Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California

Programa de Posgrado en Ciencias de la Tierra con orientación en Sismología

Forecast of large earthquakes through semi-periodicity analysis of labeled point processes

Tesis
para cubrir parcialmente los requisitos necesarios para obtener el grado de Doctor en Ciencias

Presenta:

Claudia Beatriz Mercedes Quinteros Cartaya

Ensenada, Baja California, México
2016

Tesis defendida por

Claudia Beatriz Mercedes Quinteros Cartaya

y aprobada por el siguiente Comité

_____

Dr. Fidencio Alejandro Nava Pichardo
Director de Tesis

Miembros del comité:

Dra. Ewa Glowacka

Dr. Enrrique Gómez Treviño

Dra. Renata Dmowska

Dr. Francisco Ramón Zúñiga Dávila-Madrid

_____

Dr. Juan García Abdeslem
Coordinador del Posgrado en Ciencias de la Tierra

_____

Dra. Rufina Hernández Martínez
Directora de Estudios de Posgrado

Resumen de la tesis que presenta **Claudia Beatriz Mercedes Quinteros Cartaya** como requisito parcial para la obtención del grado de Doctor en Ciencias de la Tierra con orientación en Sismología.

Pronóstico de grandes sismos mediante el análisis de semiperiodicidad de procesos puntuales etiquetados

Resumen aprobado por:

_____
Dr. Fidencio Alejandro Nava Pichardo
Director de tesis

Los grandes sismos, ocurridos en una región sismogénica determinada, son el resultado de procesos críticamente auto-organizados de acumulación y relajación de esfuerzos; y por tanto, conforman secuencias semiperiódicas con tiempos de recurrencia que varían ligeramente de la periodicidad exacta. Estudios previos han mostrado que es posible identificar estas secuencias mediante el análisis de Fourier de series de tiempo de ocurrencia de grandes sismos en una región determinada; considerando que no necesariamente todos los sismos ocurridos en la región pertenecen a la misma secuencia, ya que puede haber más de un proceso de acumulación y relajación de esfuerzos. La identificación de secuencias puede ser usada para pronosticar la ocurrencia de sismos con intervalos de confianza bien determinados. Este trabajo presenta mejoras en el método mencionado sobre identificación de secuencias y pronóstico: a) Considera la influencia del tamaño de los sismos en el análisis espectral para la identificación de secuencias semiperiódicas, lo cual significa que los tiempos de ocurrencia de sismos son tratados como procesos puntuales etiquetados. b) Utiliza estimación de la probabilidad de no aleatoriedad mejorada. c) Mejora la estimación de los límites superiores de incertidumbre utilizados en el pronóstico. d) Aplica análisis Bayesiano para la evaluación de resultados de *postnósticos* (pronósticos a posteriori). e) Evalúa la robustez de los pronósticos mediante simulaciones tipo Monte Carlo de ruido en las magnitudes de los datos. Este método mejorado fue probado exitosamente con datos sintéticos y luego aplicado a datos reales de regiones específicas: la costa suroeste de México y el noreste del Arco de Japón. Secuencias semiperiódicas con alta probabilidad de no aleatoriedad fueron identificadas: una secuencia de nueve eventos con $M \geq 7.4$ en México y una secuencia de cuatro eventos con $M \geq 8.0$ en Japón. Los postnósticos fueron acertados para los últimos eventos de cada una de las secuencias identificadas y la probabilidad Bayesiana de los postnósticos fue comparada con la probabilidad actualizada del pronóstico. Las probabilidades de los pronósticos, para un intervalo de incertidumbre con un 95% de confianza, son mayores que las probabilidades Poissonianas , y las ganancias de probabilidad y de información son significativas.

Palabras clave: Pronóstico de sismos, Semiperiodicidad, Procesos puntuales etiquetados.

Abstract of the thesis presented by **Claudia Beatriz Mercedes Quinteros Cartaya** as a partial requirement to obtain the degree of Doctor of Science in Earth Sciences with orientation in Seismology.

Forecast of large earthquakes through semi-periodicity analysis of labeled point processes

Abstract approved by:

_____

Dr. Fidencio Alejandro Nava Pichardo
Thesis Advisor

Large earthquakes have semi-periodic behavior as a result of critically self-organized processes of stress accumulation and release in seismogenic regions. Hence, large earthquakes in a given region constitute semi-periodic sequences with recurrence times varying slightly from periodicity. In previous studies, it has been shown that it is possible to identify these sequences through Fourier analysis of the occurrence time series of large earthquakes from a given region, by realizing that not all earthquakes in the region need belong to the same sequence, since there can be more than one process of stress accumulation and release in the region. Sequence identification can be used to forecast earthquake occurrence with well determined confidence bounds.This work presents improvements on the above mentioned sequence identification and forecasting method: a) Considers the influence of earthquake size on the spectral analysis, and its importance in semi-periodic sequences identification, which means that earthquake occurrence times are treated as a labeled point process. b) Uses an improved estimation of non-randomness probability. c) Improves the estimation upper limit uncertainties to use in forecasts. d) Uses Bayesian analysis to evaluate *aftcast* (forecast done a posteriori) performance. e) Estimates the forecast robustness through Monte Carlo simulation of noise in magnitude data. This improved method was successfully tested on synthetic data and subsequently applied to real data from some specific regions: the southwestern coast of Mexico and the northeastern Japan Arc. Semi-periodic sequences with high non-randomness probability were identified: 1 sequence of nine events with $M \geq 7.4$ in Mexico and 1 sequence of four events with $M \geq 8.0$ in Japan. Aftcasts were successfully done for each of the last events in the identified sequences, and the aftcast probabilities were upgraded through Bayesian analysis and compared with the updated forecast probability for the sequence including the last event. The forecast probabilities for intervals two standard deviations wide are larger than the corresponding Poissonian occurrence probabilities, and the probability gains are significant.

Keywords: Earthquake forecasting, Semi-periodicity, Labeled point processes.

Dedication

To my parents, Mercedes Cartaya and Gerardo Quinteros

Thank you for all your support along my life.

To my mentor and friend, F. Alejandro Nava.

I couldn't have done this without you.

In memory of my grandfather

Rafael Angel Cartaya

Acknowledgements

I express my sincere thanks to CONACyT that founded my postgraduate studies through the scholarship 242919. I am very grateful with CICESE, especially with Scholar Services Department and Postgraduate Coordination of the Earth Sciences Division.

I want to give infinity thanks to my thesis advisor, Dr. F. Alejandro Nava, for giving me the opportunity to work with him. He has been the best professor that I have had ever!

Thanks so much to my committee of thesis: Dr. Ewa Glowacka, Dr. Enrique Gómez Treviño, Renata Dmowska and Francisco Zúñiga, for their time dedicated to this work, teachings and valuable contributions.

Also, I want to thank you to the technical and administrative staff of Earth Science Division, especially to Sergio Arriegui, Humberto Benítez, José Mojarro, Ana Rosa Soto and Gabriela Hernández.

Of course, I want to thanks to my schoolfellows and friends: Florian Neumann, Francisco del Toro, Olaf Arroyo, Priscy Luque, Julia Canizales, Lenin Avila, Xochitl Torres, Omar Marquez, Marlyne Jaimes, Usama Yarbuh, Erik Ramirez, Mario Fuentes, Ana Gabriela Bravo, Renee Gonzalez, Rocío Arellano, Daniela Tazzo, Javier Avendaño, and to many other equally important.

Finally, thanks to my loving family for their important support.

Thanks!!!!!

Claudia Beatriz Mercedes Quinteros Cartaya

If I have seen further it is by standing on the shoulders of Giants.

(Isaac Newton 1676)

Table of contents

List of figures

List of tables

Sinopsis

Los grandes sismos, ocurridos en una región sismogénica determinada, son el resultado de procesos críticamente auto-organizados de acumulación y relajación de esfuerzos y, por tanto, conforman secuencias semiperiódicas con tiempos de recurrencia que varían ligeramente de la periodicidad exacta.

La identificación de secuencias semiperiódicas no aleatorias en series de tiempo de ocurrencia de grandes sismos, en una región específica, permite pronosticar el próximo evento correspondiente a cada secuencia semiperiódica.

Nava et al., 2014 y Quinteros et al., 2014 propusieron un método para identificar secuencias semiperiódicas mediante el análisis de Fourier de series de tiempo de ocurrencia de grandes sismos. A diferencia de otros estudios de recurrencia, ellos propusieron que no todos los sismos, ocurridos en una región dada, necesitan pertenecer a la misma secuencia, ya que puede haber más de un proceso de acumulación y relajación de esfuerzo en la región. También utilizaron análisis de Fourier analítico en vez de la transformada digital comúnmente utilizada.

Aunque dicho método dio resultados muy satisfactorios, se apreció la posibilidad de mejorarlo para aumentar la confiabilidad de los resultados. Este trabajo considera el tamaño de los sismos para mejorar el análisis espectral y para identificar eventos pertenecientes a secuencias semiperiódicas, lo cual significa que los tiempos de ocurrencia de sismos son tratados como procesos puntuales etiquetados. También mejora la estimación del valor tope apropiado de la incertidumbre del pronóstico, y aplica análisis Bayesiano para evaluar la bondad del pronóstico. Este método mejorado es aplicado a datos sintéticos y a datos reales de la costa suroeste de México y el noreste del Arco de Japón.

La información de los sismos utilizada en los análisis es tomada del Catálogo instrumental ISC-GEM, complementado con el Catálogo ISC disponible en línea y el Catálogo histórico GEM. Para ambas regiones, Japón y México, las series de tiempo de ocurrencia de sismos abarcan una ventana de tiempo en la cual se considera que los datos están completos.

Los sismos principales junto con otros sismos próximos a ellos en tiempo y espacio, son considerados Episodios de Liberación de Momento (MRE, por sus siglas en inglés) que tienen magnitud de momento equivalente, M_{eq}, correspondiente a la suma de los momentos de todos estos sismos. Con el tiempo de ocurrencia de cada evento principal y su correspondiente M_{eq}, es construido un catálogo de MRE que posteriormente es utilizado para el análisis de las series de tiempo.

En general, el método consiste en identificar semiperiodicidad en series de tiempo de ocurrencia de sismos mediante la transformada analítica de Fourier. Considerando ahora procesos puntuales etiquetados, donde las etiquetas corresponden al tamaño de cada sismo, es asignado un peso a cada evento, mediante una función de ponderación de magnitud (MWF, por sus siglas en inglés), para ponderar su contribución al análisis espectral.

La MWF propuesta está basada en la cantidad relativa de sismos de acuerdo a su magnitud, utilizando la función de probabilidad cumulativa de la relación Gutenberg-Richter doblemente truncada. El valor de los pesos toma valores desde 0.55, asignado a la magnitud mínima incluida en el análisis, hasta 1, asignado a la magnitud máxima existente en los datos.

Como los sismos más grandes son aquellos con mayor probabilidad de presentar comportamiento semiperiódico, para cada análisis, la mínima magnitud a incluir será la más alta posible que permita que haya suficientes sismos para encontrar por lo menos una secuencia de al menos cuatro eventos, pero que no sean tantos que resulten en falsas secuencias, las cuales son reconocibles por su alta probabilidad de aleatoriedad.

Una vez obtenido el espectro de frecuencias de las series de tiempo, es identificada una componente de periodicidad y, con esta frecuencia y la fase correspondiente, se construye un "peine" estrictamente periódico. El análisis se repite cuatro veces (cuatro pases), excluyendo en cada pase los eventos más alejados del peine. En cada pase la tolerancia de la distancia evento-peine es cada vez más estricta.

Para el último pase, es posible que haya más de un candidato cercano a alguno de los dientes del peine, pero, a diferencia del método anterior, tomar en cuenta las magnitudes de los eventos puede hacer que el evento más cercano al diente del peine no sea necesariamente el mejor candidato a pertenecer a la secuencia; por lo que en el cuarto pase, cada posible combinación de los eventos candidatos a la secuencia es evaluada mediante el error de ajuste ponderado.

Considerando que un proceso genera sismos dentro de un rango de magnitud característico, el error de ajuste es ponderado de acuerdo a la desviación estándar normalizada de las magnitudes de los eventos en la combinación. La combinación con menor error de ajuste ponderado es la secuencia semiperiódica seleccionada finalmente.

Ya que los errores en tiempo entre los eventos de la secuencia y el peine son una pequeña muestra de una población asociada al proceso semiperiódico, la desviación estándar medida podría no ser la más representativa del proceso. Un nuevo valor de desviación estándar es estimado con un 90 % de confianza de no estar subestimando su valor representativo de la población, considerando que estos

errores provienen de una población normal, con $\mu = 0$, cuyas varianzas tienen distribución χ^2. Esta desviación estándar poblacional estimada es utilizada para la evaluación de probabilidad del pronóstico.

Adicionalmente, es evaluada la significancia del pronóstico, a partir de las ganancias de probabilidad y de información.

Para la región de Japón, una secuencia semiperiódica de 4 eventos fue identificada, con probabilidad de no aleatoriedad $P_c = 0.992$. Al comparar estos resultados con los obtenidos mediante el análisis sin etiquetar (considerando procesos puntuales simples), las secuencias identificadas difieren en el último evento, siendo la secuencia identificada en el análisis etiquetado la que presenta mejor ajuste y probabilidad.

Para la región de México, una secuencia semiperiódica de 9 eventos fue identificada, con probabilidad de no aleatoriedad $P_c = 0.971$. Tanto para México como para Japón, se realizó pruebas de postnóstico (pronóstico a posteriori), que lograron identificar exitosamente los eventos previos al último de cada secuencia, y postnosticar éste acertadamente.

Posterior a los postnósticos obtenidos, se revisó la probabilidad de los pronósticos, mediante el análisis Bayesiano, calculando la probabilidad de que exista el proceso semiperiódico dado que un sismo pronosticado ocurrió en un tiempo dado. Esto permitió actualizar y mejorar la estimación la probabilidad de no aleatoriedad de las secuencias.

Ya que las magnitudes referenciadas en los catálogos están usualmente redondeadas a un decimal, además de que las magnitudes muchas veces difieren entre catálogos, es posible tener errores de ± 0.1 en las magnitudes utilizadas. Por tanto, para evaluar los efectos de estos posibles errores sobre los resultados obtenidos, 100 mil realizaciones introduciendo ruido aleatorio en las magnitudes fueron hechas para ambas regiones. El resultado de estas pruebas de error muestran la robustez de los pronósticos realizados, ya que la mayoría corresponden con los pronósticos obtenidos.

El método de análisis de semiperiodicidad, tomando en cuenta el tamaño de los sismos, eficientemente ha identificado, en las regiones de estudio, secuencias semiperiódicas en series etiquetadas de tiempo de ocurrencia de sismos, de manera igual o mejor que la versión inicial del método que no considera la magnitud. La ponderación según las magnitudes y el criterio para evaluar el error de ajuste ponderado para todas las posibles combinaciones, resultó muy útil en la selección de la mejor posible secuencia. Además la desviación estándar estimada, en lugar de la medida, permite tener un intervalo de tiempo de pronóstico mucho más realista.

Chapter 1. Introduction

1.1 Problem definition

Many instances of considerable human and material losses have been caused by great earthquakes. In order to contribute to risk mitigation, numerous studies have attempted to find some method for earthquake forecasting; however, results so far have not been completely satisfactory.

Based on the Elastic Rebound theory (Reid H., 1910), it is assumed that earthquakes occur due to sudden stress release when the accumulated shear stress overcomes the resistance of the rocks and, once this stress has been released, a process of stress accumulation begins anew.

The relative motion of the tectonic plates is the main source of stress generation in the crust of the Earth (e. g. Morgan, 1968; Cox, 1973; Richardson et. al., 1979). This motion is not the same for all tectonic boundaries, but in every case the motion rate may be considered constant over hundreds, and even thousands, of years, while earthquakes due to the associated stress accumulation repeatedly occur.

But, beyond this simplistic model of stress accumulation and release, earthquakes can be seen to be the result of a critically self-organized process (e. g. Bak et al., 1988; Bak et al., 1994; Ito and Matsuzaki, 1990; Bak, 1996). Such a process evolves, presenting small seismic events, until it reaches a critical state, which leads to collapses corresponding to great earthquakes accompanied by secondary seismic activity. These collapses occur repeatedly, but not in a perfectly periodical way nor having always the same size, since it is a non-linear process and there are physical and geological complexities of the system that cause variations in its behavior.

Many probabilistic studies of earthquake forecasting are based on time-dependent models, which consider that the time of occurrence of the future next earthquake depends on the occurrence times of previous earthquakes, hence, the probability of occurrence of an earthquake increases with time, contrary to time-independent models (for example, those that assume Poissonian probability) that use the average rate of recurrence only, regardless of the time elapsed from last earthquake.

Time-dependent models are consistent with the seismic gap hypothesis, which assumes that the segments of plate boundaries with the longest times lapsed from the last rupture are the ones most likely to rupture in a near future (Scholtz, 1990).

One of the main motivations for the development of time-dependents models has been the observation of recurrence times of characteristic earthquakes in fault segments. Nishenko and Buland (1987) explain characteristic earthquakes as those that break repeatedly the selfsame fault segment and whose source has dimensions that define this segment. Other definitions of characteristic earthquakes assume that, in some region, earthquakes with characteristic size or magnitude occur more frequently than would be expected according to the Gutenberg-Richter relation, although not necessarily on the same fault segment.

Quinteros (2012), Nava et al. (2014), and Quinteros et al. (2014) assumed that the occurrence of great earthquakes should constitute semi-periodic sequences associated to seismogenic processes; that is, recurrence times vary slightly form perfect periodicity. However, unlike other studies, they also considered the possibility of there being, in a given region, several seismogenic processes that lead to more than one semi-periodic sequence of earthquakes and, in addiction, the possibility of there being earthquakes that do not correspond to an identifiable semi-periodic process. Another feature of the above-mentioned works is the identification of semi-periodic sequences of great earthquakes through Fourier analysis using the analytic, instead of the digital, transform.

A sequence of six earthquakes of magnitude $M \geq 6.0$, occurred between 1857 and 1966, having an apparent periodicity of ~22 years and considered by some to be characteristic earthquakes, was observed at the Parkfield segment of the San Andres fault, in California, USA. This lead to the forecast of an earthquake to occur in 1993, with 0.95 probability (Bakun and Lindh, 1985); however, the earthquake occurred eleven years late. Application of the semi-periodicity method by Nava et al. (2014) to the Parkfield case, considering the same database used by Bakun and Lindh (1985), identified a sequence composed of four seismic events; two events, that in other studies had been included as part of the sequence, were correctly excluded by this analysis. From the identified sequence, the next earthquake was aftcast (forecast a posteriori) for 2005.63 ± 9.10 (± 2 standard deviations) with 0.85 probability. This aftcast was satisfactorily very close to the actual time of the earthquake, which occurred on September 28, 2004.

The method was also applied to regions of Japan and Venezuela (Quinteros et al., 2014). As in the Parkfield case, the method proved capable of identifying semi-periodic sequences in different seismogenic regions. For Japan, two sequences of $M \geq 8.0$ earthquakes were identified: the first with four earthquakes and the second with three. For Venezuela, semi-periodic sequences, comprising up to six earthquakes, were identified for $M \geq 5.6$ earthquakes in two regions, and for $M \geq 6.0$ earthquakes in other two regions. The estimated forecast probabilities were high compared to those from Poissonian probability; so that this method gives information gains over the Poissonian estimations.

In other aftcasting study for northeastern Venezuela done by Quinteros and Nava (2014), a sequence composed of six events with magnitudes $M \geq 5.7$ was identified, and the next earthquake was aftcast for 2014.67 ± 3.05 (± 2 standard deviations) with 0.8 probability. The earthquake occurred on October 11, 2013, which corresponds to $\Delta t = 0.89$ yr before the aftcast central time.

Nava et al. (2016) applied the Bayes' formalism to use information from the occurrence or non-occurrence of a forecast earthquake to validate the use of the non-randomness probability as the forecast probability, to quantify the forecast accuracy, to support or negate the semi-periodic sequence hypothesis, and to identify the occurrence of forecast earthquakes. This formalism was applied to the Parkfield (Nava et al., 2014) and northeast Venezuela cases (Quinteros and Nava, 2014).

Application of the semi-periodicity forecasting method described above yielded quite satisfactory results, but we considered that some aspects of the method could be improved, and this thesis is devoted to exploring possible improvements.

In the present work we explore the influence of earthquake size on identification of the semi-periodic sequences. To this effect, the earthquake occurrence series is no longer treated as a simple point process in time, but as a labeled point process. The effects of the earthquake magnitudes on the spectral identification of sequences and on the selection of events belonging to them are considered. The standard deviation estimation for forecasts is revised. Bayes' Theorem is applied to judge the performance of the forecast method when applied to data series amenable to aftcasting. Finally, the effect of possible rounding errors in the catalog magnitudes is evaluated using a Monte Carlo approach in order to analyze the reliability of the results.

1.2 Hypotheses

Hypotheses 1 to 3, stated below, were used in the MSc thesis (Quinteros, 2012), and are also used in this study. In this thesis, these three hypotheses are complemented by hypotheses 4 and 5, which lead to improvements in the semi-periodicity forecasting method:

1. The processes of stress accumulation and release in some seismogenic region, which result in large earthquakes, are critically self-organized processes so that large earthquakes have semi-periodic behavior.
2. There can be more than one process of stress accumulation and release in a given region; therefore, not all large earthquakes that occur in a region need be associated with the same semi-periodic process.
3. Semi-periodic sequences of earthquakes, within a time series representing the occurrence times of the large earthquakes in a given region, can be identified through Fourier analysis.
4. The larger the size of the earthquake, the greater its influence on the semi-periodic process; thus, earthquake occurrence times will not be considered to constitute simple point processes, but will rather be treated as labeled point processes, where labels correspond to the event magnitudes.
5. The earthquakes generated by each particular semi-periodic process may have different magnitudes, but within some characteristic magnitude range.

1.3 Objective

The objective is to identify non-random semi-periodic sequences within occurrence time series of large earthquakes occurring in some specific region, in order to forecast the next earthquake corresponding to each semi-periodic sequence. Semi-periodicities in the time series are identified using the analytic Fourier spectrum. Possible influences of the size of each earthquake in the analysis are explored and evaluated. This method will be applied to specific regions of Mexico and Japan.

Chapter 2. Theory

2.1 Time series:

The occurrence times of earthquakes in a seismogenic region, can be considered as a point process that constitutes a time series $t_e = \{t_j ; j = 1:N\}$; where t_j is the occurrence time of the $j'th$ event and N is the number of earthquakes that occurred within the observation time.

An earthquake sequence corresponds to a periodic process if the occurrence times are expressible as

$$t_j = t_0 + j\tau \ ,$$

where t_0 is some initial time, and the period is $\tau_j = t_j - t_{j-1} = \tau$.

According to Nava et al. (2014) and Quinteros et al. (2014), an earthquake sequence is semi-periodic if the occurrence times satisfy

$$t_j = t_0 + j\tau + \theta_j \ , \qquad (1)$$

where θ_j is a realization of a random variable such that $|\theta_j| \ll \tau$ (we assume it smaller than $\sim \tau / 6$).

Since we consider that not all earthquakes in the time series necessarily belong to the same semi-periodic sequence, there may be more than one semi-periodic sequence of earthquakes, as well as earthquakes extraneous to these sequences (Figure 1), the number of earthquakes contained in the time series is thus

$$N = K_1 + K_2 + \cdots + K_n + R \ ,$$

assuming that there are n semi-periodic sequences and $i = 1:n$, then K_i is the number of earthquakes that belong to the $i'th$ semi-periodic sequence and R is the number of earthquakes extraneous to these sequences.

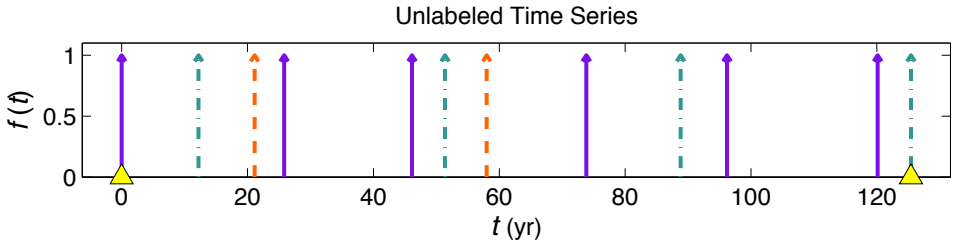

Figure 1. Example of unlabeled time series; that is, earthquakes occurrence times are considered as simple point processes. Time series represents synthetic data: a sequence of six earthquakes (solid violet lines), another sequence of four earthquakes (dashed green lines), and two extraneous earthquakes (dashed orange lines). Generally, in a real case, the sequences are not easy to recognize. The yellow triangles indicate t_0 and t_f.

2.2 Analytical Fourier transform:

Since a time series such as the one describing the occurrence in time of earthquakes is not amenable to the digital Fourier transform, it was decide to use the analytical Fourier transform to look for semi-periodicities in the point processes; to do so, Nava et al. (2014) represented the time series by the function

$$f(t)=\sum_{j=1}^{N}\delta(t-t_j) \ ,$$

where $\delta(t)$ is the Dirac delta and $f(t)$ is recognized as a section of the function corresponding to the infinite time series

$$f_\infty(t)=\sum_{j=-\infty}^{\infty}\delta(t-t_j) \ ,$$

so that

$$f(t)=f_\infty(t)\,\Pi\left(\frac{t-t_c}{T}\right) \ ,$$

where $\Pi(t-t_c/T)$ is the box function centered at time $t_c=(t_0-t_f)/2$, with length corresponding to the observation time $T=t_f-t_0$; t_0 and t_f are, respectively, the initial and final time of the time window containing the time series.

The Fourier spectral analysis method used to identify semi-periodic sequences of earthquakes, described by Nava et al. (2014) and Quinteros et al. (2014), is now adapted to the case of earthquake

occurrence time series considered as labeled point processes, where the labels indicate the size of each earthquake, through its magnitude.

These labels are used to assign to each datum a weight, $w(M)$, that quantifies its importance in the Fourier spectral analysis, by means of a *Magnitude Weighting Function* (MWF), which will be explained in detail below. Thus, each earthquake occurrence time is represented now by a pulse with area proportional to its weight (Figure 2); so that, the function corresponding to the time series is now

$$f(t)=\sum_{j=-\infty}^{\infty} w(M_j)\,\delta(t-t_j)\,\Pi\left(\frac{t-t_c}{T}\right)=\sum_{j=1}^{N} w(M_j)\,\delta(t-t_j)\ , \qquad (2)$$

where $w(M_j)$ is the weight assigned to the earthquake occurred at time t_j with magnitude M_j.

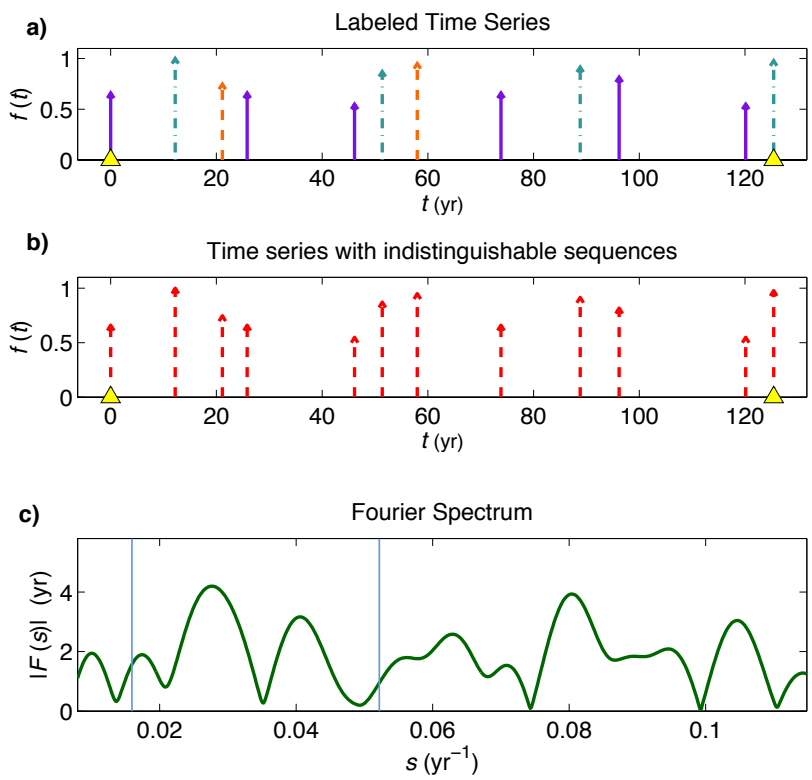

Figure 2. Example of labeled time series, using the same synthetic data of Figure 1. a) Time series as a labeled point process, with symbol lengths (arrows) proportional to weights. Conventions are as in Figure 1 b) Time series as a labeled point processes where sequences are not recognizable. c) Fourier spectrum of the labeled time series; the solid vertical blue lines are the guideline frequencies s_{min} and s_{max} (Nava et al., 2014).

The FT of (2) is thus

$$F(s)=\int_{-\infty}^{\infty}\sum_{j=-\infty}^{\infty}w(M_j)\,\delta(t-t_j)\,\Pi\left(\frac{t-t_c}{T}\right)e^{-i2\pi ts}\,dt=\sum_{j=1}^{N}w(M_j)\,e^{-i2\pi t_j s}\ , \qquad (3)$$

where s is frequency.

2.3 Magnitude Weighting Function (MWF):

Physical measures of the earthquakes size are given by the seismic moment or the seismic energy, but these parameters cannot be used for weighting because only the largest earthquakes would be significant. The MWF cannot be expressed as a simple function of magnitude, because the magnitude is not a physical measure, but is related linearly to the logarithm of the seismic moment.

It is well known that the magnitude distribution of earthquakes follows the Gutenberg-Richter (GR) relation, $\log_{10}N(M)=a-b(M-M_{\min})$, where the b-value is a constant that relates the relative numbers of small to large earthquakes, $N(M)$ denotes the cumulative number of earthquakes of magnitude greater than or equal to a given magnitude M and $M_{\min}$ is the minimum magnitude for which the GR relation behaves linearly. The GR relation implies that the probability density function of earthquake magnitude is an exponential one

$$f(M)=\beta e^{-\beta(M-M_{\min})}\ ,$$

where $\beta=b\ln(10)$ and b-value is estimated using Utsu's (1965) formula:

$$b=\frac{\log_{10}e}{\bar{M}-\left(M_{\min}-\dfrac{\Delta M}{2}\right)}\ , \qquad (4)$$

where $\Delta M=0.10$ is the usual magnitude uncertainty.

We propose a MWF based on the relative abundance of earthquakes according to their magnitude, using the cumulative probability function in terms of the doubly truncated Gutenberg-Richter (GR) relation:

$$F(M)=\Pr(m\le M)=\frac{e^{-\beta M_1}-e^{-\beta M}}{e^{-\beta M_1}-e^{-\beta M_2}}\ ,$$

where M_1 and M_2 are the minimum and maximum magnitudes in the data set, respectively. If we assign a minimum weight, w_1, to M_1, and a maximum weight, w_2, to M_2, the MWF can be defined as

$$w(M) = w_1 + (w_2 - w_1)F(M) \ . \tag{5}$$

Since w_1 and w_2 can be assigned arbitrarily, we use $w_2 = 1$ to indicate full importance, and $w_1 = 0.55$ (a value arrived at from the results of many tests) for the least important events (Figure 3). This value range is enough to obtain a weighting scale that distinguishes between data with different magnitudes while avoiding weights so small that would make the role of the smaller earthquakes negligible. The shape of the MWF (Figure 3) depends mainly on the magnitude range, and depends only weakly on the *b*-values in the range 0.7 and 1.2.

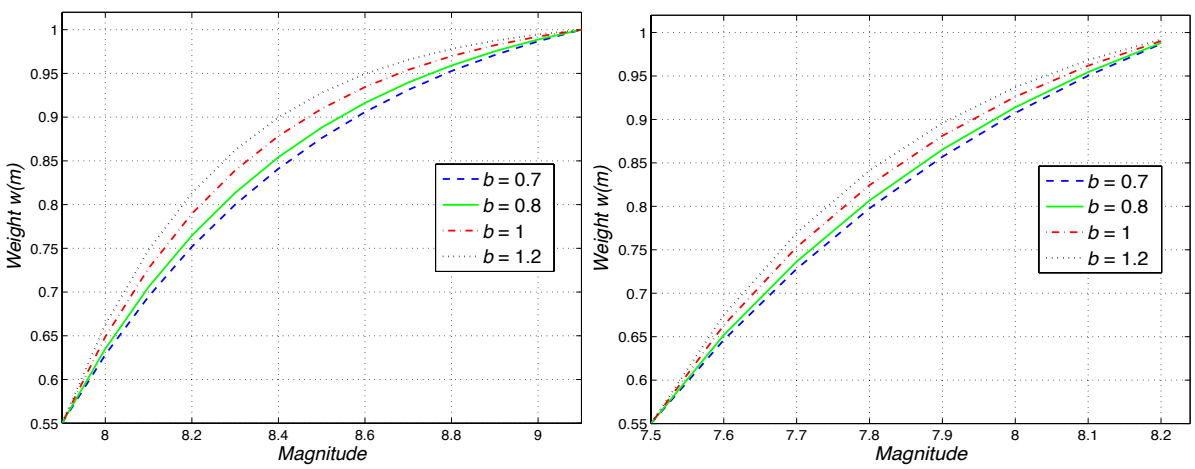

Figure 3. Examples of the MWF, $w(M)$, for various *b*-values and different magnitude ranges.

After assigning MWF weights according to (5), the Fourier spectrum is obtained from (3), using the weighted time series (2). An example of differences between the amplitude spectra of the unlabeled time series and the labeled one, is shown in Figure 4; these spectra differ in the amplitudes value and, in general, the amplitude peaks are associated to a frequency that lightly differs from one spectrum to another, but, sometimes, this difference can be enough to identify semi-periodic sequences that are not the same in both case.

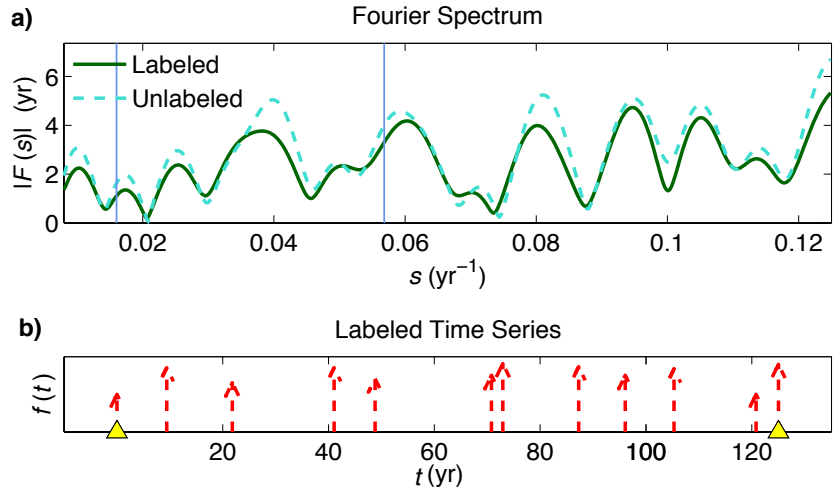

Figure 4. Comparison of labeled and unlabeled time series Fourier spectrum. The solid vertical blue lines are the guideline frequencies s_{min} and s_{max} (Nava et al., 2014). b) Labeled time series, with symbol lengths (red arrows) proportional to weights; the yellow triangles indicate t_0 and t_f.

Chapter 3. Data

3.1 Seismic catalog:

The earthquake occurrence data come from the ISC-GEM Global Instrumental Earthquakes Catalogue, the ISC on line Catalogue, and the GEM Historical Earthquakes Catalogue. Only the time window within which the information is complete is analyzed. Moment magnitudes are used throughout.

3.2 Minimum magnitude selection:

As mentioned above, large magnitude earthquakes are the ones most likely to show semi-periodic behavior; therefore, for each analysis we select as M_2 the largest magnitude in the appropriate space and time windows. The minimum magnitude, M_1, is chosen so that there is a sufficient number of earthquakes to do the analysis, i. e. enough events for at least one possible sequence of at least four events, but not so many events that may result in false sequences, which are recognizable as such because their probability of random occurrence is high.

3.3 Moment Release Episodes (MRE) and the equivalent magnitude:

A large earthquake, together with its aftershocks and other associated events in its immediate vicinity, are considered a Moment Release Episode (MRE), having an equivalent moment magnitude, M_{eq} corresponding to the sum of the moments of the main event and its associated events (Quinteros et al., 2014).

In most cases, the equivalent magnitude is (after rounding) equal to the magnitude of the main event of the MRE, and the data set does not change. But when an event is close to the minimum magnitude of acceptance, and there are other large events in the MRE, the equivalent magnitude could be large enough for the event to be included in the data set. Hence our semi-periodicity analysis is done considering MRE events instead of individual earthquakes.

The equivalent magnitudes are calculated based on the Hanks and Kanamori (1979) relation

$$\log_{10} M_0 = 16.095 + 1.5 M_w \ ,$$

where M_0 is the seismic moment measured in dyne·cm.

Thus, for each MRE the equivalent magnitude is

$$M_{eq}^{I} = \frac{2}{3} \log_{10} \sum_{i \in I} 10^{1.5 M_i} \ , \qquad (6)$$

where I is the set of indices of the events constituting it. MREs are built by associating to each earthquake that does not already belong to an MRE, those events or MREs having magnitudes smaller than that of the earthquake in question, and lying within given spatial and temporal windows with sizes that depend on the earthquake's magnitude.

The spatial closeness criterion correspond to ellipses that depend on the orientation of each earthquake, with the major axis along the strike of the local fault system, and the length of the axes are given by the Wells and Coppersmith (1994) empirical relations between magnitude and fault rupture length.

Even though aftershock activity can go on for years (time span varies with region and magnitude), the largest part of the energy or moment is liberated within the first days or weeks of the occurrence of the main event. Thus, the temporal closeness criterion used is 60 days.

The time of each MRE is that of the largest event in the set.

Chapter 4. Data processing

4.1 Spectral analysis

In the amplitude spectrum obtained from the analytical Fourier transform (3), large eligible peaks (Nava et al., 2014) correspond to frequencies that are possibly associated with semi-periodic sequences in the time series (Figure 5). These peaks are explored, from highest to lowest frequencies, in order to find frequencies that correspond to valid semi-periodic sequences. The reason for starting from the highest frequencies, is twofold: low-frequency peaks could be submultiples of the real frequency of a sequence; in time series including many seismic events, it is highly likely to find spectral peaks associated to false long-period sequences.

The frequency s_c, corresponding to the chosen peak, is used to build a strictly periodic sequence that we call a "comb", which is constituted by K teeth with period $\tau_c = 1/s_c$ and begins at time $t_{0c} = \left(-\phi_c / 2\pi s_c\right) + t_0$, where $\phi_c = \arctan\left\{\operatorname{Im} F\left(s_c\right) / \operatorname{Re} F\left(s_c\right)\right\}$ is the spectral phase corresponding to s_c.

For a sequence of K earthquakes to be considered semi-periodic, the absolute difference between the time of each comb tooth, $t_i^c; i = 1:K$ and that of, at least, one of the N earthquakes contained in the series, $t_j^e; j = 1:N$, should be less than a small fraction of the period

$$\Delta t = \left| t_i^c - t_j^e \right| < r\tau_c \ ,$$

where $r \leq 1/4$ indicates the fraction of the period to consider, as discussed below.

Since the frequency corresponding to the maximum value of a peak is influenced both by the events in the corresponding sequence, and by all other events in the series, the frequency determination is refined by iteratively eliminating those events that cannot possibly belong to the sequence.

Thus, the spectral analysis (Figure 5) is carried out three times (three passes) and, in each pass, the tolerance of the difference between events and comb, $\Delta t < r\tau_c$, is progressively stricter; we used for each pass $r = \left\{1/4, 1/4.5, 1/5\right\}$, respectively. If, during any pass the tolerance criterion is not satisfied, then the frequency being explored is deemed unviable.

When there is more than one event within the tolerance interval $\left[t_i^c-\left(r\tau_c\right),t_i^c+\left(r\tau_c\right)\right]$, for a given tooth, such as in the case of the first and fourth teeth in the example shown in Figure 5c, it is necessary to decide which of the events that are close to the tooth to choose for the semi-periodic sequence. For the unlabeled analysis the event closest to the tooth was the preferred one (Nava et al., 2014; Quinteros et al., 2014; Quinteros and Nava, 2014), so that there was only one sequence to choose from.

For the case of labeled time series, after the third pass, all possible combinations of events that are within the tolerance interval are considered as candidate sequences to be analyzed in the fourth pass (Figure 6).

The fourth pass is done separately for each candidate sequence, considering an even smaller tolerance, $\Delta t<\left(1/6\right)\tau_c$. Finally, the combinations that fulfill the closeness criterion, are evaluated through a weighted fit error (explained in the next section). The combination with the smallest weighted fit error is selected as the identified sequence.

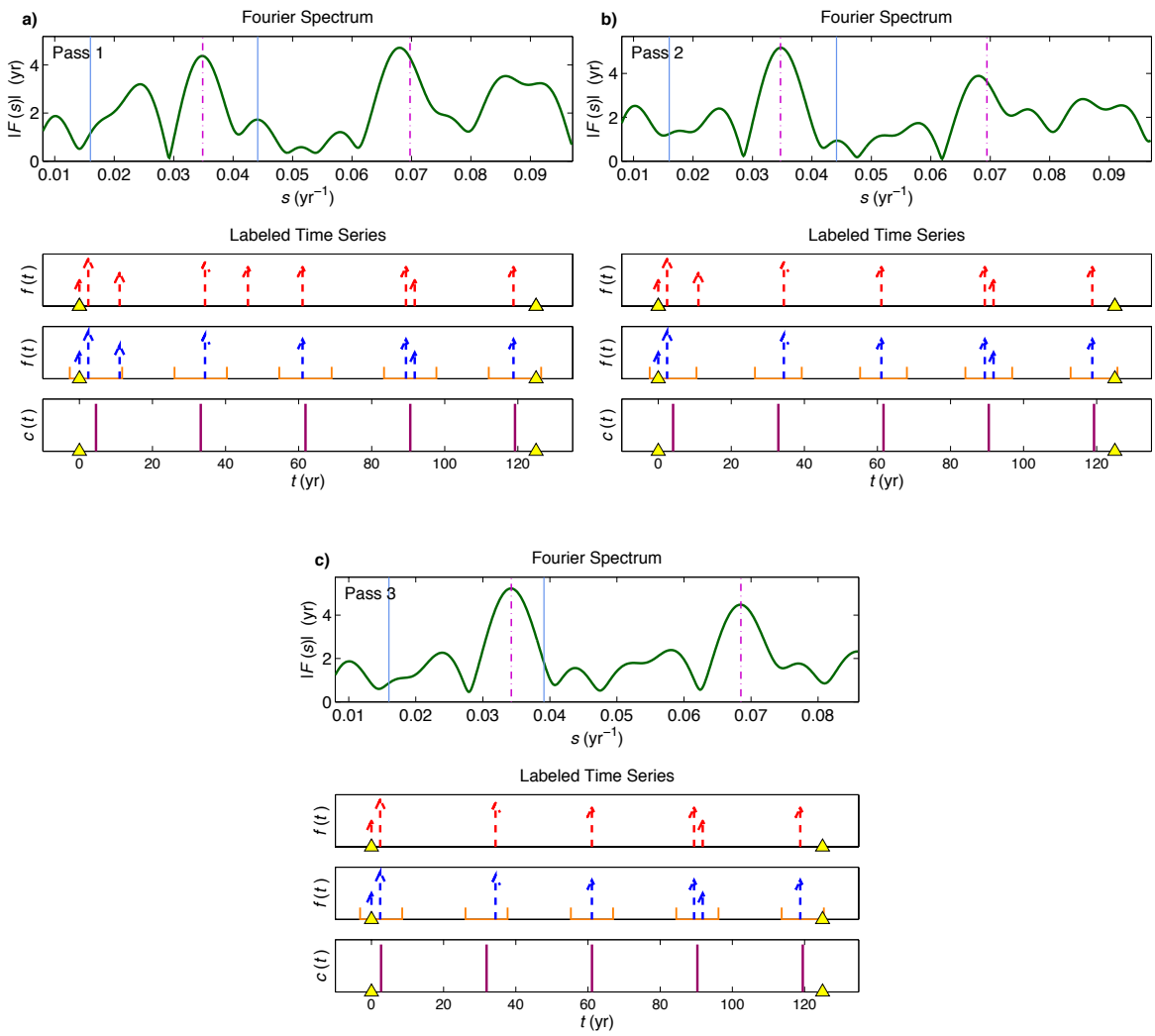

Figure 5. a) First pass, b) Second pass, and c) Third pass of the spectral analysis. (Top) Fourier spectrum of the time series shown immediately below; the vertical dot-dashed purple lines, at the left, indicate the chosen frequency and one of its multiples at the right; the solid light blue vertical lines are the guideline frequencies $s_{\min}$ and $s_{\max}$ (Nava et al., 2014). (Middle) the labeled time series in where dashed red arrows represent each event, and the $f(t)$ bellow shows as dashed blue arrows those events that are within the confidence interval (solid orange lines) and could belong to the sequence. (Bottom) $c(t)$ is the comb, whose teeth are shown by solid dark purple lines. The yellow triangles indicate t_0 and t_f.

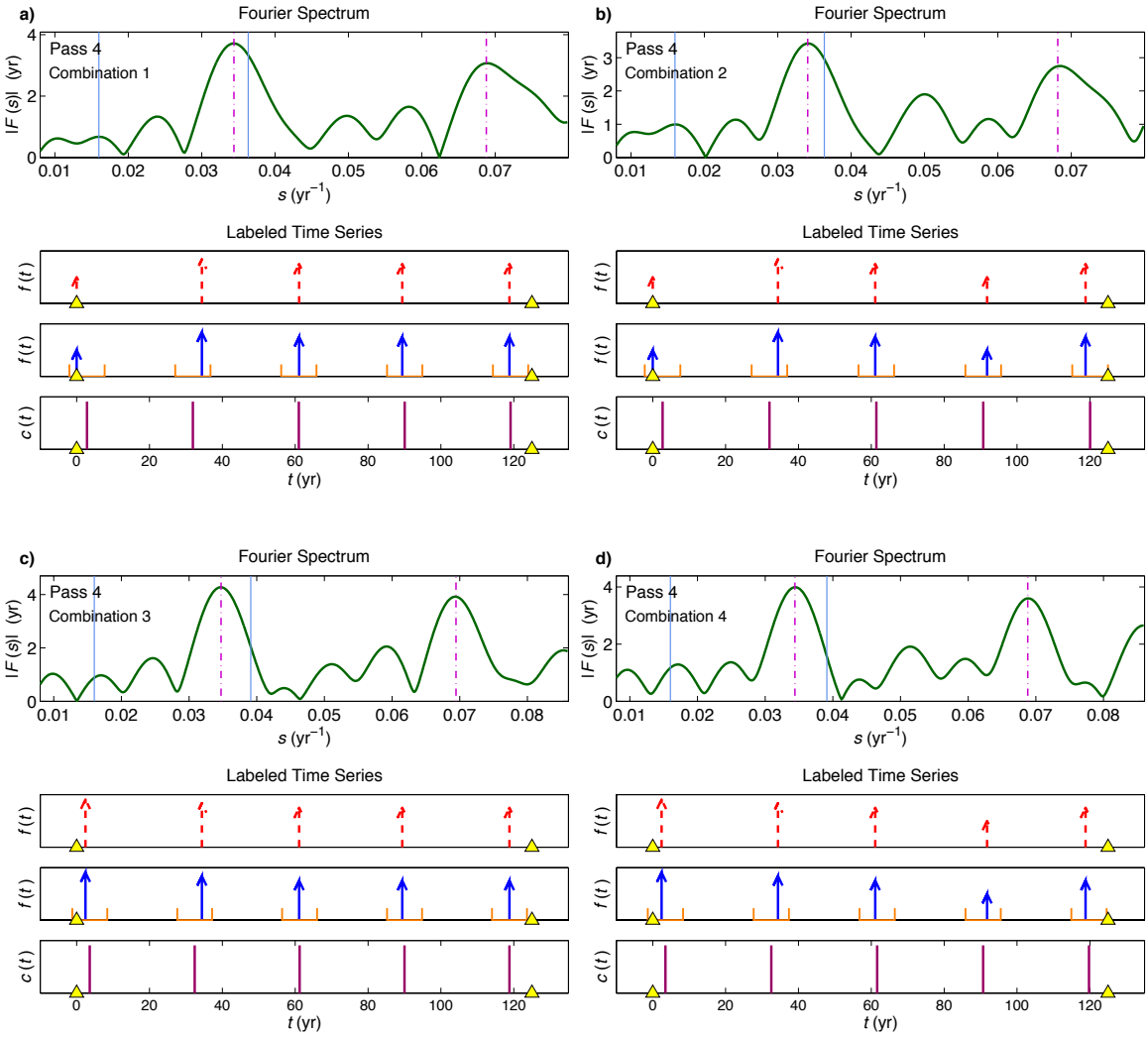

Figure 6. Fourth pass of spectral analyses of four combinations. Each combination corresponds to a possible sequence. Conventions are as in Figure 5. Combination 3 shows the best representation of a semi-periodic sequence.

4.2 Weighted fit error in the semi-periodic sequence selection:

We consider the possibility that some process may generate earthquakes within some preferred magnitude range and the time fit error is weighted according to the normalized standard deviation of the event magnitudes in the sequence.

Since every possible sequence has the same *K* number of earthquakes, the weighted fit error measure used to compare them, is

$$\varepsilon_{\psi} = \sum_{i=1}^{K} \varepsilon_i \psi \tag{7}$$

where ε_i is the fit error of event times respect to the comb teeth times,

$$\varepsilon_i = \left| t_i^c - t_i^e \right| \ , \tag{8}$$

and ψ is a factor that penalizes magnitude departures from the mean of the magnitudes of the events in a determined combination,

$$\psi = \left(1 + \frac{\sigma_{M_{eq}}}{\bar{M}_{eq}} \right) ,$$

where $\sigma_{M_{eq}}$ and $\bar{M}_{eq}$ are the standard deviation and the mean of the magnitudes in the combination, respectively:

$$\sigma_{M_{eq}} = \sqrt{\frac{\sum_{i=1}^{K} \left(M_{eq}^i - \bar{M}_{eq} \right)}{K-1}} \ ,$$

$$\bar{M}_{eq} = \frac{\sum_{i=1}^{K} M_{eq}^i}{K} \ .$$

The combination with the smallest ε_{ψ} is selected as the identified sequence. The minimum possible value of ψ equals 1, corresponding to a combination of earthquakes that have all the same magnitude, no matter how many events there are in the sequence, and the maximum possible value of ψ depends on *K* and on the ratio of minimum to maximum magnitudes of the combination; so that, for a given magnitude ratio, the longer the sequence, the smaller the weight.

4.3 Standard deviation estimation

Since we consider the fit errors, ε_i (8), to be realizations of θ (1), the measured standard deviation of the $\varepsilon_i = \theta_i$ for the selected sequence, is

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{K} \theta_i^2}{K}} \ ,$$

with K degrees of freedom because θ is postulated to have zero mean, independently of the θ_i values.

However, these errors are a very small sample from a population associated to the semi-periodic process, so that the measured standard deviation may not correctly reflect the standard deviation of the process.

In order to have a more reliable estimation of the true standard deviation, since θ_i can be considered to come from a normal population, with variance $\hat{\sigma}^2$ and mean $\mu = 0$, their variances will be distributed as χ^2

$$\sum_{i=1}^{K} \frac{\theta_i^2}{\sigma^2} = \frac{K\hat{\sigma}^2}{\sigma^2} \sim \chi_K^2 \ .$$

Hence, in order to have *P* confidence that the population standard deviation is not underestimated, defining $p = 1 - P$, the estimated standard deviation is given by

$$\sigma = \sqrt{\frac{K}{x_p}} \hat{\sigma}$$

where x_p is the value χ^2 corresponding to confidence P. For 90% confidence, $x_{0.10} = 1.06362$ and $\sigma = 1.93926\,\hat{\sigma}$; we will consider this to be the estimated standard deviation of the distribution, σ, used here for estimating the forecast probabilities.

4.4 Sequence probabilities:

To test the hypothesis of semi-periodicity, the null hypothesis would be that earthquakes occur randomly in time so that the number of events, *n*, in any interval τ is distributed according to Poisson's distribution

$$\Pr(n) = \frac{(\lambda\tau)^n e^{-\lambda\tau}}{n!} \ , \tag{9}$$

where $\lambda = N/T$, N is the number of earthquakes that occurred over the interval $[t_0, t_f]$ that corresponds to the observation time $T = t_f - t_0$. Nava et al. (2016) estimated the probability of the observed sequence occurring at random to be

$$P_0 = \left(1 - e^{-\lambda\Theta}\right)\left(1 - e^{-\lambda\theta}\right)^{K-1} \ , \tag{10}$$

where K is the number of events in the sequence. The period of the sequence must be between $T/K + \xi$ and $T/(K-1)$, where ξ is a very small quantity introduced to ensure that no more than *K* elements fit within time *T*. In the worst case that results in the largest random probability, when considering the shortest period and taking into account the uncertainty, at least one event should occur within an interval of length $\Theta = (T/K) - \xi + q\sigma$, $\xi \ll T/K$, and the other events should occur each within an interval of length $\theta = 2q\sigma$. These time lengths, Θ and θ , correspond to the uncertainty interval of the event occurrence times, where σ is the estimated standard deviation of the sequence and the parameter q establishes how rigorous is the probability estimate; we use $q = 2$ for a confidence level of 95.45%.

Hence, the probability that the sequence did not occur by chance, i. e., the non-randomness probability, is

$$P_c = 1 - P_0 \ . \tag{11}$$

Chapter 5. Forecast

5.1 The next event belonging to a sequence

If the identified sequence is indeed due to a semi-periodic process, the occurrence time of the next event belonging to the sequence is

$$t_{nxt} = t_{0c} + K\tau_c \pm q\sigma \ .$$

Since observed sequences include too few events to fully characterize the probability density function (pdf), $p(t)$, of the variations from periodicity, invoking the Central Limit Theorem, we use, as a working hypothesis for $p(t)$, the normal distribution, centered on t_{nxt}, with standard deviation σ, truncated at the occurrence time of the most recent event (Figure 7), and normalized so that the area under the curve is equal to the non-randomness probability of the sequence, P_c. Thus, the probability that the next event belonging to the comb occurs within the interval $\left[t_{nxt} - q\sigma, t_{nxt} + q\sigma\right]$ is

$$P_{cq} = P_c \operatorname{erf}\left(q\right) \ . \tag{12}$$

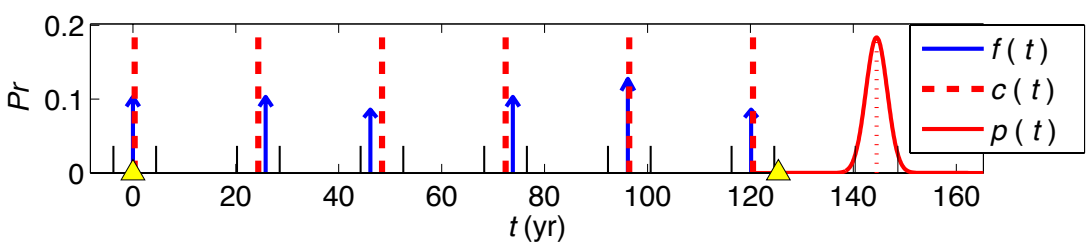

Figure 7. Example of forecast. Earthquakes are blue arrows and comb teeth are dashed red lines the forecast pdf $p(t)$ is shown by the red curve and the dotted line at its center indicates t_{nxt}, while the short vertical black lines, on both sides of the teeth and the forecast, indicate $\pm 2\sigma, (q=2)$. The triangles indicates t_0 and t_f.

5.2 Significance

To estimate how significant the results of the labeled analysis are, we evaluated the probability and information gains (Vere-Jones, 1998; Harte and Vere-Jones, 2005).

Following Nava et al. (2014) and Quinteros et al. (2014), the probability gain is obtained by comparing the probability of having a comb event and/or a non-comb event during a given interval $\left[t_{nxt}-q\sigma, t_{nxt}+q\sigma\right]$, centered on time t_{nxt}, calculated through $P_{cq}+\pi^{*}_{1+}-P_{cq}\pi^{*}_{1+}$, with the Poissonian probability of at least one earthquake occurring during the given interval:

$$P_G=\frac{P_{cq}+\pi^{*}_{1+}-P_{cq}\pi^{*}_{1+}}{\pi_{1+}},$$

where P_{cq} is the occurrence probability of the next event belonging to the sequence within the forecast confidence interval (12), π^{*}_{1+} is the Poissonian occurrence probability of at least one event within the forecast confidence interval, considering that there is a determined semi-periodic sequence and the occurrence rate of events not belonging to the sequence is $\lambda^{*}=\left(N-K\right)/T$, and $\pi_{1+}=1-e^{-\lambda 2q\sigma}$ is the Poissonian probability, in the absence of a forecast, of an earthquake occurring during the given interval, with occurrence rate $\lambda=N/T$.

The information gain (Fano, 1961; Harte and Vere-Jones, 2005), $I_G=\log_2\left(P_G\right)$, is the difference in the Shannon (1948) information of the comb forecast probability and the background probability.

5.3 Bayesian probability

Nava et al. (2016) propose that a forecast probability can be revised, after the forecast event has occurred, through the Bayesian formalism (e.g. Parzen, 1960; Winkler, 2003). Considering as event *A* the presence of semi-periodicity, evidenced by the identification of a semi-periodic sequence with non-randomness probability $\Pr\left(A\right)=P_c$, after the occurrence of a later earthquake at observed time t_O, event *B*, Bayes formalism can be used to test whether *B* is indeed the aftcast event and whether its occurrence supports the semi-periodicity assumption.

For Bayes' well-known formula, we can calculate the probability that a semi-periodic process exists, given that the forecast earthquake occurred at a given time:

$$\Pr(A|B)=\frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A)+\Pr(B|\bar{A})\Pr(\bar{A})} \ , \tag{13}$$

where $\Pr(B|A)$ is the probability given by the $p(t)$ forecast (Figure 7) for occurrence in an interval around the actual occurrence

$$\Pr(B|A)=P_{cw}+\pi^*_{1+w}-P_{cw}\pi^*_{1+w} \ ,$$

where P_{cw} is the probability that an earthquake belonging to the sequence occurs within the finite interval $\left[t_O-\frac{w}{2},t_O+\frac{w}{2}\right]$, centered on t_O that is the time in which occurred the aftcasted earthquake, and with length w, discussed below :

$$P_{cw}=P_c\int_{t_o-\frac{w}{2}}^{t_o+\frac{w}{2}}p(t)\,dt \ .$$

For the case $p(t)=\mathrm{N}(t_{nxt},\sigma)$ illustrated in Figure 8,

$$P_{cw}=\frac{P_c}{\sqrt{2\pi}}\int_{t_1}^{t_2}e^{-x^2/2}\,dx=P_c\left\{\mathrm{sgn}(t_2)\,\mathrm{erf}\left(|t_2|\sqrt{2}\right)-\mathrm{sgn}(t_1)\,\mathrm{erf}\left(|t_1|\sqrt{2}\right)\right\} \ ,$$

where $t_1=\left[\,|t_{nxt}-t_O|-(w/2)\right]/\sigma$ and $t_2=\left[\,|t_{nxt}-t_O|+(w/2)\right]/\sigma$, for the normalized time $(t-t_{nxt})/\sigma$.

π^*_{1+w} is the occurrence probability of at least one earthquake that does not belong to the sequence, within a finite interval of length w

$$\pi^*_{1+w}=1-e^{-\lambda^* w} \ ,$$

where $\lambda^*=(N-K)/T$.

The probability of the semi periodic process does not exist is $\Pr\left(\bar{A}\right)=1-\Pr\left(A\right)$, which represents the null hypothesis (10), and given $\bar{A}$, the probability of the event *B* is just the Poisson probability

$$\Pr\left(B|\bar{A}\right)=1-e^{-\lambda w} \ ,$$

where $\lambda = N/T$.

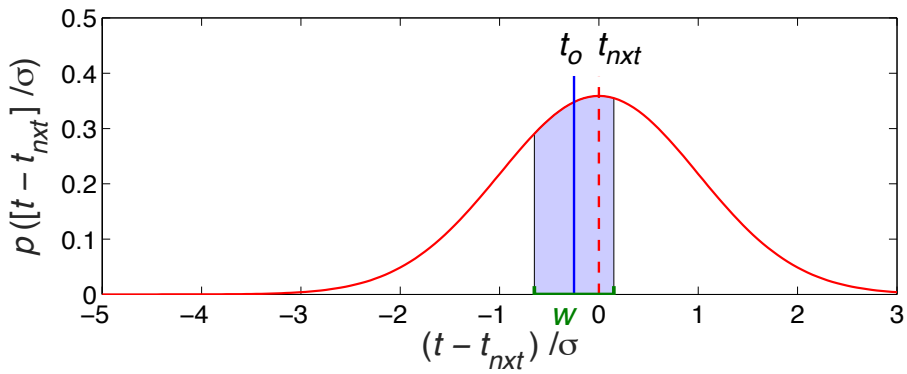

Figure 8. The forecast probability density function (pdf), *p*(*t*) with normalized time. (Modified from Nava et al., 2014).

The length of the finite interval, wherein the probability is evaluated, could take any value that meets $w \le \sigma/4$, because, for these values, the differences in probability estimations are ~10^{-4}. For $w \le \sigma/10$, the probability can be considered independent of *w*, and for practical purposes $w \le \sigma/40$ is used.

The updated non-randomness probability, P_c^U, is estimated through (10) and (11), for the whole series including all events up to and including the aftcast one.

The validity of the proposed Bayesian analysis can be judged by application to *aftcasts*. An aftcast is the forecast of an event that has already occurred, on the basis of information previous to it, i. e. a *forecast a posteriori*. For several aftcast cases Nava et al. (2014) showed that the posterior probability $\Pr\left(A|B\right)$ agrees quite well with the updated P_c^U, the new non-randomness probability estimated from the whole data set that includes the aftcast event and other events occurred since the aftcast.

Aftcasting the last event of an identified sequence can be used to judge how well the forecasting method works and whether the sequence is indeed semi-periodic or not. The Bayes approach can be also used to judge whether a given earthquake can be the forecast one.

Chapter 6. Results and discussions

6.1 Synthetic data

The analysis method for identification of semi-periodic behavior in time series was tested using synthetic data, in order to see whether the method is able to correctly identify known sequences and to evaluate how reliable this identification is.

Since the number of events within the time series has significant influence on the effectiveness of the analysis, many tests were performed using different time series, each one constituted by different numbers of events and sequences.

Each time series was constructed as follows; we began by specifying:

1) A tentative duration time, *T*, of the whole time series.

2) The number of sequences.

3) The number of events, *K*, in each sequence.

4) The standard deviation of the θ variation from periodicity for each sequence.

5) The magnitude range for each sequence.

6) The number of extraneous events, and their magnitude range.

For each sequence, the period was obtained randomly, with uniform probability, from those satisfying *T* and *K*, $T/K+\xi \le \tau \le T/\left(K-1-2r\right)$, with $\xi \ll T/K$ (Section 4.4) and $r=1/6$ corresponding to the stricter comb-event closeness criterion (Section 4.1). Next, the time of the first tooth was chosen randomly, among the eligible ones, with uniform probability. Then, variations were applied to each event time, using a normal distribution with the specified standard deviation and truncated so that variations were within the maximum allowable variation $\theta_{\max}=r\tau$. If necessary, the tentative time interval was adjusted so that all events lay within it, to obtain the final time interval *T*. Next, magnitudes were

assigned randomly to each event, according to the G-R distribution, with a representative *b*-value = 1, for the sequence magnitude range.

Finally, the extraneous events were generated, with uniform probability over the whole time interval, and assigned G-R magnitudes over a range chosen to include the lowest magnitudes in the sequence ranges and slightly higher magnitudes than those in the ranges, to mimic the occurrence of very large earthquakes belonging to sequences not identifiable from the duration time.

For all cases, we used *T*=125 years, which corresponds, approximately, to the time for which data are complete and homogeneous in the seismic catalogs. We first tried time series comprised of one sequence of five events and three extraneous events, with magnitudes in the range from 7.4 to 8.1 (Figure 9).

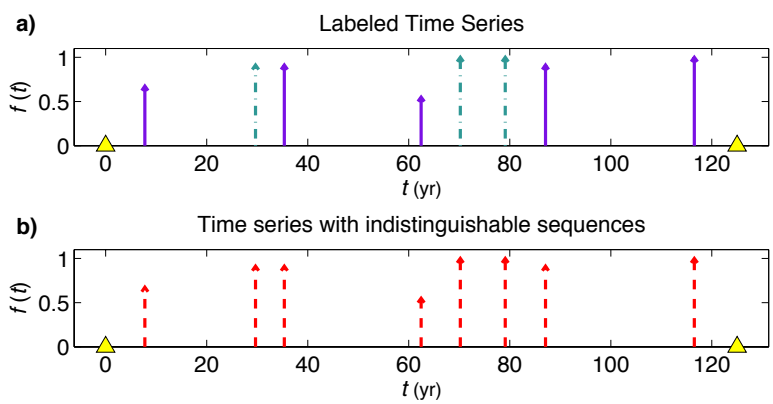

Figure 9. a) Example of the synthetic data labeled time series, constituted by one sequence of five events (violet arrows) and three extraneous events (dot-dash green arrows). b) The labeled time series with indistinguishable events (dashed red arrows); the yellow triangles indicate t_0 and t_f.

In twenty-five of thirty random realizations, the correct sequence was identified (Figure 10). Five realizations gave approximately correct results where the sequences' periods were identified, but one of the events in the sequence was substituted for an extraneous one (Figure 11).

The example in Figure 11 shows the identified sequence, where the third event of the sequence, that occurred at $t_3 = 57.1601$ yr, was substituted by the extraneous event close to it in time, occurred at $t_4 = 60.1534$ yr. The right event was replaced by the extraneous one because both events are so close that the time difference between them is less than the maximum expected variation of the sequence period.

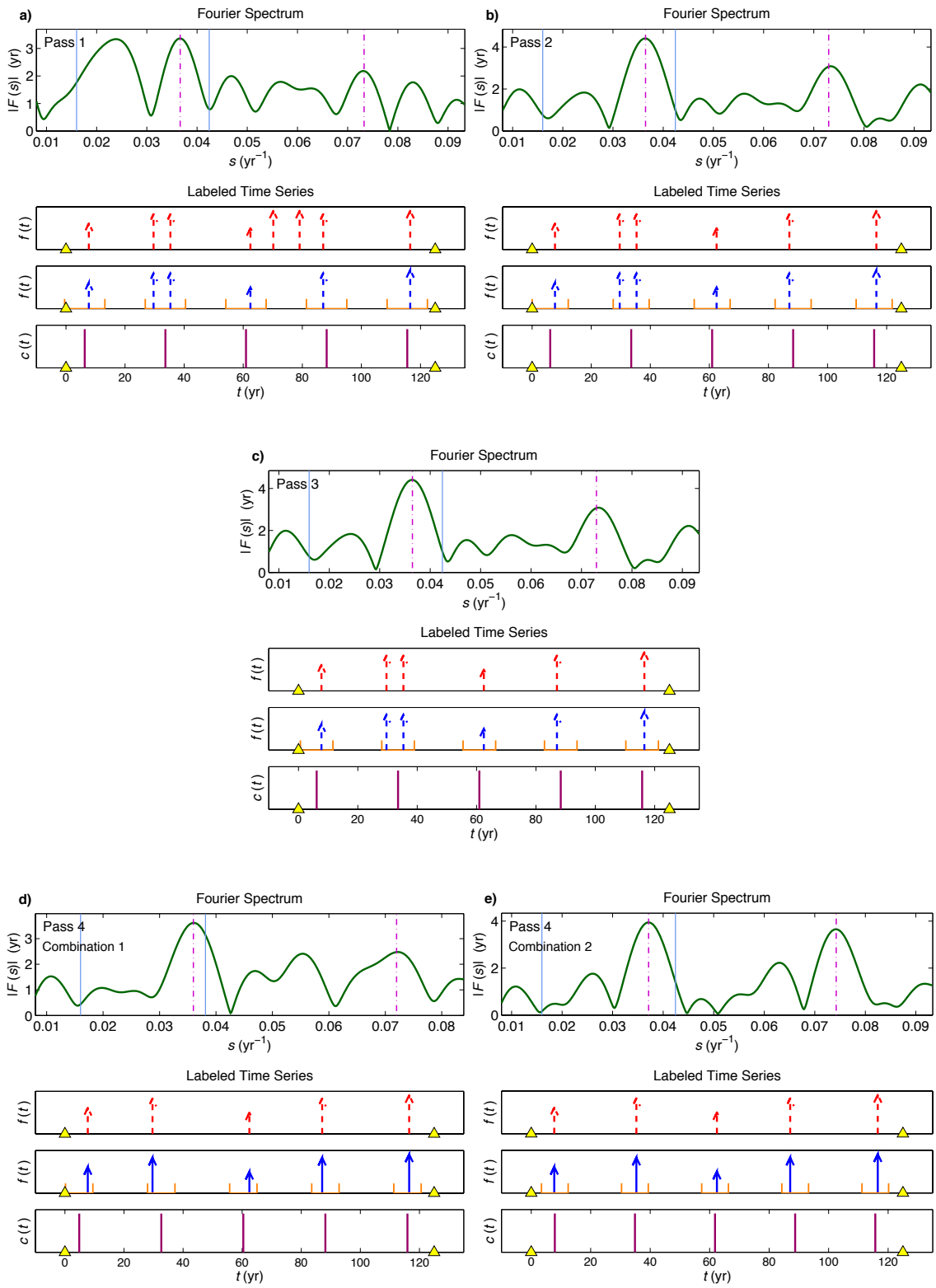

Figure 10. Spectral analysis of the time series shown in Figure 9, for each pass. Two combinations were evaluated in pass 4 and the Combination 2 was the fittest one and that corresponds to the predefined sequence. The image conventions are like those in Figure 5.

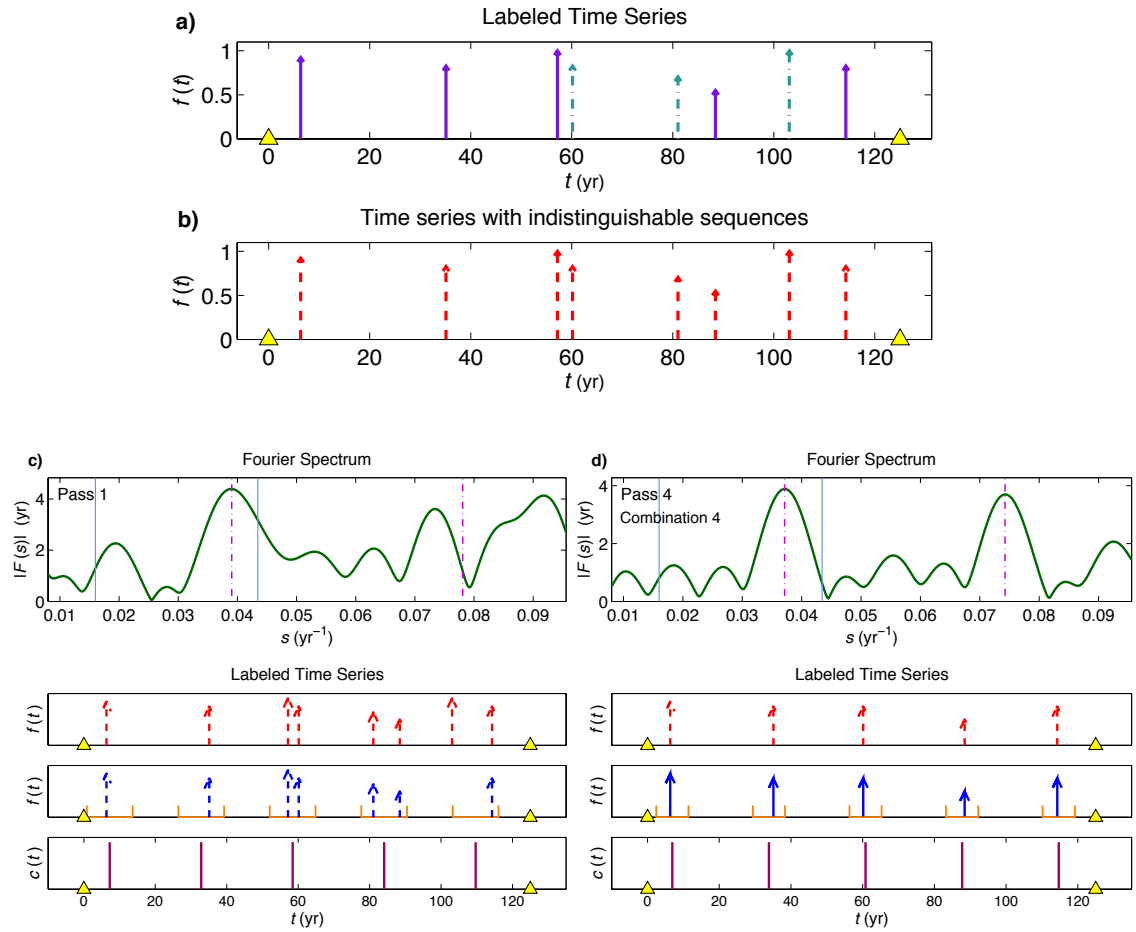

Figure 11. Example of identified sequence where the third event of the sequence was substituted by another very close to it. a) Synthetic data labeled time series constituted by one sequence of five events (violet arrows) and three extraneous events (dot-dash green arrows). b) The labeled time series with indistinguishable events (dashed red arrows). c) and d) Passes 1 and 4 of the spectral analysis, respectively. Combination 4 is the fittest in the fourth pass.

However, the differences between the periods estimated including the replaced events and the correct periods were so small that the forecast of the next events of the identified sequences were, for all practical purposes, the same as for the correct sequences.

The correct forecast corresponding to the original sequence is $t_{nxt} = 140.9213 \pm 3.5111$ yr (considering 2σ), and the forecast of the identified sequence is $t_{nxt} = 141.6194 \pm 2.7323$ yr. The difference between both forecasts (0.6981 yr) is insignificantly small with respect to the period of the sequence, which is $\tau = 27.0258$ yr.

Next, we tried randomly generated time series including two sequences and one extraneous event (Figure 12). Thirty realizations were done, of which nineteen correctly identified the sequences (Figure 13), ten resulted in wrong sequences, eleven realizations resulted in sequences in which some event belonging to the sequence was substituted with an extraneous event (Figure 14), and one did not show semi-periodicity at all.

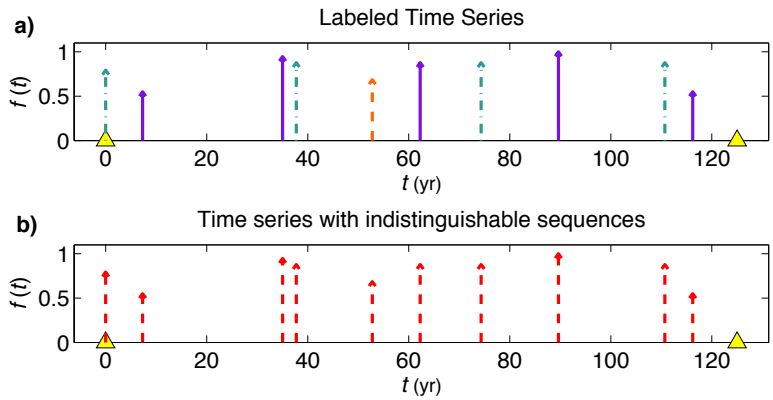

Figure 12. a) Example of the synthetic data labeled time series, constituted by one sequence of five events (violet arrows), one sequence of four events (dot-dash green arrows), and one extraneous event (dashed orange arrow) b) The labeled time series with indistinguishable events (dashed red arrows).

The example in Figure 14 shows the first identified sequence, in which the third event of the sequence, occurred at $t_6 = 70.3686$ yr, was substituted by the extraneous event close to it, occurred at $t_5 = 67.2875$ yr. The correct forecast corresponding to the original sequence is $t_{nxt} = 149.9736 \pm 2.8810$ yr (considering 2σ), and the forecast of the identified sequence is $t_{nxt} = 148.5425 \pm 2.3814$ yr.

These results show that, if the number of events in a time series increases, the chance to find events as close together as to be interchangeable also increases. However, it should be remembered that the synthetic series include random variations from strict periodicity, so that if a given event is substituted by an extraneous one within a sequence but the correct period is identified, then the resulting forecast would be correct too, so that the method is performing correctly.

While any number of sequences may be postulated, since reliable observations are available for only ~125 years, the large number of events required by three or more series (of at least four events each) plus possible extraneous events, results in such a large density of events that the probability of finding apparent sequences occurring by chance is too high. Hence, we consider only the possibility of one or two sequences.

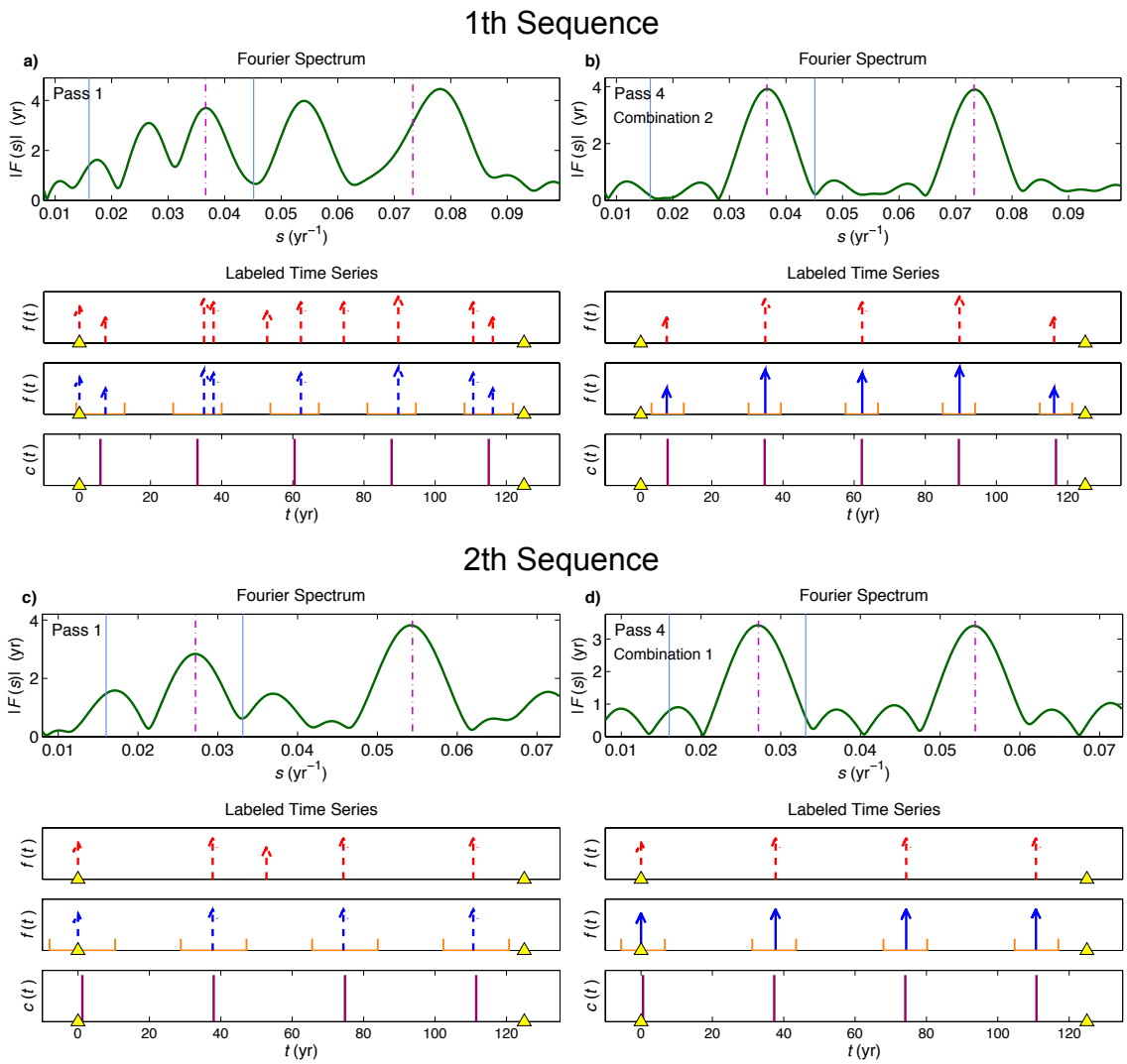

Figure 13. Two semi-periodic sequences identified through spectral analysis a) First pass of the analysis with all events in the time series. b) The fittest combination in the fourth pass corresponding to semi-periodic sequence with five events. c) First pass of the spectral analysis of the time series excluding events belonging to the previous identified sequence. d) The fittest combination in the fourth pass corresponding to the second semi-periodic sequence.

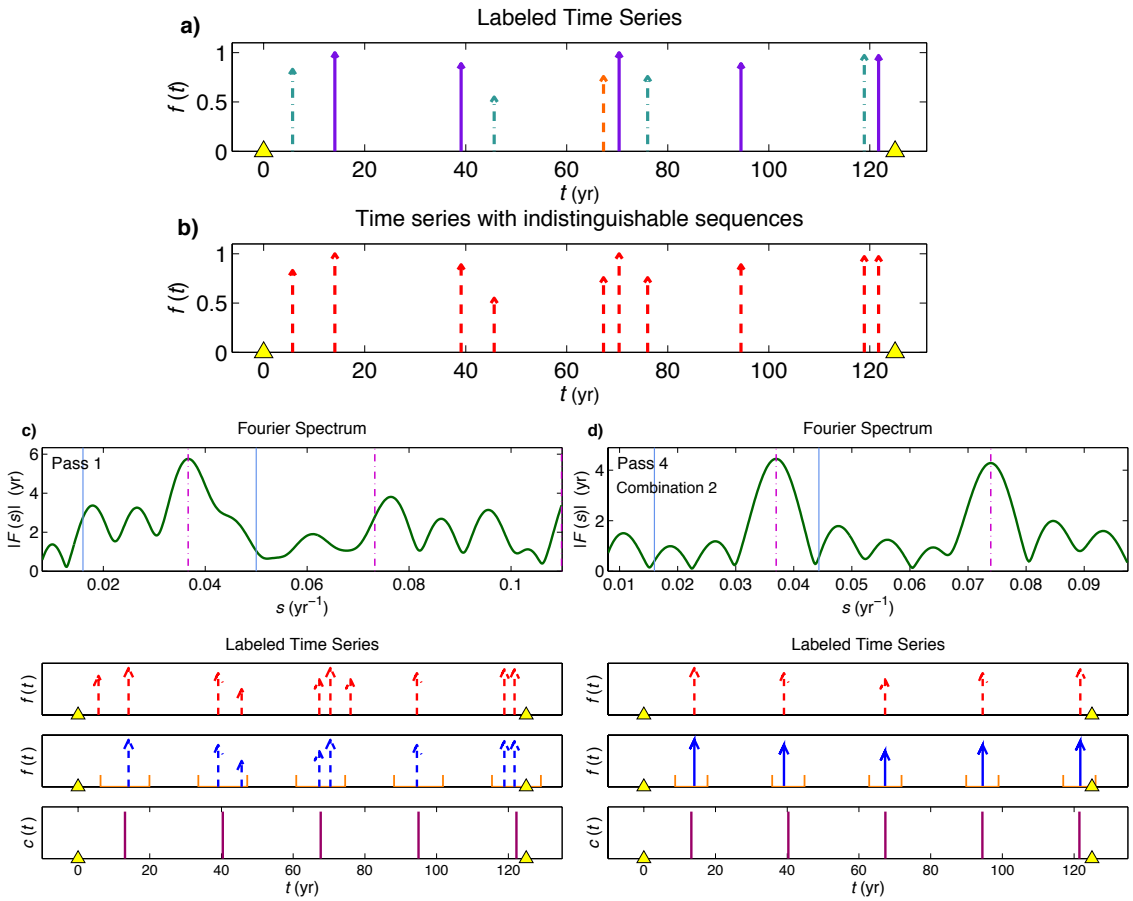

Figure 14. Example of identified sequence where the third event of the sequence of five events was substituted by another very close to it. a) Synthetic data labeled time series constituted by one sequence of five events (violet arrows), one sequence of four events (dot-dash green arrows), and one extraneous event (dashed orange arrow) b) The labeled time series with indistinguishable events (dashed red arrows). c) and d) Passes 1 and 4 of the spectral analysis, respectively. Combination 2 is the fittest in the fourth pass.

6.2 Applications to real data

6.2.1 Japan

6.2.1.1 Study region and data

The study region was defined considering the tectonic setting (Nakamura, 1983; Tamaki and Honza 1985; Taira, 2001; Bird, 2003) and the epicentral distribution of large earthquakes in Japan. The region is localized in the Northeastern Japan Arc (Figure 15), where the major seismic activity of Japan occurs as a result of the subduction of the Pacific plate beneath the Japan and Kuriles arcs, at a rate of about ~8.5 cm/yr towards the WNW.

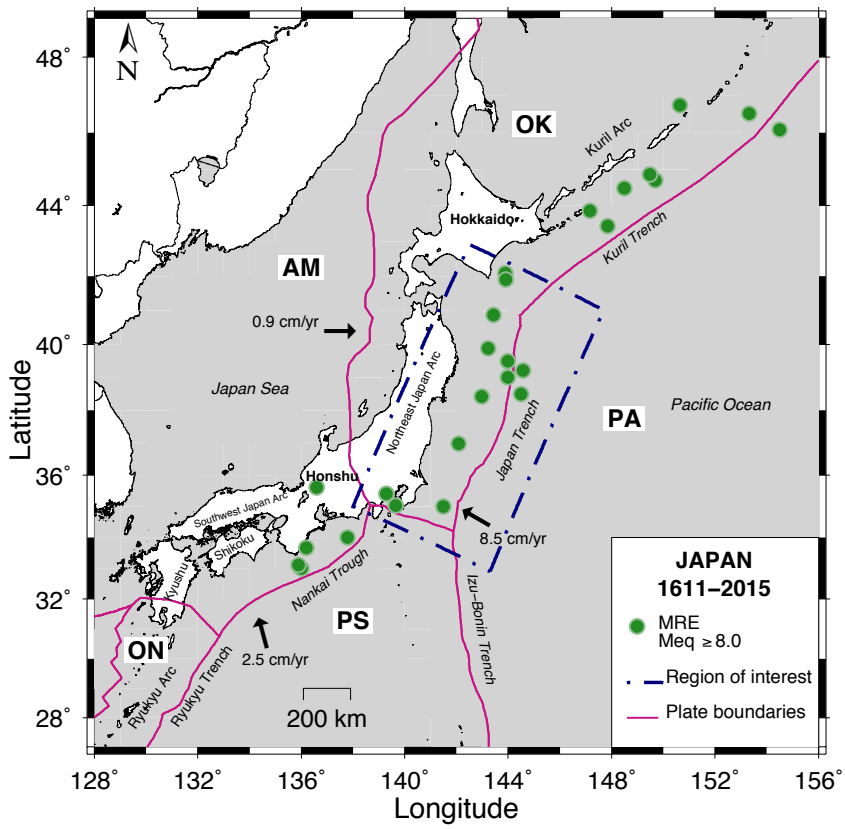

Figure 15. Map of MREs with $M_{eq} \geq 8.0$, occurred in Japan from 1611 to 2015.5 (ISCS-GEM Catalogs). Green circles indicate epicenters of the main earthquake of the MRE. Dashed blue lines enclose the study region. Solid purple lines indicate the boundaries between the Pacific (PA), Philippine (PS), Eurasia/Amur (AM), and North America/Okhotsk (OK) tectonic plates (Bird, 2003).

The Moment Release Episodes processing (MRE, section 3.3) was used to calculate equivalent magnitudes, M_{eq} (6). In some cases, M_{eq} resulted only slightly higher than the magnitude of the main earthquake (~0.01 higher than previous magnitude), but for the MRE occurred in 1938.84, which included events that occurred close in time and space with magnitudes 7.8, 7.7, 7.7, 7.6, 6.5 and 6.8, the equivalent magnitude, 8.1, resulted notably higher than the 7.8 magnitude of the main earthquake. Only MREs with $M_{eq} \geq 8.0$ were included in the analysis.

There is information about earthquakes that occurred in the study region since 1611 (Table 1), but as can be seen in Figure 16a, there are conspicuous gaps in the data before 1896, which we interpret as missing data. Thus, the earthquakes included in the time series occurred from 1896 to 2015.5; the catalog may be considered complete from this date on for MREs with $M_{eq} \geq 8.0$ (Figure 16b).

Magnitude weights (5), from 0.55 to 1.00, were assigned to events occurred from 1896 to 2015.5 with magnitudes between 8.0 and 9.1 (Table 1), using a *b*-value of 0.93. This *b*-value (4) was estimated from 1820 earthquakes occurred between 1949-2015, with $5.6 \geq M_w \geq 8.2$ for which the G-R histogram behaved linearly. It should be pointed out that the actual *b*-value is not critical since weights do not vary greatly for different, reasonable, *b*-values.

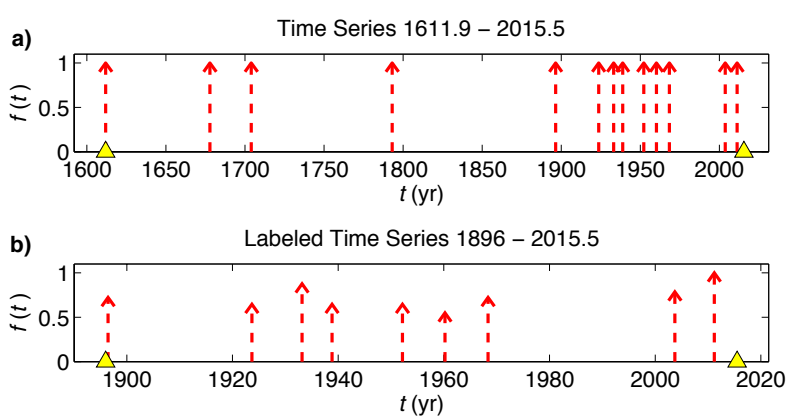

Figure 16. MREs occurrence time series, indicated by red arrows, for events with $M_{eq} \geq 8.0$ in the study area of Japan. a) Occurrence time series of events from 1611.9 to 2015.5. b) Labeled occurrence time series of events from 1896 to 2015.5; the yellow triangles indicate t_0 and t_f.

Table 1. MREs occurred in the study region of Japan, with $M_{eq} \geq 8.0$, from 1611 to 2015.5 (ISC-GEM Catalogs).

No	Time (yr)	Latitude (°)	Longitude (°)	M_{eq}	Depth (Km)	Weight $w(M_{eq})$
1	1611.9194	39.00	144.00	8.1		
2	1677.8434	35.00	141.50	8.0		
3	1703.9973	35.03	139.66	8.2		
4	1793.1304	38.50	144.50	8.2		
5	1896.4558	39.50	144.00	8.2		0.72
6	1923.6658	35.41	139.30	8.1	15.00	0.65
7	1933.1644	39.21	144.59	8.5	15.00	0.88
8	1938.8438	36.97	142.09	8.1	35.00	0.65
9	1952.1721	42.08	143.90	8.1	45.00	0.65
10	1960.2159	39.87	143.23	8.0	15.00	0.55
11	1968.3716	40.86	143.44	8.2	29.90	0.72
12	2003.7315	41.90	143.92	8.3	30.00	0.79
13	2011.1890	38.28	142.55	9.1	20.00	1.00

6.2.1.2 Analysis results

Labeled time series 1896-2015.5

In the analyzed time series there are $N = 9$ events (Figure 16b) that occurred over 119.5 years $[t_0 = 1896, t_f = 2015.5]$. The comb determined in the first pass (Figure 17a) was constituted by $K = 4$ teeth with period $\tau_c = 37.8165$ yr ($s_c = 0.0264$ yr^{-1}). For the second pass of the analysis Event 9 (all event numbers refer to Table 1) was dropped from the time series, because it did not satisfy the comb-event closeness criterion. For the second pass (Figure 17b), the period of the comb was $\tau_c = 37.5786$ yr ($s_c = 0.0266$ yr^{-1}), and event 10 was dropped from the time series for the next pass.

In the third pass (Figure 17c) the comb period was the same as that obtained in the previous pass, but event 6 was dropped from the analysis, and the six events that satisfied the comb-event closeness criterion: 5, 7, 8, 11, 12, and 13, became candidates for the four-event sequence. These six candidate events resulted in four combinations of four events (Figure 18, Table 2), and the fourth pass was carried out for each combination.

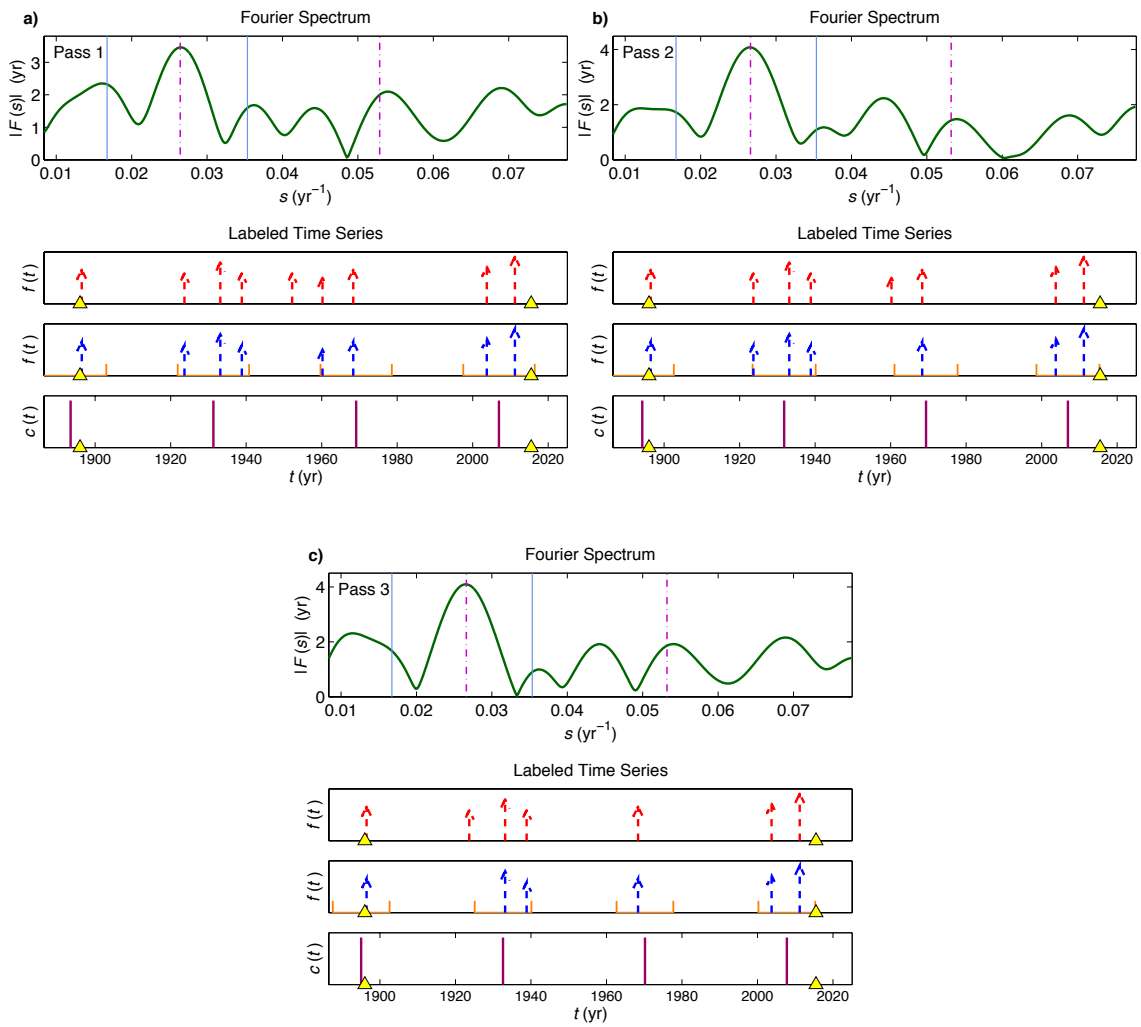

Figure 17. Spectral analysis of Japan data from 1896 to 2015.5. a) First pass, b) Second pass, and c) Third pass. Conventions are the same as in previous spectral analysis figures: (Top) Fourier spectrum of the time series shown immediately below; the vertical dot-dashed purple lines, at the left, indicate the chosen frequency and one of its multiples at the right; the solid light blue vertical lines are the guideline frequencies $s_{\min}$ and $s_{\max}$ (Nava et al., 2014). (Middle) the labeled time series where dashed red arrows represent each event, and the $f(t)$ bellow shows as dashed blue arrows those events that are within the confidence interval (solid orange lines) and could belong to the sequence. (Bottom) $c(t)$ is the comb, whose teeth are shown by solid dark purple lines. The yellow triangles indicate t_0 and t_f.

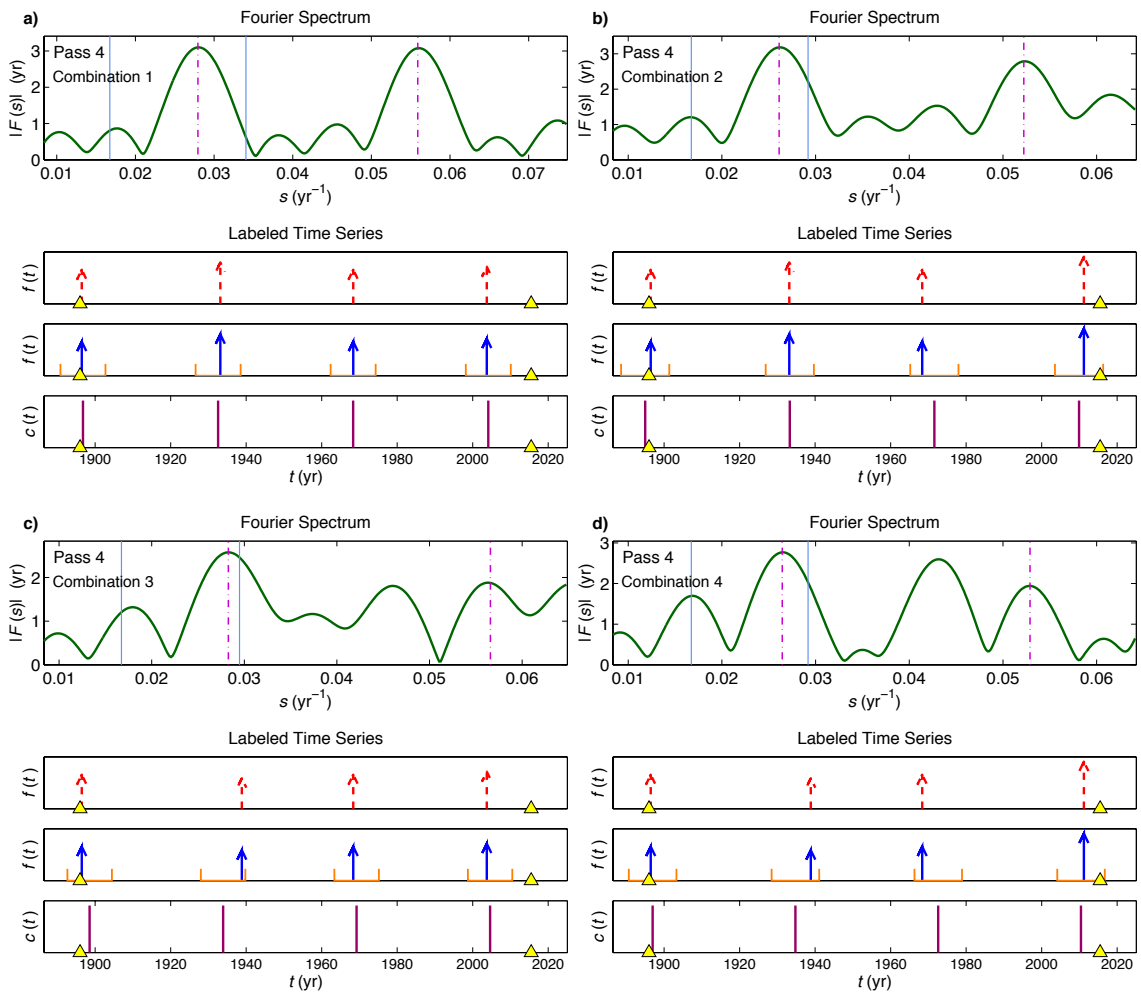

Figure 18. Spectral analyses of the four combinations in the fourth pass to identify the semi-periodic sequence for Japan region. The chosen semi-periodic sequence corresponds to combination 1, which has the lowest weighted fitting error. Conventions are like those in Figure 17.

Table 2. Event combinations analyzed in pass 4. The shaded combination is the selected semi-periodic sequence.

Combination	Events (yr)	Weighted fitting error (yr)	Combination	Events (yr)	Weighted fitting error (yr)
1	5 1896.45578 7 1933.16438 11 1968.37158 12 2003.73151	1.3686	2	5 1896.45578 7 1933.16438 11 1968.37158 13 2011.18904	6.3729
3	5 1896.45578 8 1938.84384 11 1968.37158 12 2003.73151	8.8690	4	5 1896.45578 8 1938.84384 11 1968.37158 13 2011.18904	10.0355

The selected sequence was that of combination 1, which had the lowest weighted fitting error (Table 2), with a comb period $\tau_c = 35.7784$ yr ($s_c = 0.0280$ yr^{-1}), starting at $t_{0c} = 1896.7864$ yr ($\phi_c = -0.1381$). The measured standard deviation of θ was $\hat{\sigma} = 0.3943$ yr, and the estimated distribution standard deviation was $\sigma = 0.7646$ yr.

The epicenters of the events constituting the identified sequence are shown in the Figure 19, together with all other MRE with $M_{eq} \geq 8.0$ in the region. The epicenters of the events in the sequence are clustered in the NE part of the study region.

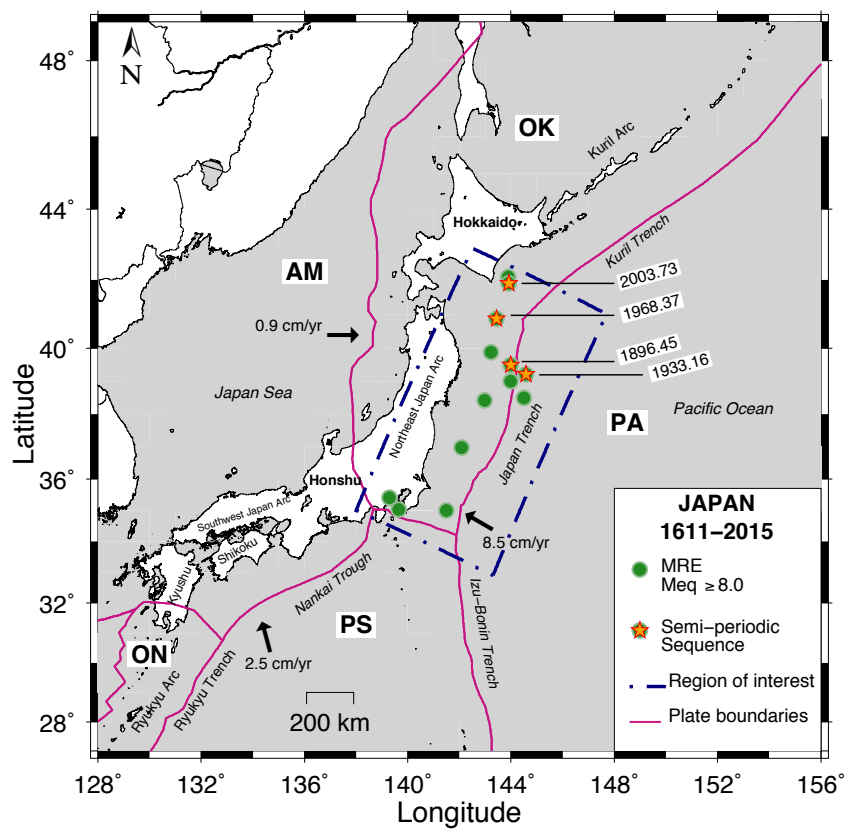

Figure 19. Map of MREs with $M_{eq} \geq 8.0$, occurred in study region of Japan from 1611 to 2015.5 (ISCS-GEM Catalogs). Orange stars indicate epicenters of the events belonged to the semi-periodic sequence, and green circles show epicenters of that not belonging to it. The image conventions are the same as in Figure 15.

We would expect the next $M_{eq} \geq 8.0$ event belonging to the sequence around $t_{nxt} = 2039.9001 \pm 1.5292$ yr (with 95.45% confidence, $q = 2$). The probability of random occurrence of the semi-periodic sequence is $P_0 = 0.008$, so that the sequence/comb probability is $P_c = 0.992$.

The identified sequence and the pdf of the forecast, $p(t)$, are presented graphically in Figure 20, together with the survival and the hazard functions (Nava et al., 2014) that help to visualize some consequences of the forecast.

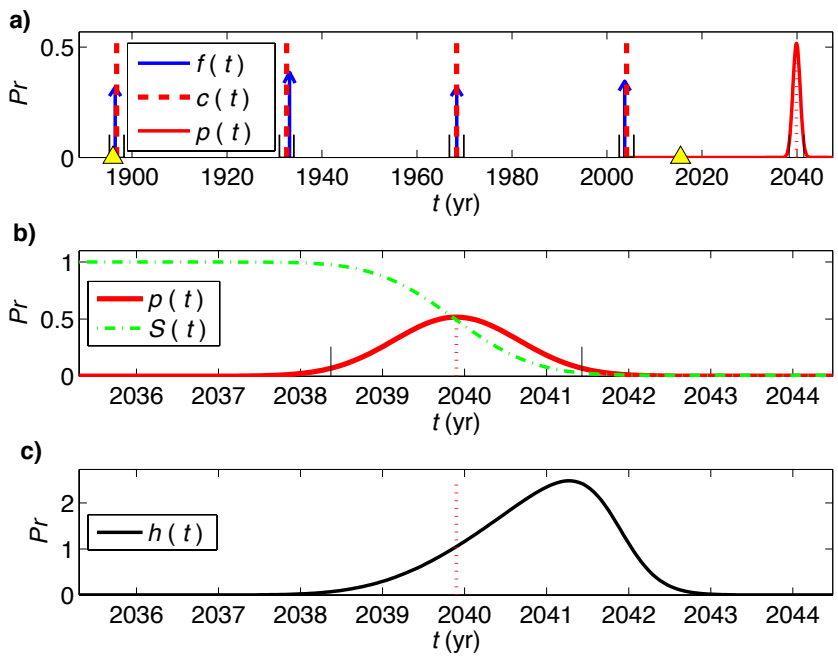

Figure 20. Japan forecast. a) Forecast based on the sequence identification shown in Figure 18a; same convention as in Figure 7. b) Close-up of the pdf, $p(t)$, red curve; and the survival function of the forecast, $S(t)$, dot-dash green curve. c) Hazard function $h(t)$.

The occurrence probability of the next event belonging to the sequence within a $\pm 2\sigma$ interval, centered on time t_{nxt}, is $P_{cq} = 0.947$; the Poissonian occurrence probability of at least one event within a $\pm 2\sigma$ interval, considering the occurrence rate of the whole time series $\lambda = 0.075/yr$, is $\pi_{1+} = 0.206$ the Poissonian occurrence probability of at least one event within $\pm 2\sigma$, considering the occurrence rate of events not belonging to the sequence, $\lambda^* = 0.042/yr$, is $\pi^*_{1+} = 0.120$. The probability gains, evaluated from these probabilities through eq. 11 (section 5.1), for different values of *q*, are significant (Table 3).

Table 3. Probabilities, probability gains and information gains for the Japan sequence.

q	P_{cq}	π_{1+}	π^*_{1+}	P_G	I_G (bits)
1	0.677	0.109	0.062	6.410	2.680
2	0.947	0.206	0.120	4.634	2.212
3	0.989	0.292	0.175	3.393	1.763

Unlabeled vs. Labeled time series

Results obtained using the unlabeled and labeled time series analyses are compared in order to know whether the influence of the earthquakes magnitude on the analysis leads to significant differences between the forecasts.

Quinteros et al. (2014) analyzed the unlabeled time series for the Japan region with data from an older seismic catalog than the one used in the present work. However, to correctly compare results, we used the same data both in the unlabeled and the labeled analyses.

Hence, the analysis of the unlabeled time series is done here using the Japan data from the ISC-GEM catalog (Table 1) as in the labeled analysis, but without weights (Figure 21), which is the same as assigning the same magnitude to all events. So that, the unlabeled time series is constituted by $N = 9$ events that occurred over 119.5 years $[t_0 = 1896, t_f = 2015.5]$.

The semi-periodic sequence identified in the unlabeled analysis is constituted by $K = 4$ events, which are the same as the sequence identified in the labeled analysis except for the last event. In the labeled analysis the last event of the sequence is event 12 (all event numbers refer to Table 1) and the unlabeled analysis is event 13.

This difference is because in the third pass of the unlabeled analysis (Figure 22), the selected events to be used in the last pass are only the closest ones to the comb, identification is limited to only one combination of events, and in this case, event 13 is slightly closer to the comb than event 12. Since in the labeled analysis every possible combination is evaluated, not only by the closeness of time (comb-event) criterion, but also by the closeness of magnitudes in the sequence criterion, the combination that included event 12 ($M_{eq} = 8.3$) fitted the sequence better than the combination that included event 13

(Figure 23), whose magnitude ($M_{eq} = 9.1$) is notably higher than those of the other events in the sequence.

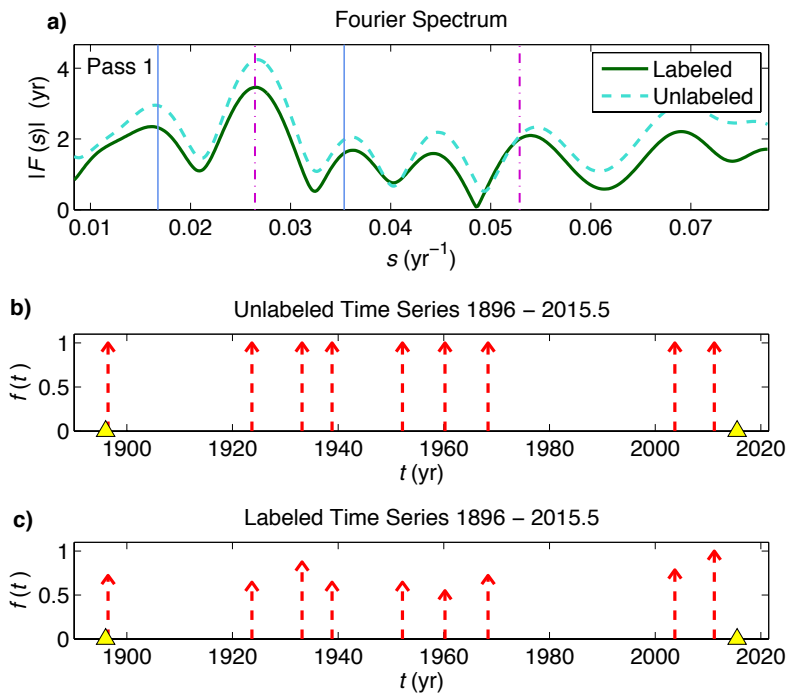

Figure 21. a) Comparison between the unlabeled and labeled spectrum analysis (pass 1) of Japan data; b) and c) are their respective time series. Conventions are like those in previous figures.

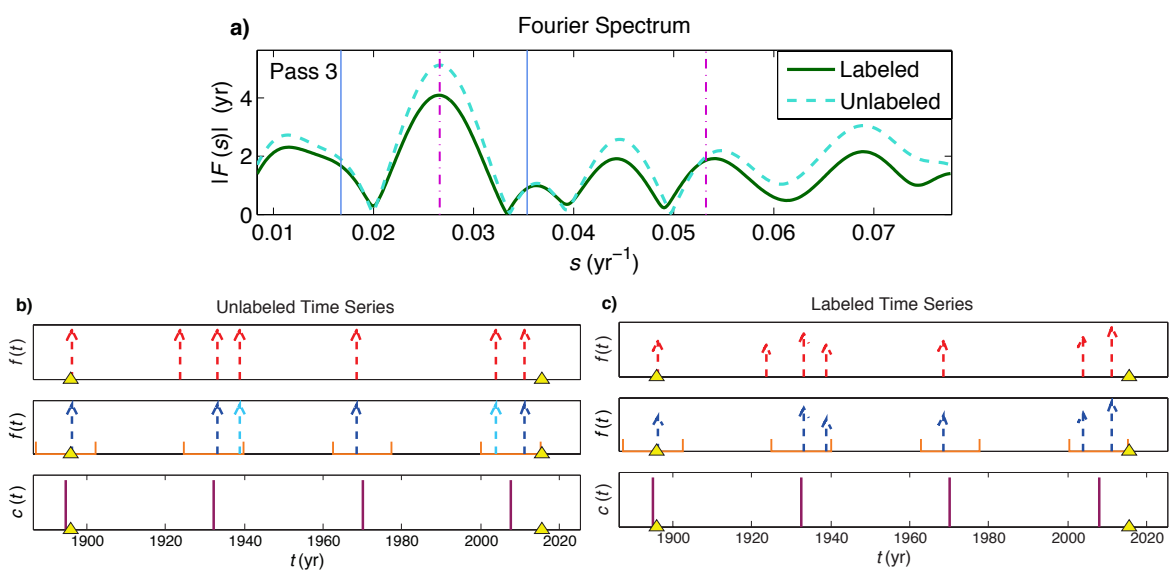

Figure 22. a) Comparison between the unlabeled and labeled spectrum analysis (pass 3) of Japan data. In the unlabeled analysis (b), the selected events to be analyzed in the pass 4 are the closest events (dark blue dashed rows) to the comb (purple lines), while in the labeled analysis (c), the events (blue dashed rows) within the acceptance interval (dark green lines) are combined in different possible sequences to be analyzed in pass 4.

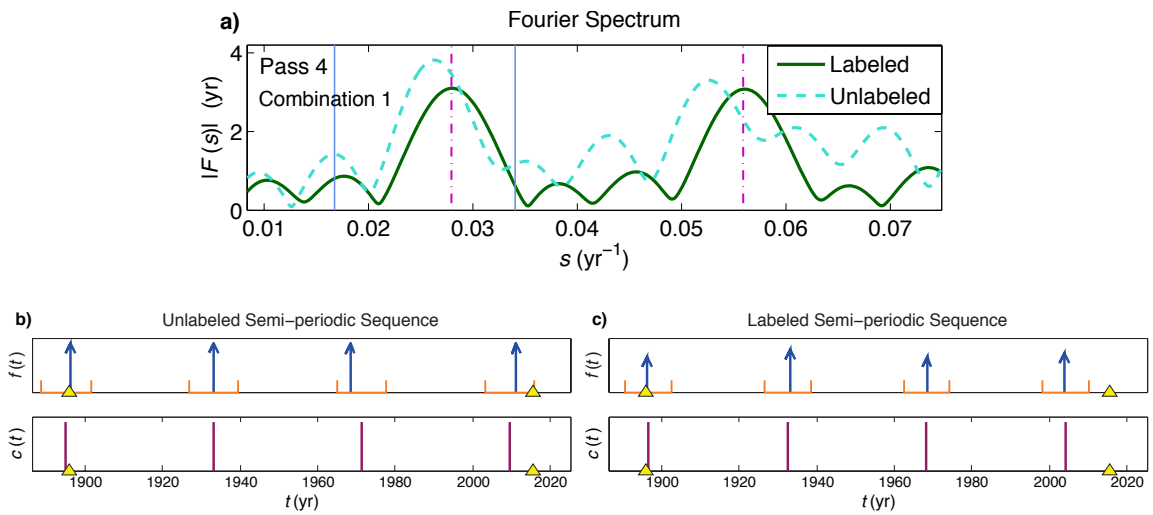

Figure 23. a) Comparison between the unlabeled and labeled spectrum analysis (pass 4) of Japan data. In b) and c) the identified sequence (dark blue dashed rows) and the comb (purple lines) are shown for the unlabeled and the labeled cases, respectively.

Thus, in the unlabeled analysis, the resulting comb period is $\tau_c = 38.0573$ yr, the phase is $\phi_c = 0.1271$, so that the origin time of the comb is $t_{0c} = 1895.2302$ yr. The difference between the period of the sequences obtained from the two different analyses is ~ 2.28 yr (~0.06 τ_c).

In the unlabeled series analysis the measured standard deviation, $\hat{\sigma} = 1.8405$ yr, was used as the standard deviation for the forecast; but, since it is better to use the estimated population standard deviation, we will do so for both methods. For the unlabeled method, the estimated population standard deviation is $\sigma = 3.5693$ yr and the forecast is $t_{nxt} = 2047.4595 \pm 7.1386$ yr (with 95.45% confidence, $q = 2$). The difference between the forecast times estimated from the two different analyses is ~ 7.56 yr, which is not very high compared to the unlabeled forecast confidence interval.

The probability of random occurrence of the sequence, considering the estimated standard deviation, is $P_0 = 0.268$, so that the sequence/comb probability, $P_c = 0.732$, is smaller then the P_c obtained in the labeled analysis. The results of the analyses, both unlabeled and labeled, are shown for comparison in Table 4.

Table 4. Results obtained with the labeled and unlabeled analyses for Japan region.

	Sequence	$\tau\,(yr)$	ϕ_c	t_{0c} (yr)	$\hat{\sigma}$ (yr)	σ (yr)	t_{nxt} (yr)	P_c	$P_{cq}\left(\pm 2\sigma\right)$
Unlabeled	1896.45578 1933.16438 1968.37158 2011.18904	38.0573	0.1271	1895.2302	1.8405	3.5693	2047.4595	0.732	0.698
Labeled	1896.45578 1933.16438 1968.37158 2003.73151	35.7784	-0.1381	1896.7864	0.3943	0.7646	2039.9001	0.992	0.947

While most comparisons of unlabeled versus labeled analyses gave very similar results, this example is particularly important, because it clearly shows how simple proximity in time to a comb's tooth is too simplistic a criterion for choosing a sequence event. The choice of an alternative event that is not quite as close to the tooth, but has a magnitude that agrees better with the other magnitudes in the sequence, yields a period that is a much better fit to the sequence events, which results in a much better forecast.

Aftcasts and Bayesian analysis

We will now present an example of an aftcast and its subsequent Bayesian evaluation.

In the forecasting example presented above, a sequence of four events was identified from events occurring from 1896 to 2015.5; we will try to aftcast the last of the sequence events, occurred at 2003.7315, using only those events that occurred before it, between $t_0 = 1896$ yr and $t_f = 1969$ yr.

The time series is constituted by $N = 7$ events (Figure 24) that occurred over 73 years. Magnitude weights, from 0.55 and 1.00 (Table 5), were assigned to magnitudes between 8.0 and 8.5.

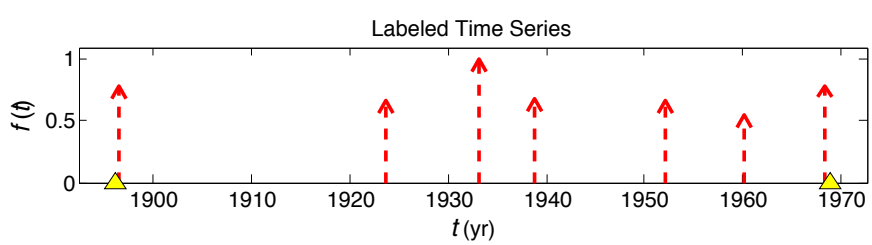

Figure 24. MREs occurrence labeled time series, indicated by red arrows, for events with $M_{eq} \geq 8.0$ in the study region of Japan, from 1896 to 1969.

Table 5. MREs used in the afcast analysis, occurred in the study region of Japan, with $M_{eq} \geq 8.0$, from 1986 to 1969 (ISC-GEM Catalogs).

No	Time (yr)	Latitude (°)	Longitude (°)	M_{eq}	Depth (Km)	Weight $w(M_{eq})$
1	1896.4558	39.50	144.00	8.2		0.79
2	1923.6658	35.41	139.30	8.1	15.00	0.68
3	1933.1644	39.21	144.59	8.5	15.00	1.00
4	1938.8438	36.97	142.09	8.1	35.00	0.68
5	1952.1721	42.08	143.90	8.1	45.00	0.68
6	1960.2159	39.87	143.23	8.0	15.00	0.55
7	1968.3716	40.86	143.44	8.2	29.90	0.79

Four passes were done (Figure 25), with two combinations considered in the fourth pass (Figure 25d and 25e). The selected sequence is that in Combination 2, with period $\tau_c = 36.1386$ yr ($s_c = 0.0277$ yr^{-1}), starting at $t_{0c} = 1896.5663$ yr ($\phi_c = -0.0985$). The measured standard deviation of θ is $\hat{\sigma} = 0.3856$ yr, and the estimated distribution standard deviation is $\sigma = 0.8737$ yr.

Thus, we would expect the next $M_{eq} \geq 8.0$ event belonging to the sequence around $t_{nxt} = 2004.9822 \pm 1.7474$ yr (with 95.45% confidence, $q = 2$). The probability of random occurrence of the sequence is $P_0 = 0.074$, so that the sequence/comb probability is $P_c = 0.926$.

The events belonging to the identified sequence are the same as the first three events belonging to the identified sequence in the previous analysis (1896-2015.5), and the aftcast clearly coincides with the event occurrence in 2003.7315 yr. The aftcast error with respect to the actual occurrence time of the event is 1.25 yr, less than the time given in the confidence interval; so that, the aftcast is considered successful.

The identified sequence and the pdf of the forecast, $p(t)$, are presented graphically in Figure 26, together with the occurrence time of the event that occurred in 2003.7315, just when the survival function begins to decline and well before the hazard function of the forecast reaches its maximum value.

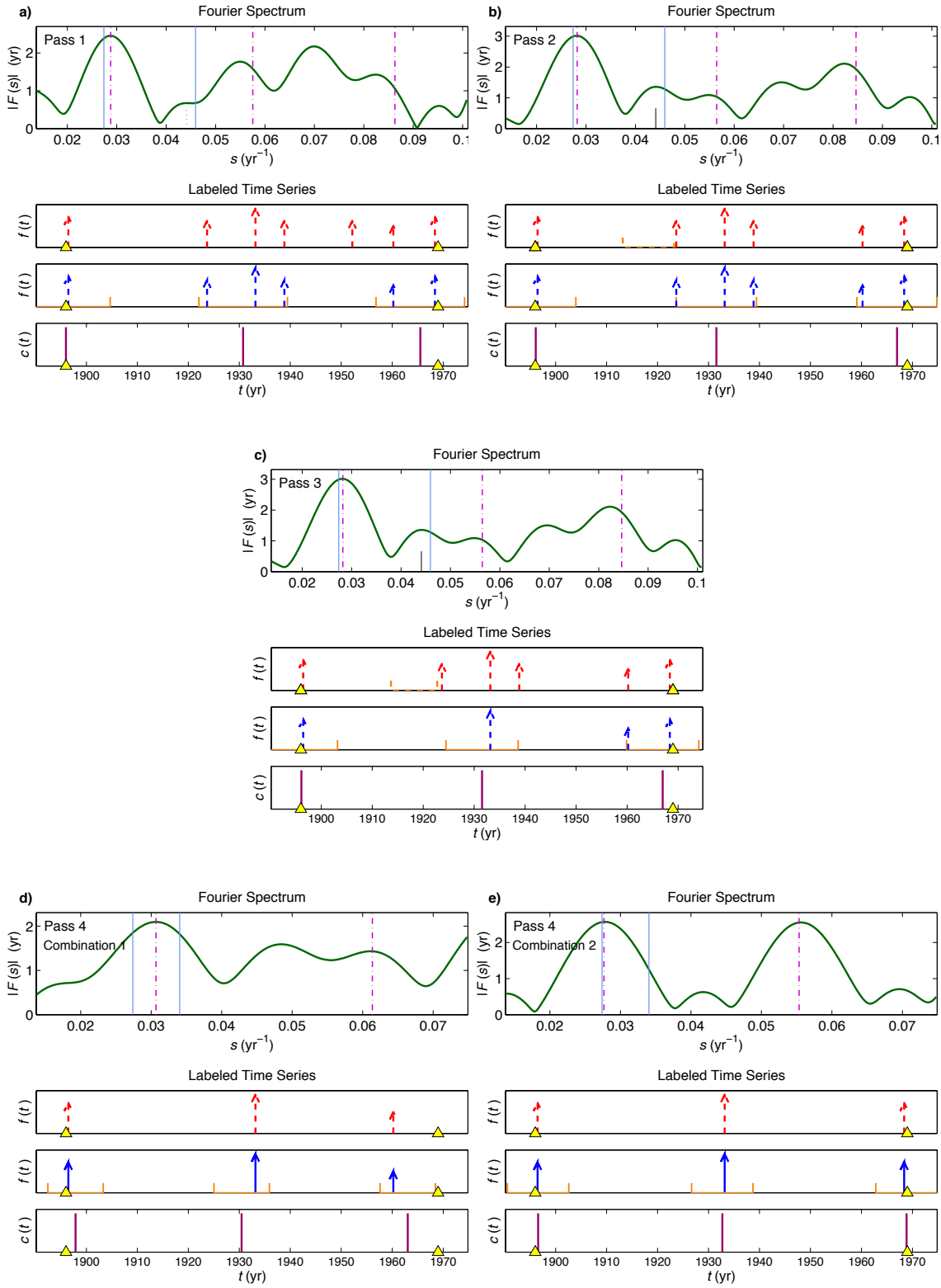

Figure 25. Spectral analysis of data from Japan from 1896 to 1969. a) First pass, b) Second pass, c) Third pass, and in d) and e) Two combinations in the Fourth pass. Conventions are the same as in previous figures. The selected sequence is that in Combination 2.

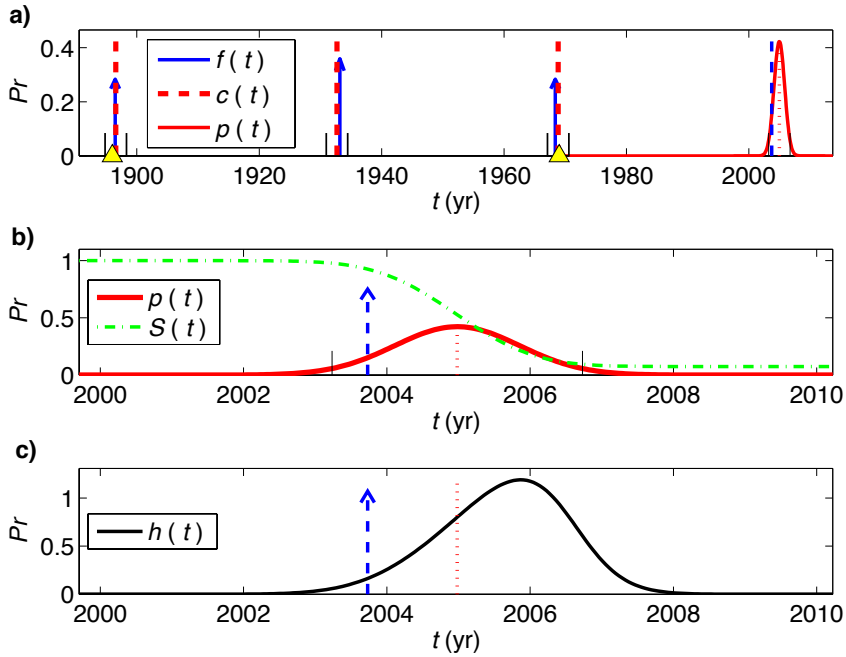

Figure 26. Japan aftcast. a) Aftcast based on the sequence identification shown in Figure 25; same convention as in Figure 20. b) Event that occurred in 2003.7315, in blue dashed row, and close-up of the pdf, $p(t)$, in thick red line; and the survival function of the forecast, $S(t)$, in dot- dash green line. c) Event that occurred in 2003.7315, in blue dashed row, and Hazard function $h(t)$.

After the occurrence of the aftcast event, Bayes formalism is used to test prior estimates or assumptions about possible semi-periodicity.

We considered three possible assumptions about the prior probability: 1) Confident: Semi-periodicity occurs and the probability of there being a non-random semi-periodic earthquake sequence in the study region is $\Pr(A) = P_c = 0.926$; 2) Undecided: Semi-periodicity may or may not occur, so that $\Pr(A) = 0.5$; 3) Skeptical: The probability of there being semi-periodicity is very small, say $\Pr(A) = 0.1$.

The results of the Bayesian evaluation are shown in Table 6. It is clear that the occurrence of the aftcast event so close to the forecast time supports the semi-periodicity thesis, since posterior probabilities are higher than prior ones, and probability gains are all greater than one; note that the gain is largest for the most skeptical approach.

Let us now compare the results of the Bayesian analysis that, from a prior probability of 0.926, resulted in a posterior probability of 0.979, with the updated semi-periodic probability for the four

events sequence $P_c^U = 0.993$. We see that Bayesian analysis gives a quite good approximation to the updated probability (it is smaller by only ~0.014). We note that coincidence between these two probabilities is even better for some other aftcasts not shown here.

Table 6. Results of aftcast of the event occurred in 2003.7315, in Japan, and the posterior comb probability and probability gains.

t_{nxt}	P_c	t_O	$\Pr(A)$ / $\Pr(A\|B)$ / P_G			P_c^U
			P_c	0.5	0.1	
			0.979	0.789	0.293	
2004.9822	0.926	2003.7315	1.058	1.578	2.933	0.993

6.2.2 Mexico

6.2.2.1 Study region and data

The study region is localized in the Pacific coast of southwest Mexico (Figure 27), south of the Colima graben and parallel to the subduction zone of the Cocos plate beneath the North American plate (e.g. Bandy et al., 1995, Demets and Wilson, 1997, Manea et al., 2013); within the region interplate velocities range from ~4.7 cm/yr in the north to ~6.4 cm/yr in the south. This region presents important seismic activity, and the earthquakes focal mechanisms are predominantly reverse near the coast and normal inland.

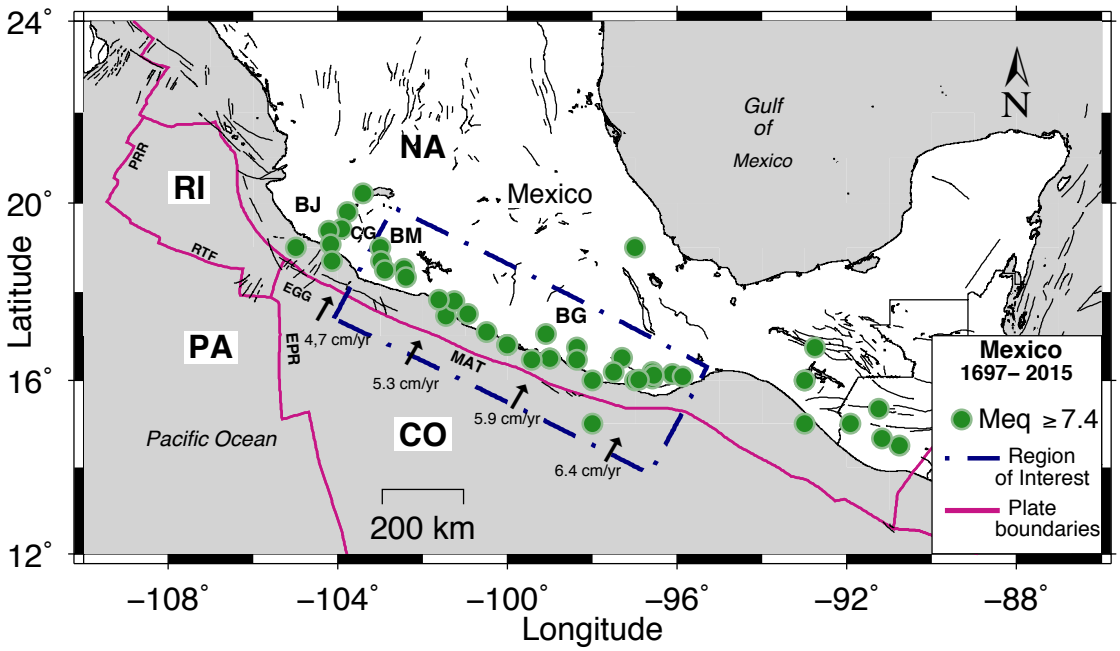

Figure 27. Map of MREs with $M_{eq} \geq 7.4$ MRE occurred in the study region of Mexico, from 1568 to 2015.5 (ISCS-GEM Catalogs). Green circles indicate epicenters of earthquakes. Dashed blue lines enclose the study region. Solid purple lines indicate the boundaries between the Rivera (RI), Cocos (CO), North America (NA), and Pacific (PA) tectonic plates (Bird, 2003).

The equivalent magnitudes, M_{eq}, of each MRE, were calculated. In the cases of the events occurred in 1985.7 and 2014.3, the respective M_{eq} was ~0.1 higher than the main earthquake magnitude. In other cases, M_{eq} is the same as that of the main earthquake of the respective MRE.

We have information about earthquakes that occurred in the study region since 1697 (Table 7), but, like in the Japan case, there are conspicuous gaps in the historical data, as can be seen in Figure 28a, which we interpret as missing data; so that the earthquakes included in the time series occurred from 1899 to 2015.5; the catalog may be considered complete from this date on for MREs with $M_{eq} \geq 7.4$ (Figure 28b).

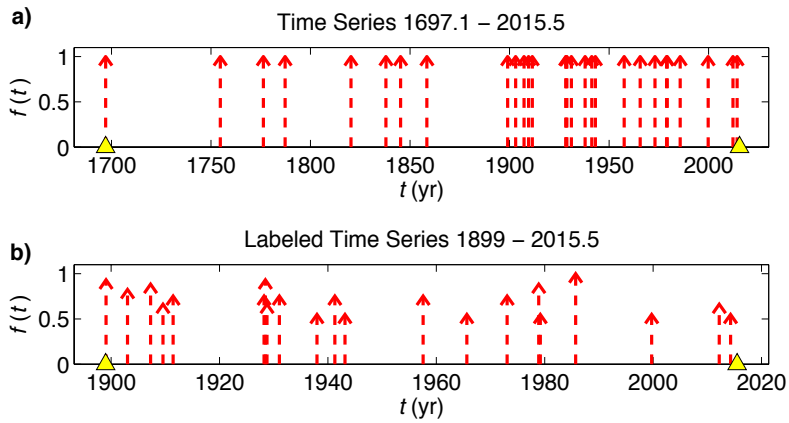

Figure 28. MREs occurrence time series, indicated by red arrows, for events with $M_{eq} \geq 7.4$ in the study area of Mexico. a) Occurrence time series of events from 1697 to 2015.5. b) Labeled occurrence time series of events from 1899 to 2015.5.

Magnitude weights, from 0.55 to 1.00 (Table 7), were assigned to events occurred from 1899 to 2015.5 with magnitudes form 7.4 to 8.1, using a *b*-value of 0.94. This *b*-value (eq. 3) was estimated from 189 data, occurred between 1950-2015, with $5.8 \leq M_w \leq 7.4$ for which the G-R histogram behaved linearly.

Table 7. MREs occurred in the study region of Mexico, with $M_{eq} \geq 7.4$, from 1697 to 2015.5 (ISC-GEM Catalogs).

No	Time (yr)	Latitude (°)	Longitude (°)	M_{eq}	Depth (Km)	Weight $w(M_{eq})$
1	1697.1014	16.50	-99.00	7.5		
2	1754.6658	16.80	-100.00	7.8		
3	1776.3033	16.80	-100.00	7.7		
4	1787.2356	16.00	-97.00	8.6		
5	1820.3388	16.50	-99.00	7.8		
6	1837.8932	16.00	-98.00	7.7		
7	1845.2630	16.80	-100.00	8.3		
8	1858.4630	19.00	-103.00	7.5		
9	1899.0657	17.10	-100.50	7.9		0.93
10	1903.0356	15.00	-98.00	7.7		0.83
11	1907.2849	16.51	-97.30	7.8	30.00	0.88
12	1909.5753	16.47	-99.43	7.5	20.00	0.66
13	1911.4301	18.52	-102.44	7.6	30.00	0.75
14	1928.2213	16.14	-96.11	7.6	15.00	0.75
15	1928.4590	16.18	-96.58	7.9	20.00	0.93
16	1928.7705	16.19	-97.50	7.5	25.00	0.66
17	1931.0384	16.04	-96.58	7.6	35.00	0.75
18	1937.9753	16.75	-98.36	7.4	25.00	0.55
19	1941.2849	18.69	-102.99	7.6	30.00	0.75
20	1943.1425	17.46	-101.45	7.4	20.00	0.55
21	1957.5698	17.05	-99.09	7.6	37.80	0.75
22	1965.6411	16.08	-95.87	7.4	25.00	0.55
23	1973.0794	18.49	-102.89	7.6	35.30	0.75
24	1978.9095	16.11	-96.55	7.8	20.00	0.88
25	1979.1973	17.79	-101.25	7.4	30.00	0.55
26	1985.7151	18.34	-102.39	8.1	20.00	1.00
27	1999.7452	16.01	-96.90	7.4	40.00	0.55
28	2012.2159	16.47	-98.37	7.5	19.40	0.66
29	2014.2932	17.50	-100.94	7.4	10.00	0.55

6.2.2.2 Analysis results

Labeled time series 1899-2015.5

In the analyzed time series there are $N = 21$ events (Figure 28b) that occurred over T=116.5 years, $[t_0 = 1899, t_f = 2015.5]$. The comb determined in pass one (Figure 29a) is constituted by $K = 9$ teeth with period $\tau_c = 14.3120$ yr ($s_c = 0.0699$ yr^{-1}). Event 10, 11, 12, 18, 22, 24 and 25 (all event numbers refer to Table 7) were dropped from the time series for the second pass of the analysis, because they did not satisfy the comb-event closeness criterion.

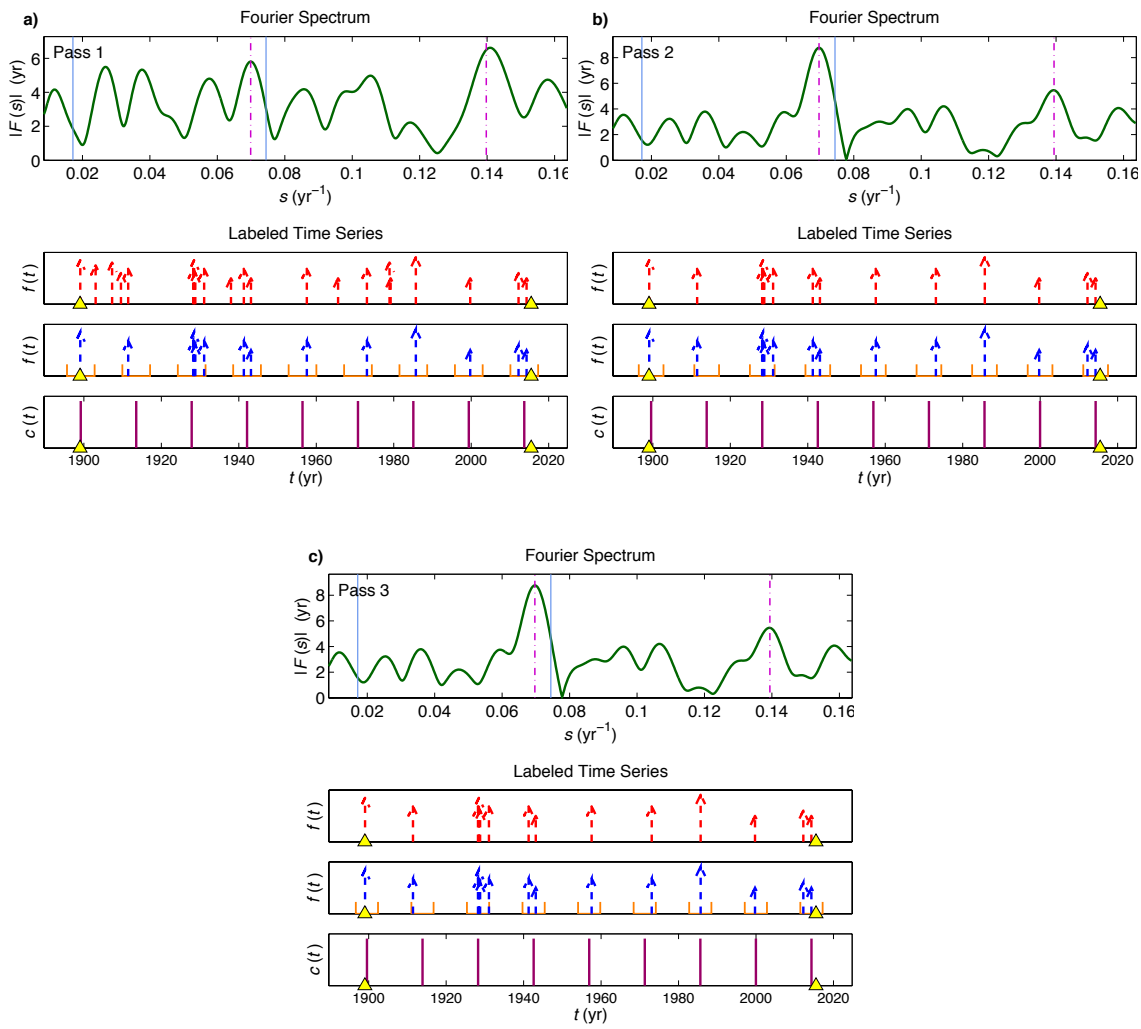

Figure 29. Spectral analysis of Mexico data from 1899 to 2015.5. a) First pass, b) Second pass, c) Third pass. Conventions are the same as in previous figures.

In pass two (Figure 29b), the period of the comb is $\tau_c = 14.3473$ yr ($s_c = 0.0697\,\text{yr}^{-1}$), and no event was dropped from the time series for the next pass.

In the third pass (Figure 29c) the comb period is the same as in the pass two. No event was dropped from the time series; so that the fourteen events closest to the comb: 9, 13, 14, 15, 16, 17, 19, 20, 21, 23, 26, 27, 28, and 29, became candidates to the nine event sequence. The fourth pass was applied to 16 combinations of nine events each, of which combination 4 had the lowest weighted error. Figure 30 and Table 8 show results for the best four combinations.

Table 8. The best four combinations analyzed in the pass 4. The shaded combination is the selected semi-periodic sequence.

Combination	Events (yr)		Weighted fitting error (yr)	Combination	Events (yr)		Weighted fitting error (yr)
	9	1899.0657			9	1899.0657	
	13	1911.4301			13	1911.4301	
	14	1928.2213			15	1928.4590	
	20	1943.1425			19	1941.2849	
4	21	1957.5699	7.475	6	21	1957.5699	7.996
	23	1973.0795			23	1973.0795	
	26	1985.7151			26	1985.7151	
	27	1999.7452			27	1999.7452	
	29	2014.2932			29	2014.2932	
	9	1899.0657			9	1899.0657	
	13	1911.4301			13	1911.4301	
	15	1928.4590			16	1928.7705	
	20	1943.1425			20	1943.1425	
8	21	1957.5699	7.688	12	21	1957.5699	7.988
	23	1973.0795			23	1973.0795	
	26	1985.7151			26	1985.7151	
	27	1999.7452			27	1999.7452	
	29	2014.2932			29	2014.2932	

The period of the comb from the spectral analysis of the selected combination is $\tau_c = 14.5262$ yr ($s_c = 0.0688\,\text{yr}^{-1}$), starting at $t_{c0} = 1898.8443$ yr ($\phi_c = 0.0674$). The measured standard deviation of θ is $\hat{\sigma} = 0.9814$ yr, and the estimated distribution standard deviation is $\sigma = 1.4421$ yr.

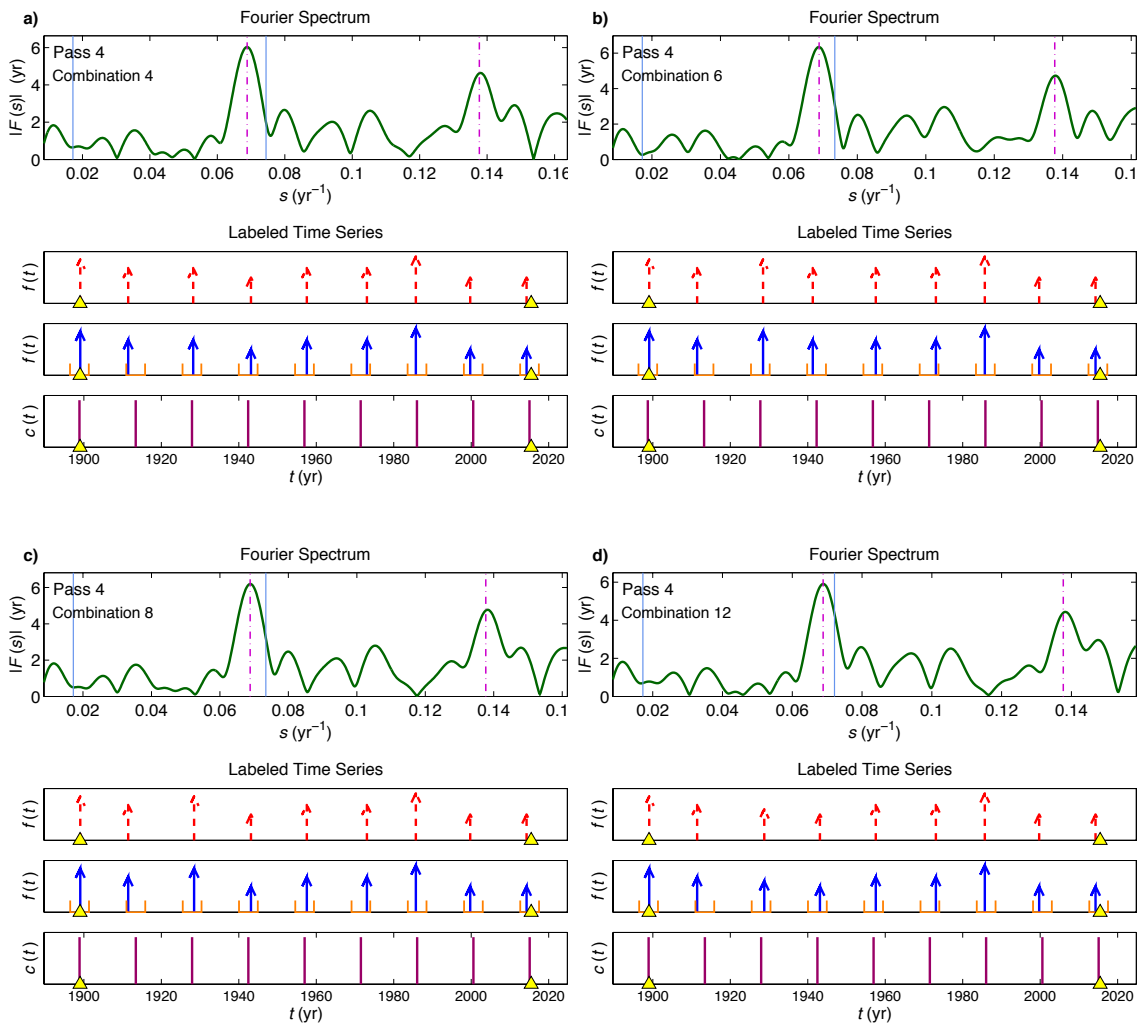

Figure 30. Spectral analyses of the best 4 combinations of the 16 made to identify the first sequence for Mexico data. The chosen semi-periodic sequence corresponds to combination 4, which has the lowest weighted fitting error (Table 8). Conventions are the same as in previous figures.

Thus, we would expect the next $M_{eq} \geq 7.4$ event belonging to the sequence around $t_{nxt} = 2029.5799 \pm 2.8842$ yr (with 95.45% confidence, $q = 2$). The probability of random occurrence of the sequence is $P_0 = 0.029$, so that the sequence/comb probability is $P_c = 0.971$.

The identified sequence and the pdf of the forecast, $p(t)$, are presented graphically in Figure 31, together with other probability functions that help to visualize some consequences of the forecast: the survival function and the hazard function of the forecast.

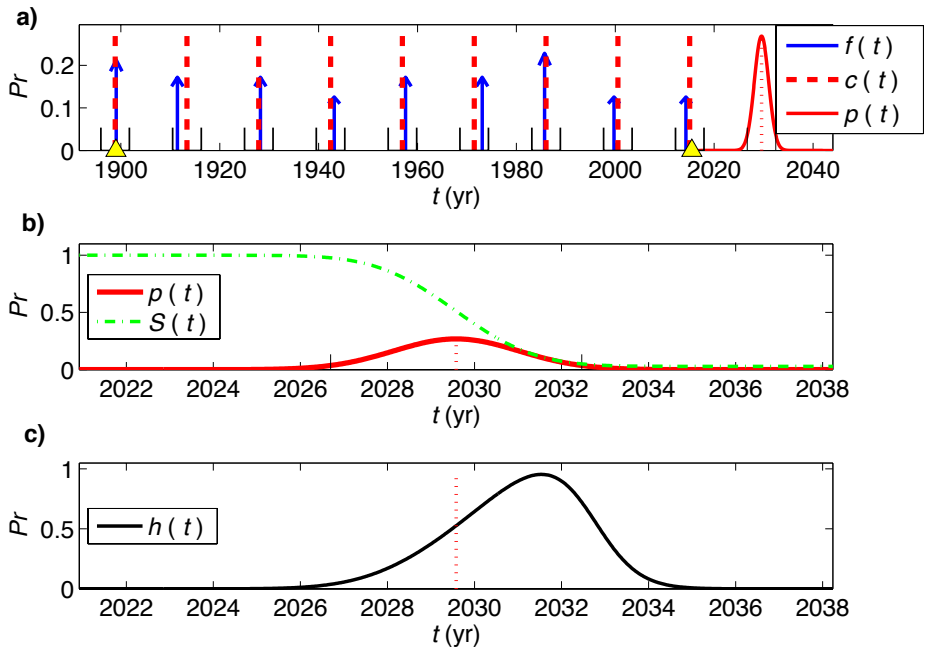

Figure 31. Mexico forecast. a) Forecast based on the sequence identification shown in Figure 30a; same convention as in Figure 20. b) Close-up of the pdf, $p(t)$, in thick red line; and the survival function of the forecast, $S(t)$, in dot- dash green line. c) Hazard function $h(t)$.

The occurrence probability of the next event belonging to the sequence within $\pm 2\sigma$, centered on time t_{nxt}, is $P_{cq} = 0.927$; the Poissonian occurrence probability of at least one event within a $\pm 2\sigma$ interval, considering the occurrence rate of the whole time series $\lambda = 0.180/\text{yr}$, is $\pi_{1+} = 0.647$; the Poissonian occurrence probability of at least one event within $\pm 2\sigma$, considering the occurrence rate of events not belonging to the sequence, $\lambda^* = 0.103/\text{yr}$, is $\pi^*_{1+} = 0.448$. The probability and information gains, evaluated from these probabilities, for different values of *q*, are significant (Table 9).

Table 9. Probabilities, probability gains and information gains for the Mexico sequence.

q	P_{cq}	π_{1+}	π^*_{1+}	P_G	I_G (bits)
1	0.6777	0.1608	0.1103	4.4349	2.1489
2	0.9475	0.2958	0.2085	3.2403	1.6961
3	0.9900	0.4090	0.2958	2.4275	1.2795

The epicenters of the events constituting the identified sequence are shown in Figure 32, together with all other large, $M_{eq} \geq 7.4$ MRE in the region.

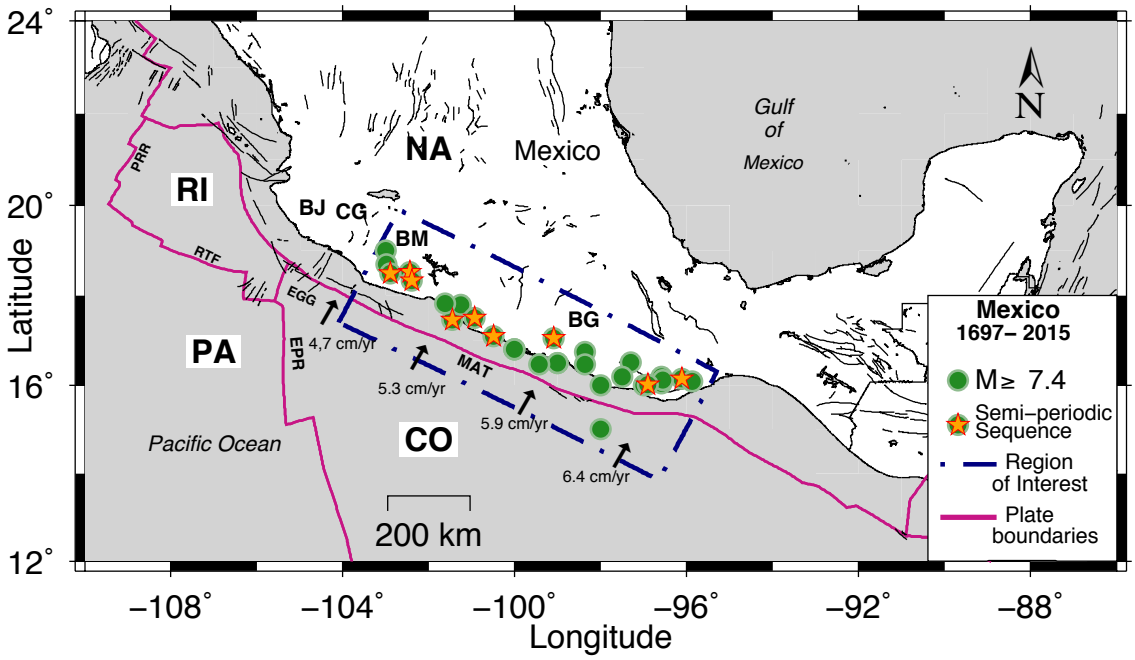

Figure 32. Map of MREs with $M_{eq} \geq 7.4$, occurred in the study region of Mexico from 1697 to 2015 (ISCS-GEM Catalogs). Orange stars indicate epicenters of the earthquakes belonged to the semi-periodic sequence, and green circles show epicenters of earthquakes not belonging to it. The image conventions are the same as in Figure 27.

Aftcasts and Bayesian analysis

The nine event sequence presented above was identified from events occurring from 1899 to 2015.5; we aftcast the last of the sequence events that occurred at 2014.2932, using only those events that occurred before it, between $t_0 = 1899$ yr and $t_f = 2000$ yr.

The time series is constituted by $N = 19$ events occurred over 101 years (Figure 33). Since the magnitude range of the events is between 7.4 and 8.1, the same as the analysis including earthquakes from 1899 to 2015.5, magnitude weights assigned to each event from 1899 to 2000 are the same as Table 7 shows.

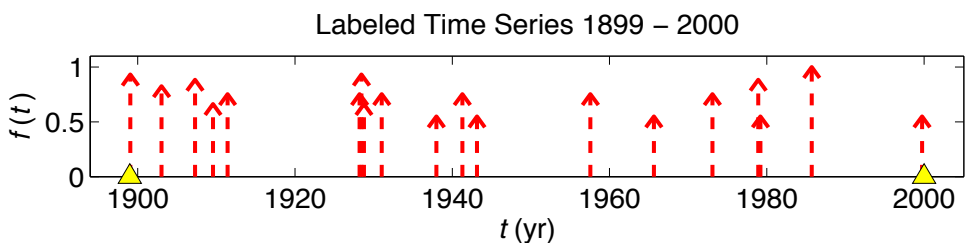

Figure 33. MREs occurrence labeled time series, indicated by red arrows, for events with $M_{eq} \geq 7.4$ in the study region of Mexico, from 1899 to 2000; the yellow triangles indicate t_0 and t_f.

Four passes were done (Figure 34) and the fourth pass was applied to eight combinations of eight events each one, of which Combination 2 had the lowest weighted error (Figure 34d). Table 10 shows results for the best four combinations.

The period of the comb obtained from the selected combination is $\tau_c = 14.5954$ yr ($s_c = 0.0685$ yr^{-1}), starting at $t_{c0} = 1898.6815$ yr ($\phi_c = 0.1371$). The measured standard deviation of θ is $\hat{\sigma} = 0.9950$ yr, and the estimated distribution standard deviation is $\sigma = 1.5066$ yr.

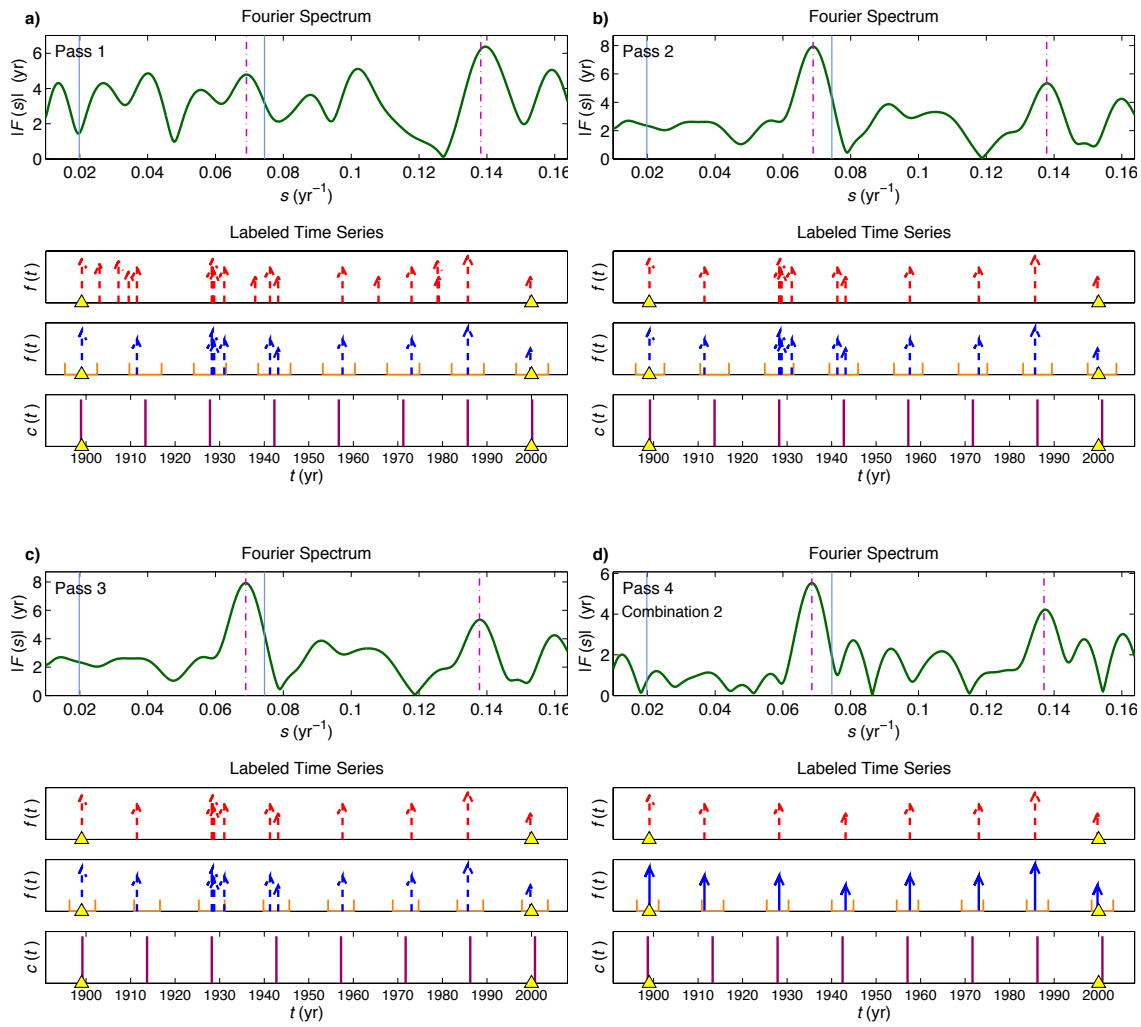

Figure 34. Spectral analysis of data from Mexico from 1899 to 2000. a) First pass, b) Second pass, c) Third pass, and d) Fourth pass of Combination 2 that is the selected sequence. Conventions are the same as in previous figures.

Thus, we would expect the next $M_{eq} \geq 7.4$ event belonging to the sequence around $t_{nxt} = 2015.4445 \pm 3.0132$ yr (with 95.45% confidence, $q = 2$). The probability of random occurrence of the sequence is $P_0 = 0.053$, so that the sequence/comb probability is $P_c = 0.947$.

The events belonging to the identified sequence are the same as the first eight events belonged to the identified sequence in the previous analysis (1896-2015.5), and the aftcast clearly coincides with the event occurrence in 2014.2932 yr. The aftcast error with respect to the actual occurrence time of the

event is 1.151 yr less than the time given in the confidence interval; so that, the aftcast is considered successful.

The identified sequence and the pdf of the forecast, $p(t)$, are presented graphically in Figure 35, together with the occurrence time of the event that occurred in 2014.2932, when the survival function begins to decline and well before the hazard function of the forecast reaches its maximum value.

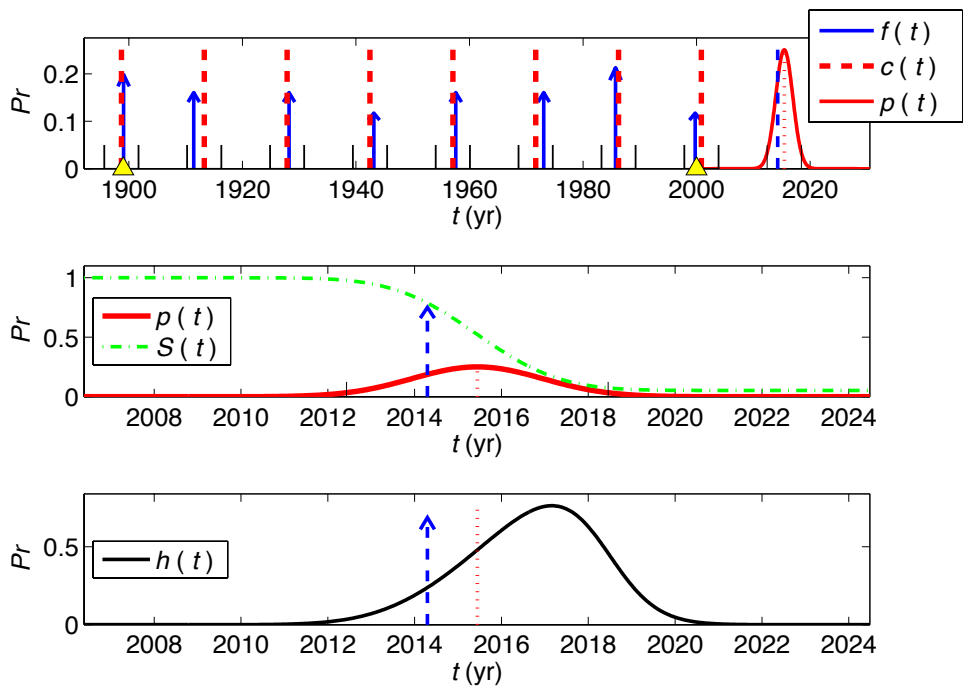

Figure 35. Mexico aftcast. a) Aftcast based on the sequence identification shown in Figure 34d; same conventions as in Figure 20. b) Close-up of the pdf, $p(t)$, in thick red line, and the survival function of the forecast, $S(t)$, in dot- dash green line. c) Hazard function $h(t)$.

The results of the Bayesian evaluation are shown in Table 10. The occurrence of the aftcast event so close to the forecast time, as in the Japan case, supports the semi-periodicity thesis, since posterior probabilities are higher than prior ones, and probability gains are all greater than one.

From a prior probability of $P_c = 0.947$, resulted in a posterior probability of $\Pr(A|B) = 0.979$, with the updated semi-periodic probability for the nine events sequence $P_c^U = 0.971$. We see that Bayesian analysis gives a quite good approximation to the updated probability (it is smaller by only ~0.008).

Table 10. Results of aftcast of the event occurred in 2014.2932, in Mexico, and the posterior comb probability and probability gains.

t_{nxt}	P_c	t_O	$\Pr(A)$ / $\Pr(A\|B)$ / P_G			P_c^U
			P_c	0.5	0.1	
			0.979	0.720	0.222	
2015.4445	0.947	2014.2932	1.033	1.440	2.221	0.971

6.3 Test of the effects of possible errors in rounded magnitudes on forecasts

The magnitudes referred in catalogs are usually rounded to one decimal place because of the uncertainties inherent in magnitude determinations, and in many cases magnitudes differ among catalogs and among sources, so that rounded magnitudes may be in error by more than ± 0.1. Thus, since the effects of possible errors in location and time are not significant, catalog magnitudes are the main source of uncertainty for the labeled point process semi-periodicity analysis, because due to these possible errors, there may be earthquakes erroneously excluded from, or included in, the analysis, and the magnitude weights may be wrong.

We studied effects of possible rounding error in magnitudes with a Monte Carlo approach, adding noise to the magnitudes listed in the catalogs, from ISCS-GEM, used for the analyses for Mexico and Japan. Normally distributed noise was added to the original catalog magnitudes, and the resulting magnitudes were rounded to obtain variations of the original catalogs.

$N_r = 100000$ realizations of modified catalogs were made to give a reliable statistical estimation, with noise $\mu_n = 0$ and $\sigma_n = 0.05$, rounded to one decimal place. This standard deviation is sufficiently large, because for both Japan and Mexico, all realizations resulted in catalogs that differed in some aspect from the original one; the number of modified catalogs $N_n = N_r$.

Many of the Moment release episodes (MRE) time series resulting from the *Nn* modified catalogs

were not equal to the original MRE time series; the number of MRE time series differing from the original one were $N_m = 97404$ for Japan and $N_m = 99994$ for Mexico

Forecasts based on the N_m time series were determined in order to test whether the original forecast was robust in the face of uncertainties in the magnitudes, by observing how many new forecasts differed significantly from it. We consider that a given forecast differs significantly from the original one, when the forecast times differ by more than 1/6 of the forecast period (the comb-event closeness criterion used in the last pass of the analysis).

For Mexico $N_0 = 71742$ realization resulted in forecasts very close to the original one, having mean forecast time $t_{nxtm} = 2029.4121$ yr, with $\sigma_t = 0.2884$ standard deviation, a mean that differs by only $\Delta t_{nxt} = 0.1678$ yr (~0.01 of the sequence period) from the original forecast $t_{nxt} = 2029.5799$ yr. Hence, we are assured that the original forecast was not an artifact of a particular combination of data, and that it is a robust estimation since the probability of obtaining essentially the same forecast in the face of possible errors in the magnitudes is $P_f = N_0 / N_m = 0.7571$

The robustness for the Japan original forecast is even better. $N_0 = 97404$ realizations resulted in forecasts very close to the original for all the MRE time series; so that the probability of obtaining practically the same forecast given magnitude noise in the MRE time series is $P_f = 1$. The mean of these forecasts is $t_{nxtm} = 2039.8700$ yr, and the standard deviation of the forecast values is $\sigma_t = 0.1222$. The mean of the forecasts differs only by $\Delta t_{nxt} = 0.0301$ yr (~0.0008 of the sequence period) from the original forecast $t_{nxt} = 2039.9001$ yr.

The above mentioned robustness estimates do not include the MRE time series that did not differ from the original ($N_n - N_m$), because they result in exactly the same forecast as the original one. If we include these cases, the probability of having practically the same forecast when there is magnitude noise in the catalog is $P_f = (N_0 + N_n - N_m)/N_n$, that is probability $P_f = 1$ for Japan and $P_f = 0.7175$ for Mexico, which are extremely good confidence estimates.

Chapter 7. Conclusions

7.1 About results

The semi-periodicity analysis method, modified to take into account the earthquake sizes, effectively identified semi-periodic sequences within labeled occurrence time series of large earthquakes, with results equal or better than the old version of the method. The weighting by magnitude introduced in the analysis, and also the criterion to evaluate the weighting fit error for all possible event combinations, were useful in selecting the best possible sequence. Also, the use of estimated population standard deviation of semi-periodicity, instead of the measured one, resulted in more realistic occurrence time forecasts.

We consider that sequences with less than four events and (usually) low non-randomness probability, are not reliable. Hence, although more than one semi-periodic sequence was identified in most study regions, only sequences with more than three events and high non-randomness probability were considered as valid forecasts, and worth showing.

The identified semi-periodic sequence, in the Japan case, is constituted by four events and the forecast is $t_{nxt} = 2039.9001 \pm 1.5292$ yr with $P_c = 0.992$. The semi-periodic sequence of Mexico is constituted by nine events and the forecast is $t_{nxt} = 2029.5799 \pm 2.8842$ yr with $P_c = 0.971$.

The most notable case to illustrate the influence of the weights in the analysis, was the Japan case, in which the event occurred in 2011.1890 ($M_{eq} = 9.1$) was the last event of the sequence identified in the unlabeled analysis; but in the labeled analysis, the last event of the identified sequence was the one occurred in 2003.73151 ($M_{eq} = 8.3$), and this choice resulted in a much better forecast. Aftcasts from both labeled and unlabeled analyses were closer to the 2003.73151 event than to the 2011.1890 one; this bolsters the assumption that this last event, which has a magnitude considerably higher than the others belonging to the sequence, could really belong to another sequence with much longer period.

As in the Japan case, aftcasts done for Mexico, using the new method, were successful and the posterior probabilities of the sequence, obtained through the Bayesian analysis, were high and very close to the new P_c^U.

The tests of the effects of possible rounding error in magnitudes showed the robustness of our forecasts, because in spite of noise in the magnitudes, that causes different events to be included in or excluded from the analyses, or causes the same events to have different magnitudes and weights that result in different spectra, most of the forecasts are very close to the corresponding noiseless ones.

7.2 About the method

The method is a mathematical/computational embodiment of the assumptions stated at the beginning of this work; the main assumption is, of course, the existence of physical processes that produce semi-periodic sequences of strong earthquakes. Experience shows that the method does find semi-periodic sequences, supporting our main assumption, and the very high non-randomness probabilities of the identified sequences further upholds it.

The method still includes some arbitrary aspects and parameters: the way in which study regions are chosen, the parameters used for event association into MREs; the actual way in which magnitude weights for the FT are assigned, the closeness criteria for discarding non-comb events at each pass, the way of weighting for magnitude spread. These arbitrary aspects and values have been set to result in the best and most reliable results.

The method is limited in its possible applications by the length and quality of seismic catalogs, which for many regions are not long or homogeneous enough.

As before, the main limitation to forecasting using this method is that there is no assurance that the identified sequences in fact correspond to physical semi-periodic processes that will continue to produce large earthquakes with the same sequence periods. This is a limitation common to all statistical methods: if the observed processes are changing in some unpredictable manner, all forecasts based on their history will be useless.

An important factor to consider is that the forecast of an earthquake indicated by some identified sequence, does not imply at all that other earthquakes, not related to the given sequence, may not occur at any time before, during, or after the forecast time.

7.3 Future work

We will continue to work on improvements to our method for identifying semi-periodic sequences of large earthquakes, and the method will be applied to other interesting regions. If and when a reliable short-term forecast is obtained; it will be presented to the international scientific community for judging, and, if it is judged reliable enough, it will be communicated to the appropriate authorities.

References

Bak, P., Tang, C., and Wiesenfeld, K. 1988. Self-organized criticality. *Phys. Rev. A* 38(1), 364–374.

Bak, P., Christensen, K., and Olami, Z. 1994. Self-Organized Criticality: Consequences for Statistics and Predictability of Earthquakes. *Nonlinear Dynamics and Predictability of Geophysical Phenomena*, 69-74.

Bak, P. 1996. How Nature Works*:* The Science of Self-Organized Criticality. Copernicus, New York .

Bakun, W. H., and Lindh A. G. 1985. The Parkfield, California, earth- quake prediction experiment, *Science*, 229, 619–624.

Bandy W.L., Mortera-Gutierrez C., Urrutia-Fucugauchi J., Hilde T.W.C. 1995. The subducted Rivera-Cocos plate boundary: Where is it, what is it, and what is its relationship to the Colima rift?. *Geophys. Res. Lett*., 22, 3075-3078.

Bird, P. 2003. An updated digital model of plate boundaries. *Geochem. Geophys. Geosyst.* 4(3), 1027.

Cox, A. 1973. Plate Tectonics and Geomagnetic Reversals. W. H. Freeman, San Francisco.

Demets, C. and Wilson, D.S., 1997. Relative motions of the Pacific, Rivera, North American, and Cocos plates since 0.78 Ma. *J. Geophys. Res*. 102, 2789–2806.

Fano, R. 1961. Transmissions of Information, A statistical Theory of Communications. John Wiley and Sons, Inc., New York, and M.I.T. Press, Cambridge, Mass.

Hanks, T. and Kanamori, H. 1979. A moment magnitude scale, *J. Geophys. Res*. 84, No. B5, 2348-2350.

Harte, D., and Vere-Jones, D. 2005. The entropy score and its uses in earthquake forecasting*. Pure Appl. Geophys*. 162, 1229–1253.

Ito, K., and Matsuzaki, M. 1990. Earthquakes as self-organized critical phenomena, *J. Geophys. Res*., 95, 6853–6868.

Manea, V. C., Manea M. and Ferrari L. 2013. A geodynamical perspective on the subduction of Cocos and Rivera plates beneath Mexico and Central America. *Tectonophysics*. *609*, 56–81.

Morgan, W. J. 1968. Rises, Trenches, Great Faults, and Crustal Blocks. *J. Geophys. Res*., 1959-1982.

Nakamura, K. 1983. Possible nascent trench along the eastern Japan Sea as the convergent boundary between Eurasian and North American plates. *Bull. Earthq. Res. Inst*., Univ. Tokyo 58, 711–22.

Nava, F. A., Quinteros, C. B., Glowacka, E., and Frez, J. 2014. Semi-periodic sequences and extraneous events in earthquake forecasting: I. Theory and method, Parkfield. *Pure Appl. Geophys*. 171(7), 1355–1366.

Nava, F. A., Quinteros, C. B., Glowacka, E., and Frez, J. 2016. A bayesian assessment of seismic semi-periodicity forecasts. *Pure Appl. Geophys.* 173(1), 197–203.

Nishenko, S. P., and Buland, R. 1987. A Generic Recurrence Interval Distribution for Earthquake Forecasting, *Bull. Seismol. Soc. Am*. 77, 1382–1399.

Parzen, E. 1960. Modern Probability Theory and Its Applications, John Wiley and Sons, New York.

Quinteros, C. B. 2012. Estudio de semiperiodicidad para grandes sismos en Venezuela y Japón. Master Thesis. Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California. 138 pp.

Quinteros, C. B., and Nava, F. A. 2014. Postnóstico (pronóstico hecho a posteriori) del sismo del 11 de octubre de 2013 en Venezuela, mediante análisis de semiperiodicidad. *GEOS,* 33 (2), 350-355.

Quinteros, C. B., Nava, F. A., Glowacka, E., and Frez, J. 2014. Semi-periodic sequences and extraneous events in earthquake forecasting: II Application, forecasts for Japan and Venezuela. *Pure Appl. Geophys*. 171(7), 1367–1383.

Reid, H. 1910. The mechanics of the earthquake, the California earthquake of April 18, 1906. Report of the State Investigation Commission, Vol. 2. Carnegie Institution of Washington, Washington, D.C.

Richardson, R. M., Solomon, S. C., and Sleep, N. H. 1979. Tectonic stress in the plates. *Reviews of Geophysics and Space Physics*, 17(5), 981-1019.

Scholz, C.H. 1990. The Mechanics of Earthquakes and Faulting. Cambridge Univ. Press, New York. 439 pp.

Shannon, C. 1948. A mathematical theory of communication. The Bell System Technical Journal 27, 379-423, 623-656.

Taira, A. 2001. Tectonic evolution of the Japanese Island Arc System Tamaki. Annu. *Rev. Earth Planet*. Sci. 29, 109–34.

Tamaki, K., and Honza, E. 1985. Incipient subduction and obduction along the eastern margin of the Japan Sea. *Tectonophysics* 119, 381–406.

The Global Historical Earthquake Catalogue. Retrieved in January 2015: http://www.globalquakemodel.org/what/seismic-hazard/historical-catalogue/

The ISC-GEM Global Instrumental Earthquake Catalogue. Version 2.0. Retrieved in January 2015: http://www.isc.ac.uk/iscgem/

Utsu, T. 1965. A method for determining the value of b in formula log n = a - bM showing the magnitude-frequency relation for earthquakes. *Geophys. Bull*. Hokkaido Univ. 13, 99-103.

Vere-Jones, D. 1998. Probabilities and information gain for earthquake forecasting. Comput. Seismol. 30, 248–263.

Wells, D. L., and Coppersmith K. J. 1994. New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bull. Seism. Soc. Am.* 84, 974-1002.

Winkler, Robert L. 2003. Introduction to Bayesian Inference and Decision, 2nd Edition Probabilistic. ISBN 0-9647938-4-9