportion of resources devoted to perfective maintenance can be measured by the proportion of such changes in a moving time window. An answer to the question of window size and shape, that is specific to this type of data, is given in Section 3. The result, for the data set at hand, is shown as the solid Proportion of Perfective Maintenance Curve in Figure 1.

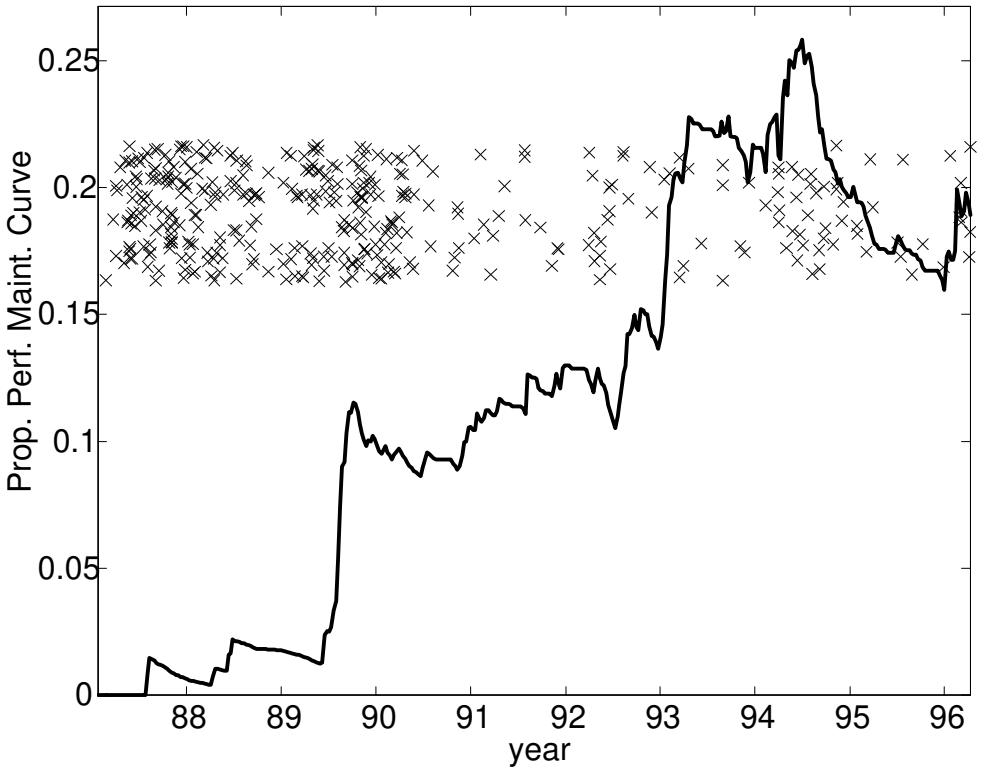

FIGURE 1: *First CHETA example. "Curve" is the Proportion of Perfective Maintenance shown with solid linetype. "Events" are the times of faults, shown as jittered x marks.*

To understand how the Proportion of Perfective Maintenance affects the rate of occurrence of faults, this curve is compared with the times of "events", i.e. changes made to fix faults (improper operation) in the software shown as $\times$ marks, that have been jittered (i.e. given a random vertical coordinate for easy visualization), in Figure 1. It appears that there are more faults at times when the Proportion of Perfective Maintenance curve is lower, but it is not clear whether this is a sampling artifact or else there is in fact an important connection between the curve and the events. This question is approached by evaluating the curve at each of the events (i.e. by "plugging the events into the curve") and then studying the resulting distribution on the vertical axis. CHETA assesses the statistical significance of the connection between the events and the curve by comparison with an appropriate corresponding null distribution. The general method is developed in Section 2. This is applied to show a significant connection between perfective maintenance and reduction of software faults, together with discussion of complicating factors and relevant details, in Section 3.

The geological data in Figure 2 suggest a connection between glaciation and volcanism. This time the "curve" is a smoothed version of the SPECMAP curve, which is a surrogate for the amount of global glaciation, discussed in more detail in Section 4. The events, again shown as jittered $\times$ marks, are times of volcanic activity in the Owens Valley region of California. The visual connection is less obvious than for Figure 1, but a careful look suggests that these events tend to occur more frequently near the valleys of the SPECMAP curve, i.e. in periods of reduced glaciation between ice ages. This is easier to see from the thin dashed curve, which is a kernel density estimate (see e.g. Wand and Jones (1995) for an introduction and discussion) that is high in regions where the data are more dense. The density estimate has peaks that often coincide with the valleys of the SPECMAP curve, McIntyre (1989). As for the software engineering example, an important question is whether this could be explained as expected random variation, or represents a systematic connection between glaciation and volcanism. In Section 4 it is seen that the CHETA method shows that this connection is statistically significant. This association was announced by Glazner, et. al. (1999), who also speculated that the mechanism behind this phenomenon could be that the increased weight of glaciers pressing down on mountains tends to inhibit volcanic eruptions. Further details of this analysis are discussed in Section 4.

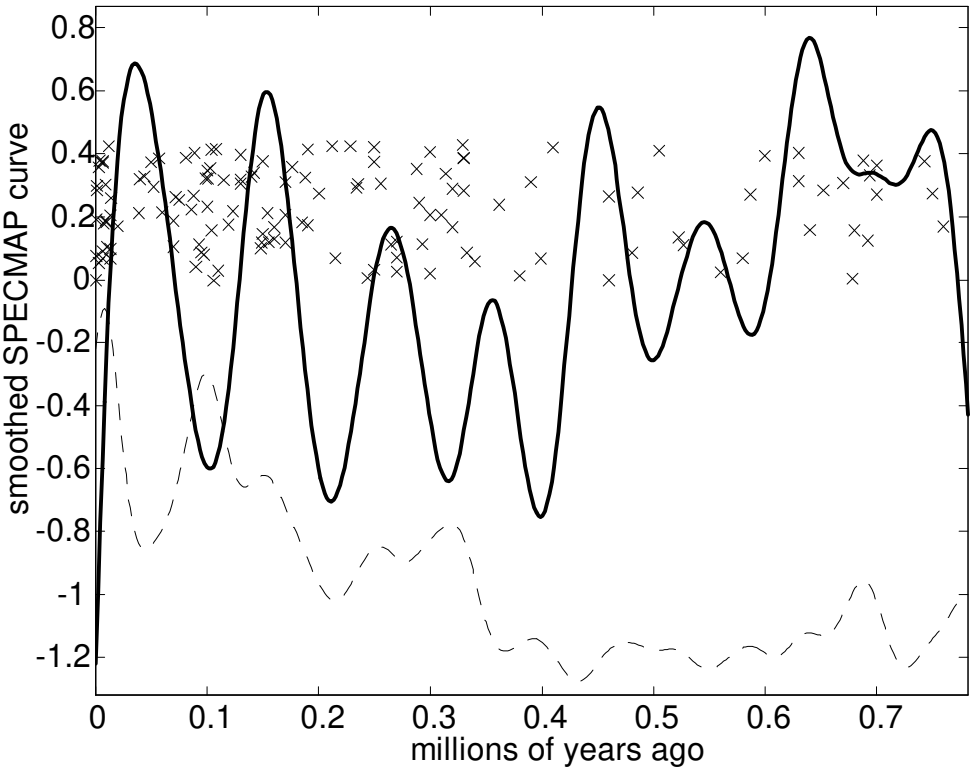

FIGURE 2: *Second CHETA example. "Curve" is smoothed SPECMAP shown with solid linetype. "Events" are the times of volcanic activity, shown as jittered x marks. Times of increased volcanic activity are highlighted with a kernel density estimate, shown with dashed linetype.*

The development of the CHETA method in this paper raises some interesting open problems in mathematical statistics that are discussed in Section 5.

2 The CHETA approach

The idea behind the CHETA approach to finding the statistical significance of the connection between a curve and occurrences in time is illustrated in Figure 3. The curve and the data shown there are a made up toy example. As in Figures 1 and 2, the event times of interest are shown as $\times$ marks on the horizontal axis. These times are plugged into the curve, and their image is shown as corresponding $\times$ marks on the vertical axis. Because there are more $\times$ marks where the curve is low, the distribution of the images on the vertical axis is more concentrated towards the lower values. This distribution is represented by a kernel density estimate, drawn vertically with solid line type. While the lower peak appears "bigger", some frame of reference is needed for calculation

of statistical significance. This comes from repeating the evaluation for a set of "background events", shown as circles in Figure 3. In some situations the background events could be simulated from the uniform distribution. More complicated "background distributions" are appropriate for the examples of Figures 1 and 2, which are discussed in Sections 3 and 4 respectively. The images obtained by evaluating the curve at the background times are shown as circles on the vertical axis. The distribution of these, shown as the dashed vertical kernel density estimate, has a more balanced number of high and low points than the corresponding main event population.

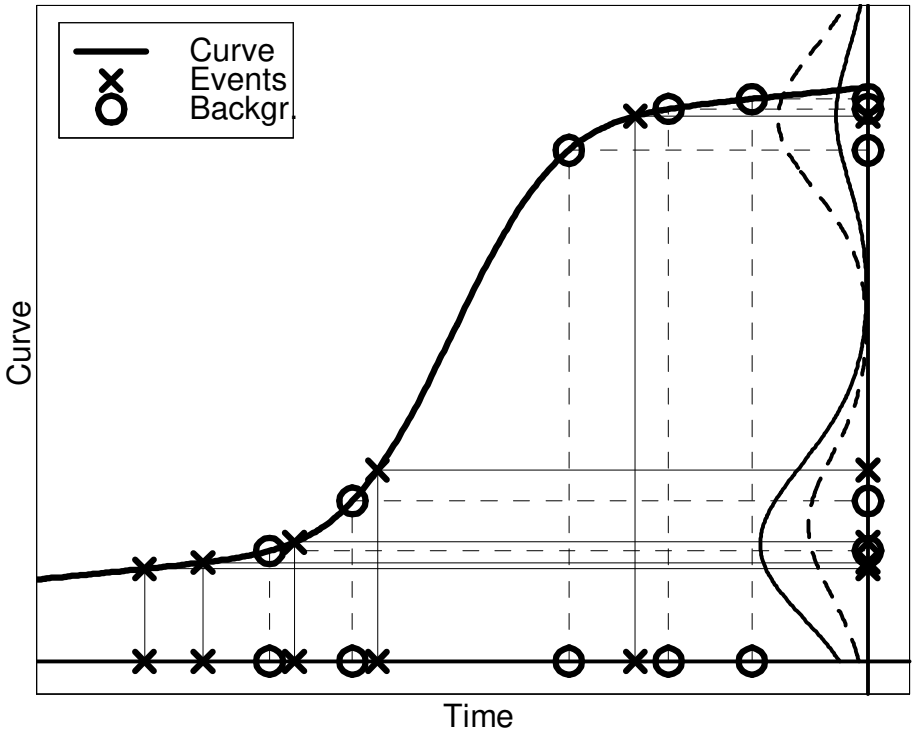

FIGURE 3: *Toy example illustrating Curve Height at Event Time Analysis. Original data points are on the horizontal axis, and their curve evaluation images are on the vertical axis. Relative densities of image points are shown by vertical kernel density estimates.*

CHETA investigates the connection between the curve and the events, by investigating the difference between these two image populations. Statistically significant differences between these populations imply that the event times are not distributed as expected according to the "background" distribution. Furthermore p values for the differences between these populations are also p

values for assessing the strength of the evidence against the null hypothesis of no connection between the event times and the curve.

The particular test of difference of distribution could be an omnibus type test, such as the Kolmogorov-Smirnov test or the Cramér-von Mises test; see Stephens (1974) for comparison of these and some related alternatives. However, for most applications of CHETA, including those of Figures 1, 2 and 3, it is anticipated that events will occur more often at times where the curve is high, or else more often where it is low. Thus more powerful inference will often be based on traditional two sample mean tests. This method is used exclusively in the next sections.

A Matlab implementation of CHETA is available from J. S. Marron (email: `marron@stat.unc.edu`) on request.

3 CHETA in Software Engineering

For the CHETA application of Figure 1, the construction of the Proportion of Perfective Maintenance curve is first considered. This could be done by using the proportion of perfective changes in a moving window, but this requires specification of the window width. This can be generalized to a smooth window shape by taking the ratio of kernel density estimates (this is equal to the proportion inside the window when the uniform kernel is used), but then both the window shape and the window width need to be chosen. A natural solution to this problem in the present case comes from the work of Graves, et. al. (1999) who studied the impact of a given change to the system, in terms of its potential for later faults. They showed that an exponential decay model is reasonable for the present software system. In particular, letting FR denote the fault rate at time t, measured in years,

$$FR(t) \propto \sum_{i=1}^{n} e^{-c(t-t_i)} 1_{\{t_i<t\}}, \tag{1}$$

where t_i is the time of the i th change made to the system, where the indicator function is

$$1_{\{t_i<t\}} = \begin{cases} 1 & t_j < t \\ 0 & t_j \geq t \end{cases},$$

and where $c = \frac{-11}{12 \log(0.5)}$ (which reflects a half life of exponential decay of about 11 months). This assumes that the effect that an adaptive change has on potentially increasing the fault rate of software, dies off at the same rate as the effect that a perfective change has on decreasing the fault rate. Note that $FR(t)$ is proportional to a kernel density estimate, whose bandwidth is determined by c, and whose kernel is the exponential density. This asymmetric kernel is very unusual in classical density estimation, but is quite appropriate here, because the impact of a given change in the code happens only after the change is made. Furthermore, the problem of choosing the window width is also solved since c

has already been determined. Thus, the Proportion of Perfective Maintenance curve shown in Figure 1, is

$$PPM(t) = \frac{PFR(t)}{FR(t)},$$

where $FR(t)$ is given in (1), and where PFR is the similar quantity, but summed only over the changes that were perfective.

The application of CHETA to the data of Figure 1 also requires choice of the background events. It is not appropriate to choose these randomly, i.e. from the uniform distribution, since changes to this part of the software system do not occur homogeneously in time. Instead there was a large burst of activity early in the development process, followed by occasional spikes as significant new features were added to this part of the system. Thus it is sensible to choose the background events to have this same distribution. One way of doing this is to draw a sample from the total set of changes. The results of this, and the corresponding CHETA analysis are shown in Figure 4.

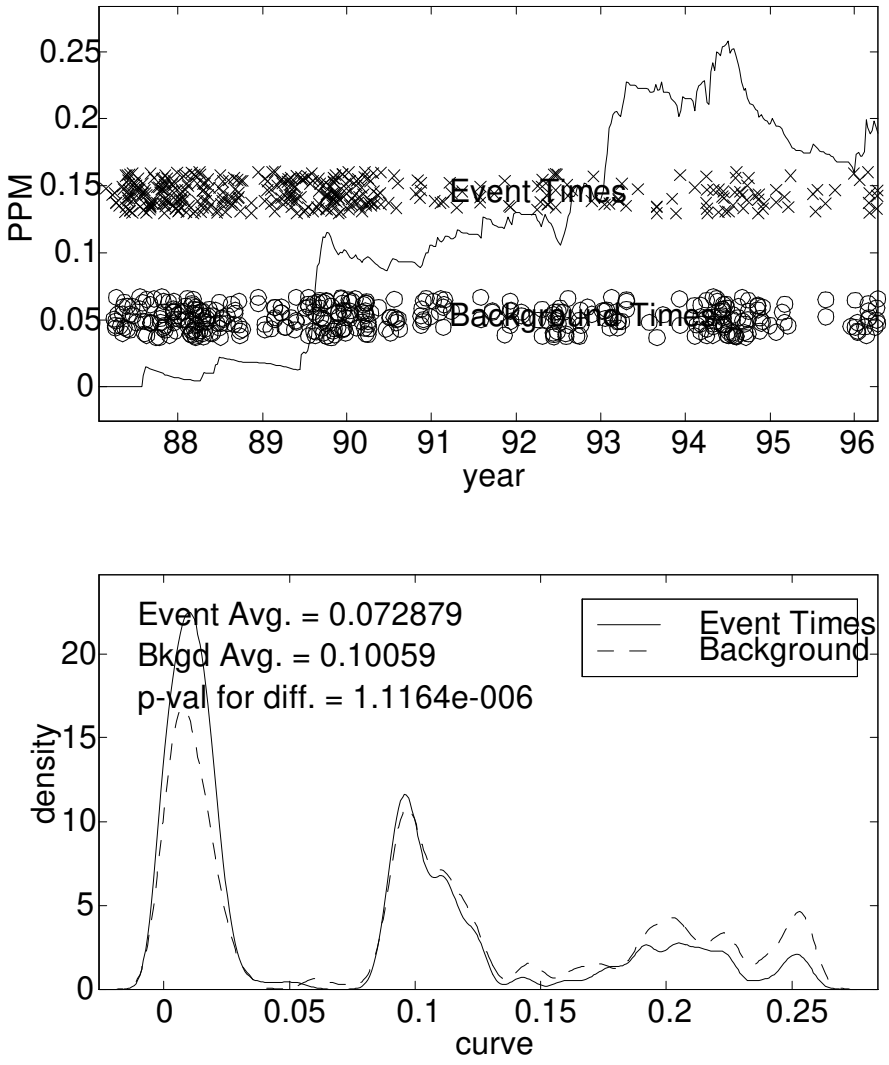

FIGURE 4: *Full CHETA analysis of software engineering data. Shows statistical significance of connection between perfective maintenance and fault rate.*

The non-homogeneity of the sample of background times in the upper panel of Figure 4 shows clearly that a uniform background distribution is not appropriate. Furthermore, while studying only the PPM curve and the event times, as in Figure 1, it appears that many events occurred in 1994, which is inconsistent with the idea that fewer events occur when the PPM curve is high. However the plot of background times shows that there was a high degree of overall activity at this time. The CHETA analysis in the lower panel of Figure 4 shows how these effects can be properly combined and understood. As in

Figure 3, kernel density estimates of the images of both types of event times are shown, but this time in the conventional horizontal fashion (not vertically as in Figure 3). The solid (event time) kernel density estimates are higher on the left, and lower on the right, suggesting that faults tend to occur more frequently when the PPM curve is lower. The strength of this is assessed in terms of statistical significance, by a simple 2 sample mean test of the null hypothesis $H_0 : \mu_{events} > \mu_{background}$. This hypothesis is rejected with p value $\approx 10^{-6}$, i.e. the evidence in favor of a connection between a high level of perfective maintenance and a low fault rate is very strong.

A limitation of CHETA analysis is that it shows only association, but not causation. For example, alternate explanations of the statistical significance observed above can be given in terms of expected features of the development process. In particular, early in the life of software, it is expected that bugs will be found simply because the software is young. At same time, during that phase of the development, there will be naturally less need for perfective maintenance. To eliminate this possibility, the analysis was repeated, using only the part of the data more than three years into the life of the software, i.e. everything after the first big step in the PPM curve. No picture is shown, because it is not much different from Figure 4. Here the p value is 0.031, still statistically significant. Hence these time effects explained only part of the above observed relationship. We suspect the rest is actually a causation effect, but different methods are needed to establish this.

4 CHETA in Geology

Here is the detailed analysis of geological data from Figure 2. The SPECMAP curve of McIntyre (1989) records oxygen isotope variations in planktonic foraminifera. These oxygen isotopes are assumed to correlate with global ice volume as noted in Shackleton (1987). The raw SPECMAP curve has higher frequency components, which do not visually connect with volcanic activity. These components are smoothed out using a Gaussian kernel smooth, with bandwidth $h = 0.02$, which gives the solid curve shown in Figure 2. While that showed the visual connection that motivated this work, the biases induced by smoothing are eliminated by using the raw SPECMAP curve in the analysis shown in Figure 5.

Details of the gathering and dating of the volcanic data are given in Glazner, et. al. (1999). The dashed curve in Figure 1 is a Gaussian kernel density estimate, with bandwidth $h = 0.18$, chosen to visually maximize the apparent correlation with the smoothed SPECMAP curve.

The CHETA analysis for these data is shown in Figure 5.

As in Section 3, the uniform distribution is not appropriate for choosing the "background events". For this geological data, this is because evidence of more recent volcanic activity is easier to find. The exponential distribution

$$f_\theta(x) = \theta e^{-x\theta}, \ x > 0,$$

is a reasonable model for the increasing difficulty of finding volcanic evidence. But the parameter θ still need still needs to be determined. The conventional maximum likelihood estimate, $\widehat{\theta} = 1/\overline{X}$, is inappropriate, since the data are truncated at $T = 0.782$ million years ago, as shown in Figure 2. For a truncated exponential model

$$f_{\theta,T}(x) = \frac{\theta e^{-x\theta}}{1 - e^{-T\theta}},\ 0 < x < T,$$

maximum likelihood estimation has no closed form, but we developed a standard Newton's method iterative solution (details available from J. S. Marron), which converged quickly to a reasonable answer $\widetilde{\theta}$. Background values were simulated from the $f_{\widetilde{\theta},T}(x)$ distribution.

The resulting CHETA analysis is shown in Figure 5.

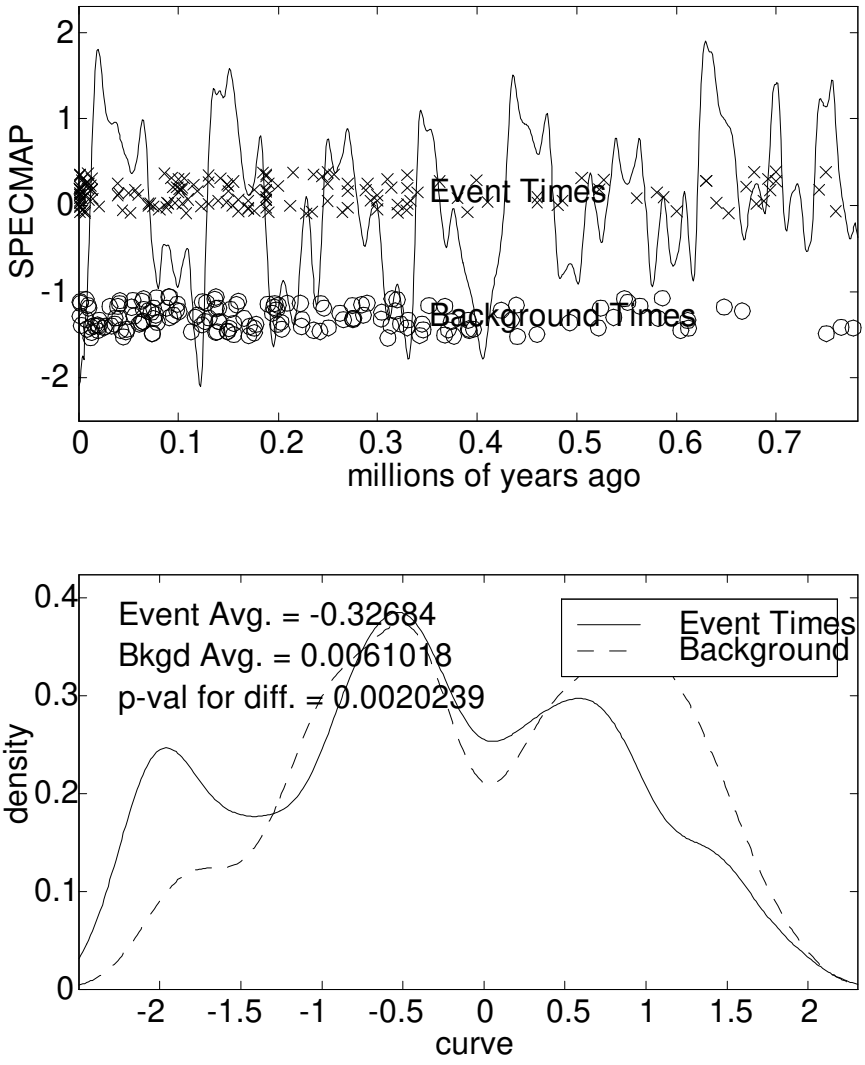

FIGURE 5: *Full CHETA analysis of geological data. Shows statistical significance of connection between glaciation and volcanism.*

Again the lower panel compares the two distributions which result from evaluating the curve at the two sets of points. The kernel density estimates show that the evaluated event times tend to occur more often when the SPECMAP curve is low. Statistical significance is assessed via the simple two sample mean test of the null hypothesis $H_0 : \mu_{events} > \mu_{background}$. Again there is a strongly conclusive result, i.e. strong evidence for the mean of the evaluated events being smaller, quantified by the p value of 0.002. This gives the conclusion of a statistically significant connection between glaciation and volcanism.

5 Open problems

The CHETA method is seen here to be a useful method of data analysis. This motivates a mathematical statistical study of the properties of the method. Here are some conjectures:

- When the true underlying curve is fixed and known (as in the geological example), the size of the test is asymptotically correct.
- When the true underlying curve is unknown, and estimated by a smoothing method (as in the software engineering example), the size of the test is asymptotically correct, when the smoothing is done in a consistent manner.
- The bandwidth problem for smooth estimates of the underlying curve can be solved by optimizing the size of the test. A source for ideas of this type is the "confidence interval coverage optimization" methods developed in a series of papers described in Chapter 4 of Hall (1992).

Another open problem is how many background events to use. In these examples, we took this to be the same as the number of regular events, because this gives the best visual impression. However, since these values are simulated, an arbitrary number could be used. When a very large number is used, the sampling error in the background densities is negligible, which will result in a more powerful test.

References

[1] Eick, S. G., Graves, T. L., Karr, A. F., Marron, J. S. and Mockus, A. (1999) Does code decay? Assessing the evidence from change management data, to appear in *IEEE Transactions on Software Engineering.*

[2] Glazner, A. F., Manley, C. R., Marron, J. S. and Rojstaczer, S. (1999) "Fire or ice: anticorrelation of volcanism and glaciation in California over the past 800,000 years", *Geophysical Research Letters*, 26, 1759-1762

[3] Graves, T. L., Karr, A. F., Marron, J. S. and Siy, H. (1999) Predicting fault incidence using software change history, to appear in *IEEE Transactions on Software Engineering.*

[4] Hall, P. (1992) *The Bootstrap and Edgeworth Expansion*, Springer Verlag, New York.

[5] McIntyre, A., Ruddiman, W. F., Karlin, K. and Mix, A. C. (1989) Surface water response of the equatorial Atlantic Ocean to orbital forcing, *Paleoceanography*, 1, 137-162.

[6] Stephens, M. A. (1974) EDF Statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association, 69*, 730-737.

[7] Shackleton, N. J. (1987) Oxygen isotopes, ice volume and sea level, *Quaternary Science Reviews*, 6, 183-190.

[8] Swanson, E. B. (1976) The dimensions of maintenance, *2nd Conference on Software Engineering*, San Francisco, CA, 492-497.

[9] Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*, Chapman and Hall, London.