Survival in Seattle: Earthquake recurrence intervals and time-dependent earthquake likelihoods from non-parametric survival analysis of paleoearthquakes in the Puget Lowland (WA, USA) and San Andreas Fault (CA, USA)

Richard Styron
Earth Analysis
richard.h.styron@gmail.com

Kate Scharer
US Geological Survey

Brian Sherrod
US Geological Survey

Abstract

abstract

1 Introduction

The characterization of earthquake occurrence in time is of obvious importance in seismic hazard analysis, and is also of substantial interest in the insight that it yields into the physics of the earthquake process: Different types of earthquake recurrence behaviour imply different types of loading, triggering and unloading (**reword**) of faults or fault systems (e.g. Faenza et al., 2003) (**say something about interaction perhaps**). These

The majority of studies seeking to characterize recurrence behavior or to make probabilistic earthquake forecasts based on recurrence behavior use standard, parametric probability distributions to represent recurrence. These may be generic distributions with widespread cross-disciplinary use, such as the normal, lognormal, Weibull and exponential distributions. Or they may be distributions that are specific to earthquake science, perhaps with some physical justification as well; the most prominent of the latter category is the Brownian Passage time (BPT) model (**???**, **???**; Matthews et al., 2002). Studies that incorporate empirical recurrence observations (from instrumental, historical or paleoseismological

records) typically fit the parameters of the recurrence distribution to the data using maximum-likelihood or other optimization routines [e.g., **REFS**] and may draw many samples from the earthquake timing or recurrence probability estimates to propagate the uncertainty in the data to the parameter estimates [e.g., **REFS**].

However, it is entirely possible and in some instances more appropriate to characterize recurrence non-parametrically. There are several reasons for doing so. A primary motivation is that the recurrence distributions listed above are typically applied to individual faults that may operate independently and may incorporate some manner of renewal (i.e. stress reloading), rather than populations of faults that may interact. Though the recurrence behavior of a given fault may be well characterized by, say, a BPT model, the statistical properties of a population of BPT faults is unlikely to be as well characterized by a BPT distribution. The major exception is the exponential distribution (the Poisson model) which depends only on the mean rate of events, and therefore can accommodate any number of faults but cannot account for fault reloading or interaction. Additionally, a researcher may simply wish to operate without specifying a particular model; nonparametric analysis of time-to-event data is extremely common across a wide range of social and natural sciences.

In this work, we demonstrate methods to characterize and analyze the interevent (i.e. recurrence) times of faults and fault zones based on their paleoseismic histories. The recurrence distributions are constructed by Monte Carlo sampling of the empirical paleoearthquake age distributions, with and without constraints from stratigraphic ordering. Then, we illustrate the use of the statistical techniques of *survival analysis* to calculate the time-dependent earthquake rates, the expected times to the next event, and related quantities, and we show how those quantities may be updated given the elapsed time since the last event.

We do not claim that for individual points or segments of a fault, the empirical recurrence distributions as derived here are more accurate characterizations of fault behavior than the aforementioned parametric models. In our methods, we embed the assumption that the recurrence behavior of these faults based on a small number of events ($n \leq 30$) is sufficient to fully characterize their behavior. A particular weakness is that they lack the capability to describe recurrence intervals that are longer than those observed. The situation is different for fault systems, as the exponential model is the only one that is capable of dealing with multiple faults; therefore, our methods are self-evidently superior in that they are applicable.

Regardless, the methods presented here are general, flexible and powerful, and may be applied to a range of problems in geoscience.

1.1 Recurrence models

need a SSAF map

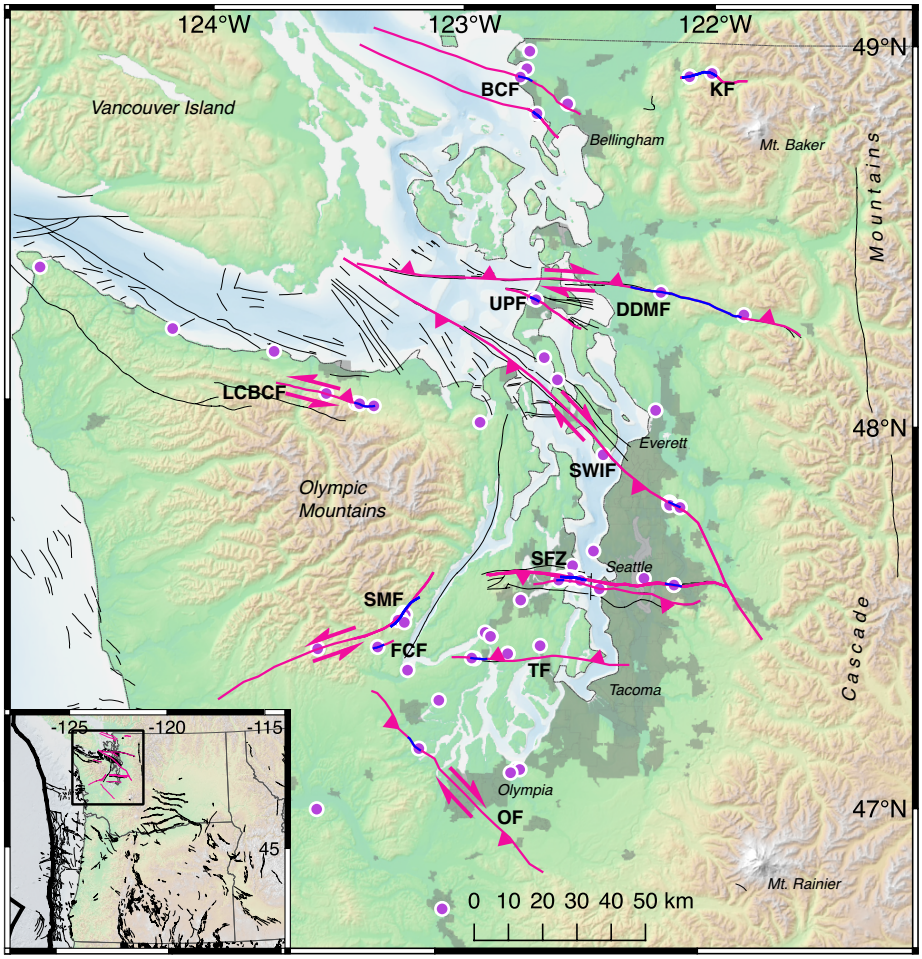

Figure 1: Map of fault traces in the Puget lowlands hosting the earthquakes analyzed here.

2 Earthquake geology of the Puget Lowlands and San Andreas

2.1 Puget Lowlands

2.2 San Andreas

3 Empirical recurrence model construction

We construct empirical earthquake recurrence models in a straightforward manner:

1. Draw a set of N random samples directly from each of the paleoearthquake age PDFs through an inverse transform sampling algorithm.
2. Make N sequences of earthquakes by ordering one sample per earthquake from each of the earthquake age sample sets (the nature of this ordering depends on availability of relative time constraints and will be discussed below).
3. Calculate the interevent times (i.e. the time interval between successive events) for all of the earthquakes in each of the N sequences.
4. Concatenate the interevent times, sort them, and create a recurrence (interevent) PDF through a kernel density estimation of the interevent times.

3.0.1 Sorting vs culling based on geologic relationships

An important consideration here is the presence of stratigraphic constraints on paleoearthquake ordering when earthquake age PDFs overlap. If the earthquakes have no stratigraphic or other geologic ordering constraints (e.g., they are from different trenches), each sequence of earthquakes can simply be sorted and then differenced to obtain the interevent time sequence. However, if the ages overlap but Earthquake A is demonstrably older than the next (for example through superposition of syn-earthquake colluvial wedges), then the sample age sequence cannot simply be sorted, because some sample earthquake sequences would place the date for the older Earthquake A *after* the age for the younger Earthquake B. In this situation, we instead reject sample earthquake sequences that are not in geologic order.

In the Puget Lowland, this is not a concern: The earthquakes from each individual fault trace are well-separated in time, and no stratigraphic or cross-cutting relationships are known to constrain relative ages of temporally overlapping events on different faults, so these can be treated independently.

However, geologic (mostly stratigraphic) relative age constraints are strong in the San Andreas data. For these analyses, we cull the sample sets that violate relative age constraints.

3.1 San Andreas recurrence

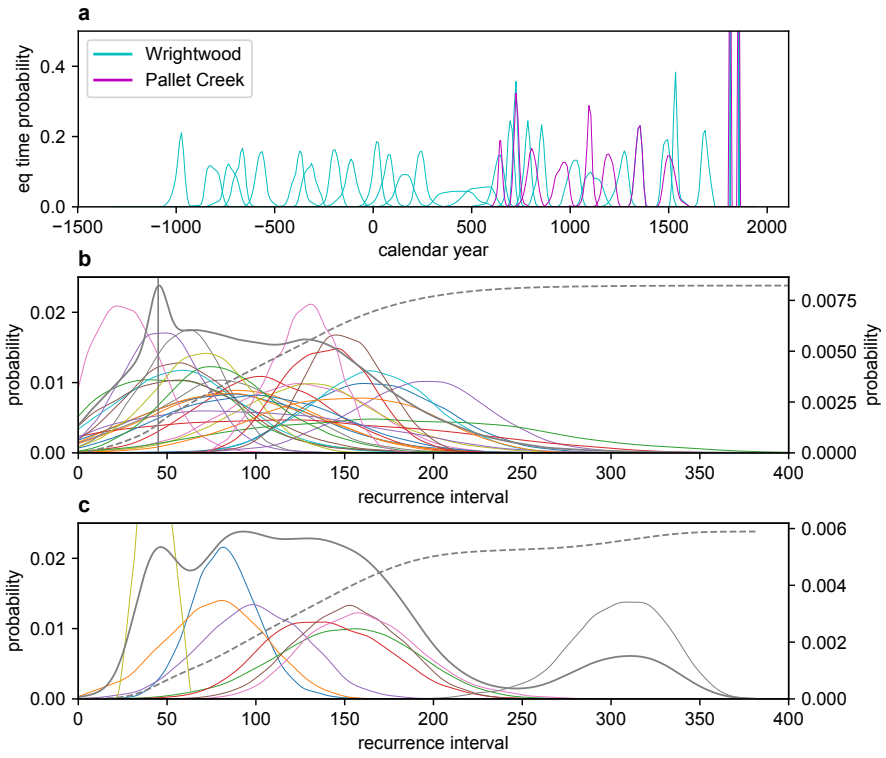

Figure 2: Earthquake timing and recurrence on the San Andreas fault.

3.1.1 Wrightwood

Wrightwood mode: (45.749148385884872, 0.0078220103914044825) WrighPuget median: 98.53337529070336 WrighPuget mean: 104.332002507

3.1.2 Pallet Creek

Pallet Creek mode: (90.50480533481911, 0.0057850008225488313) Pallet Creek median: 122.22446429602063 Pallet Creek mean: 134.510949648

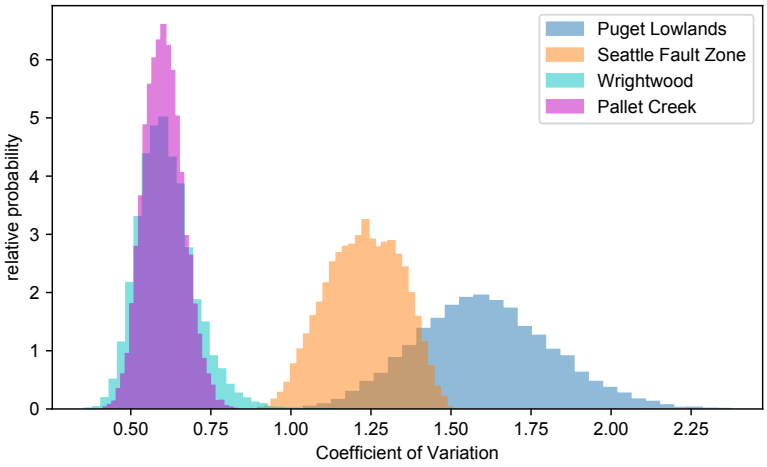

Figure 3: Histograms of oefficients of variation for fault systems considered here. Histograms are normalized, but each are based on 10,000 data points.

3.2 Puget Lowlands recurrence

Earthquakes in the Puget Lowland are dispersed throughout the past 16 ka, with most of the earthquakes happening in the past 4000 years (Figure 4a). Given the considerable overlap in the age PDFs (especially for the older events), the earthquakes appear to be grouped in time, with the oldest group at 14-10 ka, another group at 9-6 ka, and a final group from 4 ka to the present. Preliminary research suggests this represents earthquake clustering (Styron & Sherrod, 2016), perhaps due to static stress triggering, though more thorough work is ongoing.

Regardless of the nature of the clustering, the proximity of many of the earthquakes leads to a recurrence distribution that has a high x-intercept (i.e., immediately following an earthquake), and climbs steeply to a very peaked mode at 60 years. It then decays monotonically, in a quasi-exponential fashion. The median recurrence interval is 226 years, and the mean is 486 years.

The shape of this empirical recurrence distribution is quite different than a normal or log-normal distribution, as is often assumed for individual faults (particularly faults though to have 'characteristic'-type rupture behavior, with regular cycles of strain accumulation and release). This is not surprising, because the recurrence distribution shown represents aggregate behavior from fourteen fault zones, some with multiple strands capable of rupturing independently, such as the Seattle Fault Zone. Therefore, this recurrence distribution is more a reflection of fault interaction than of the specifics of the earthquake cycle on an individual

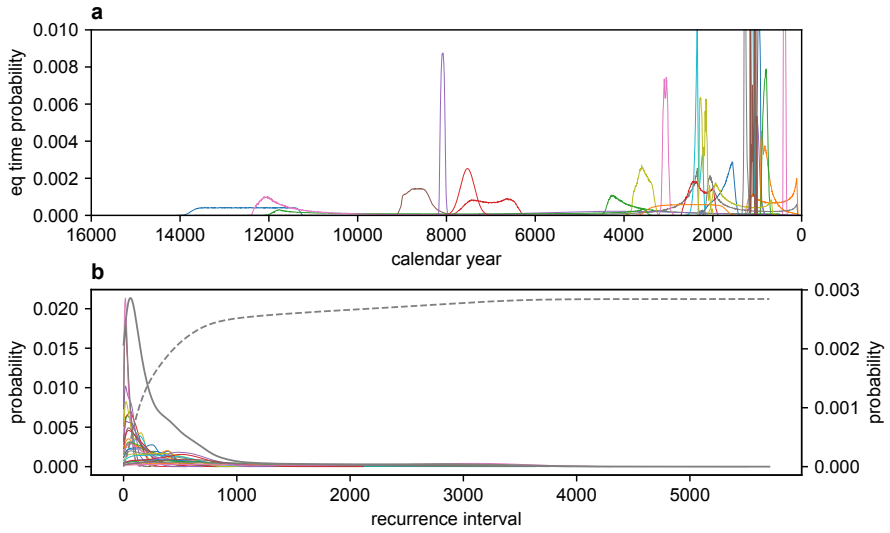

Figure 4: Earthquake timing and recurrence intervals in the Puget Lowlands. **a**: Time probabilities for each earthquake from the OxCal modeling. **b**: Interevent times for each set of consecutive earthquakes (thin lines) and kernel-density estimate for all interevent times (thick line). Note that the left *y*-axis scale corresponds to the individual interevent PDFs, while the right *y*-axis scale corresponds to the total kernel density estimated recurrence PDF.

fault.

This recurrence distribution indicates that following an earthquake, the probability of the next earthquake is high, and this probability stays high for several decades before declining. (Quantification of this will be done through survival analysis in Section 4).

3.2.1 Seattle Fault Zone

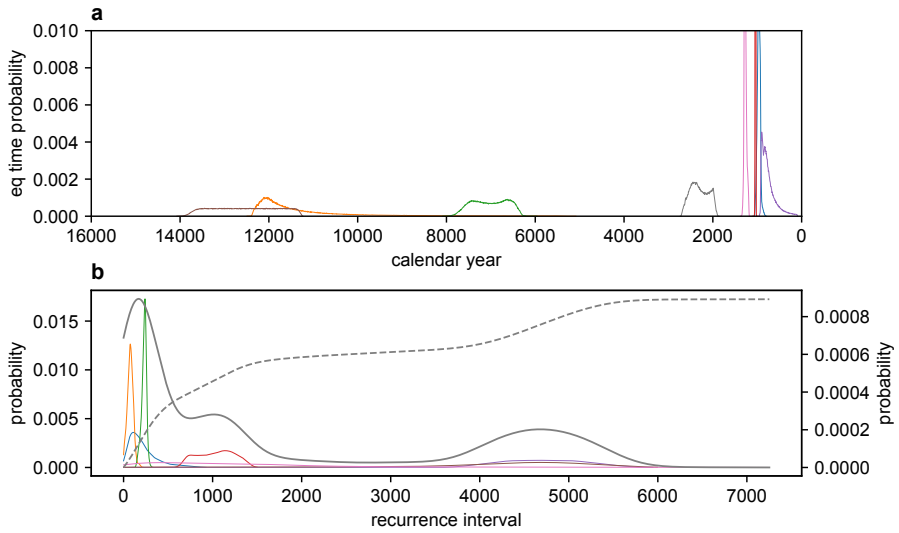

Figure 5: Earthquake timing and recurrence intervals in the Seattle Fault Zone. See Figure 4 for explanation.

The Seattle Fault Zone has a billion earthquakes per year (Figure 5).

- sfz recurrence stats:
- mode: 167 years
- median: 958 years
- mean: 1908 years

3.3 Recurrence discussion

4 Survival analysis

Survival analysis is a branch of statistics that deals with the time to events, commonly deaths (hence its name) or manufactured product failures. It is very commonly used in the social and life sciences (especially medicine and public health)

as well as in engineering (where it is often called 'reliability analysis'), and has seen application in a wide variety of fields. Unfortunately, this has lead to some variability in terminology that can make searching for certain ideas challenging; we will provide a list of common terms for the same concept where appropriate.

The techniques of survival analysis are well-suited for studying a range of geoscientific phenomena, but there seems to be little explicit mention of it in the literature or the classroom. Several of the key functions (particularly the hazard function) are used commonly enough in the earthquake science community (e.g., Davis et al., 1989; Matthews et al., 2002; Reasenberg & Jones, 1989; Sornette & Knopoff, 1997) that the terminology is not universally unfamiliar, but in our science, these tools seem to be used almost exclusively by a small set of specialists in statistical seismology who are already well versed in them. As a consequence, it is hard to find a straightforward explanation of the principles and general description of the key relationships in the geoscience literature.

Below, we give a simple overview of the aspects of survival analysis used in this paper. This should be considered an superficial introduction to a mature statistical science, but (as is often the case with applied mathematics) the basics can be quite helpful in both clarifying our ideas and giving us powerful quantitative tools. Additional techniques in survival analysis can be used if an application calls for it; for example Faenza et al. (2003) apply a proportional hazard model in a study of Italian seismicity to incorporate the effects of earthquake magnitude and spatial occurrence into a time- and space-dependent predictive model.

As we are operating on empirical PDFs that have no simple analytical form, the mathematical descriptions will be general, which should aid in comprehension.

4.1 Survival analysis mathematics

Survival analysis is based on a few simple equations incorporating the recurrence interval PDF, or $f(t)$, which is the probability that some event will occur at time t from a start time (such as a birth; for our purposes, this is the time since the last event). First, we define $T \geq 0$, a non-negative random variable representing the time between earthquakes.

Next, we can define $F(t)$, which is the probability that T is as short or shorter than some t, i.e. that a child survives for fewer than t years or that the time between earthquakes is less than t years:

$$F(t) = \begin{cases} \Pr[T \leq t] & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{1}$$

$F(t)$ is the *cumulative distribution function* of $f(t)$, and is calculated from it if $f(t)$ is known:

$$F(t) = \int_0^t f(u)\,du\,. \tag{2}$$

(Note that we are integrating from 0 instead of $-\infty$ as in the general integral for a cumulative distribution function because $f(t)$ is everywhere positive.) In the case of empirical (or non-analytical) $f(t)$, as in this study, $F(t)$ can be calculated through numerical integration.

Then, we can define the *survival function* (also known as the reliability function) $S(t)$ as

$$S(t) = \Pr[T > t] = 1 - F(t) = \int_t^\infty f(u)\,du\,. \tag{3}$$

The survival function, as the compliment of $F(t)$, describes the probability that T is longer than t, or $\Pr[T > t]$, i.e. that a child will live beyond a certain age. If $f(t)$ or $F(t)$ is not known but some some samples of T are known, $S(t)$ can be estimated directly, for example through the Kaplan-Meier Estimator (Kaplan & Meier, 1958) (which can also handle *censored data*, discussed below).

Though the survival function is foundational enough to serve as the namesake for this sort of analysis, it is not a final product in our work. Instead, it is a component of several other very useful functions that are of more immediate interest.

These functions $f(t)$ and $S(t)$ provide alternate ways of estimating a useful parameter, the polynymous *mean survival time* (most familiarly, the *life expectancy* or *expected lifetime* in demographics and engineering, and the *mean recurrence interval* for our purposes), which we denote $E(T)$. When this value is unconditional (i.e. at $t = 0$), the mean recurrence interval is the mean of the recurrence interval PDF $f(t)$, but it can also be shown to be the integral of $S(t)$:

$$E(T) = \int_0^\infty t\,f(t)\,dt = \int_0^\infty S(t)\,dt\,. \tag{4}$$

Another useful function, the *hazard function* (also called the *hazard rate function* or *failure function*) $\lambda(t)$ describes the instantaneous probability of an event at time t given that the event has not yet occurred. $\lambda(t)$ can be derived from the previous functions as:

$$\lambda(t) = f(t)/S(t)\,. \tag{5}$$

Assuming that the last date of an earthquake on a fault is known, $\lambda(t)$ may be used in probabilistic earthquake hazard analysis.

A plot of t vs. $\lambda(t)$ may assume many forms, and these reflect the processes controlling the distribution of lifetimes or interevent times, or the assumptions thereof.

4.1.1 Conditional survival

The equations above all relate to the probabilities and expectations of recurrence times (or lifetimes) starting from $t = 0$, i.e. instantaneously after an earthquake. However, while these quantities are all useful for understanding the physics of earthquakes, for time-independent probabilistic seismic hazard analysis, and many other applications, we are often very interested in how these probabilities change through time in the absence of an earthquake. The colloquial 'overdue earthquake' is an intuitive (if imprecise) representation of this; Davis et al. (1989)'s question is another.

First, we define the *conditional survival function* S_c, which is the survival function $S(t)$ conditional on $T > c$, i.e. that some time c has elapsed and the event has not occurred.

$$S_c(t) = \Pr[T_c > t] = \frac{S(c+t)}{S(c)} \tag{6}$$

Then we can calculate the *conditional mean survival time remaining* (i.e., the *conditional life expectancy*, or the *mean lifetime remaining*) $E(T_c)$, which is the mean (or expected) amount of time that remains after time $t = c$ has elapsed.

$$E(T_c) = \int_0^\infty S_c(t)\,dt \tag{7}$$

$$= \frac{\int_c^\infty S(t)\,dt}{S(c)} \tag{8}$$

$$= \frac{\int_c^\infty t\,f(t)\,dt}{S(c)}\,. \tag{9}$$

The various methods for calculating $E(T_c)$ are similar and, for the most part, do not offer different advantages in cases of limited information or computational power. However, we show these three because Equations 7 and 8 are the most common forms in the literature, while we find Equation 9 to be the most intuitive because it illustrates how the recurrence PDF $f(t)$ doesn't change its shape after some time c has passed; instead, that part of the PDF $f(t < c)$ is excised and the remainder of the PDF $f(t \geq c)$ is normalized so that it is a proper PDF, i.e. that it integrates to 1, by dividing by its area (which is equal to $S(c)$, as shown in Equation 3).

The conventional application of $E(T_c)$ is in demographic analyses of populations with a high infant mortality rate. The life expectancy at birth in these populations is lower than the life expectancy at age 5, because some significant fraction of the population dies by age 5, but those that do not may be expected to live well into adulthood.

In the case of earthquakes, there are several applications of $E(T_c)$. Two are common topics of discussion, the expectations for the interseismic time remaining given (1) 'overdue' earthquakes, i.e. when the waiting time T has exceeded the mean recurrence time $E(T)$; and (2) triggered earthquakes, when some event causes an earthquake on a fault or fault system soon after the initial event (regardless of whether the first event is an earthquake or not). The tools of survival analysis, as laid out here, allow for easy quantification of these expectations.

4.1.2 Censored data

An important topic in survival analysis is *censored data*, which is the

4.1.3 operations on KDE functions

4.2 SAF hazard

4.3 WA hazard

4.3.1 Puget Lowlands hazard

4.3.2 SFZ hazard

4.4 Expected time to failure

4.4.1 SAF

- wrightwoood = 34.96 years
- pallet creek = 65.30 years

4.4.2 Puget

- pugetL 366
- sfz 1765

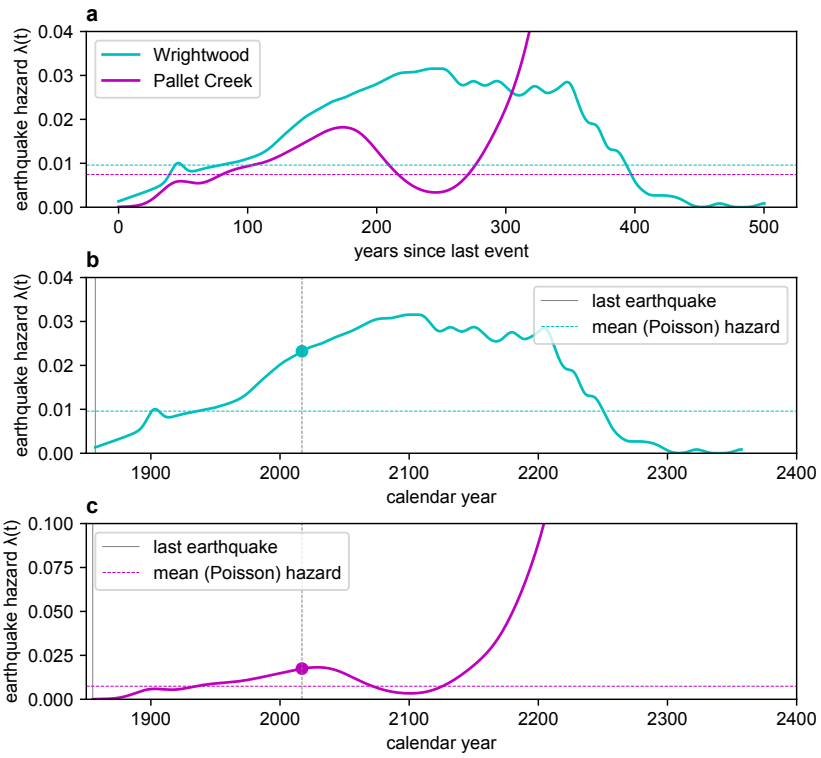

Figure 6: SAF hazards fig_saf_hazard}

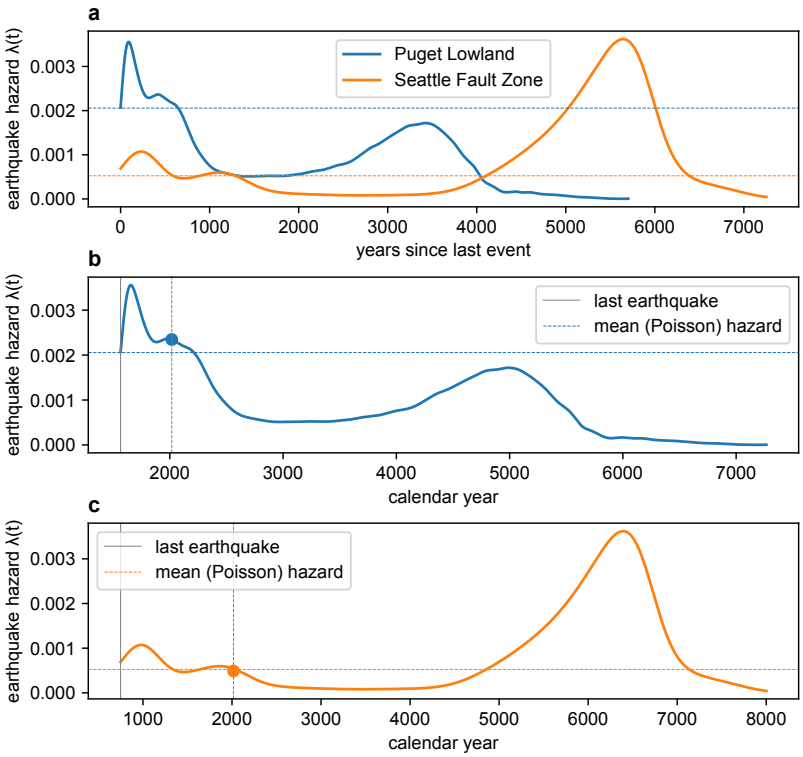

Figure 7: Earthquake hazard $\lambda(t)$ for the Puget Lowlands and Seattle Fault Zone. **a** t vs. $\lambda(t)$ for the Puget Lowlands (blue) and Seattle Fault Zone (orange) with time t since a generic last event. **b**: t vs. $\lambda(t)$ for the Puget Lowlands since the last event (on the Utsalady Fault, 384 years B.P.). **c**: t vs. $\lambda(t)$ for the Seattle Fault Zone since the last event (Seattle Fault Zone earthquake E, 748 years B.P.). Solid vertical lines in **b** and **c** indicate the last events, and dashed vertical lines indicate the hazard at the time of this writing (2017).

5 Discussion

6 Conclusions

References

Davis, P. M., Jackson, D. D., & Kagan, Y. Y. (1989). The longer it has been since the last earthquake, the longer the expected time till the next? *Bulletin of the Seismological Society of America*, *79*(5), 1439–1456. Retrieved from http://www.bssaonline.org/content/79/5/1439.short

Faenza, L., Marzocchi, W., & Boschi, E. (2003). A non-parametric hazard model to characterize the spatio-temporal occurrence of large earthquakes; an application to the italian catalogue. *Geophysical Journal International*, *155*(2), 521–531. Retrieved from https://academic.oup.com/gji/article-abstract/155/2/521/598287

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481. Retrieved from http://www.jstor.org/stable/2281868

Matthews, M. V., Ellsworth, W. L., & Reasenberg, P. A. (2002). A brownian model for recurrent earthquakes. *Bulletin of the Seismological Society of America*, *92*(6), 2233–2250. Retrieved from http://www.bssaonline.org/content/92/6/2233.short

Reasenberg, P. A., & Jones, L. M. (1989). Earthquake hazard after a mainshock in california. *Science; Washington*, *243*(4895), 1173. Retrieved from https://search.proquest.com/docview/213549358/abstract/20693C990B314449PQ/1

Sornette, D., & Knopoff, L. (1997). The paradox of the expected time until the next earthquake. *Bulletin of the Seismological Society of America*, *87*(4), 789–798. Retrieved from http://bssa.geoscienceworld.org.www2.lib.ku.edu/content/87/4/789

Styron, R. H., & Sherrod, B. L. (2016). Earthquake clustering and recurrence intervals in the puget sound region, washington: A statistical perspective. *AGU Fall Meeting Abstracts*, *33*. Retrieved from http://adsabs.harvard.edu/abs/2016AGUFM.T33D..06S