Modification of the pattern informatics method for forecasting large earthquake events using complex eigenfactors

J.R. Holliday [a,b,*], J.B. Rundle [a,b], K.F. Tiampo [c], W. Klein [d], A. Donnellan [e]

[a] *Center for Computational Science and Engineering, University of California, Davis, California, USA*
[b] *Department of Physics, University of California, Davis, California, USA*
[c] *Department of Earth Sciences, University of Western Ontario, London, Ontario, Canada*
[d] *Department of Physics, Boston University, Boston, Massachusetts, USA*
[e] *Earth and Space Sciences Division, Jet Propulsion Laboratory, Pasadena, California, USA*

Abstract

Recent studies have shown that real-valued principal component analysis can be applied to earthquake fault systems for forecasting and prediction. In addition, theoretical analysis indicates that earthquake stresses may obey a wave-like equation, having solutions with inverse frequencies for a given fault similar to those that characterize the time intervals between the largest events on the fault. It is therefore desirable to apply complex principal component analysis to develop earthquake forecast algorithms. In this paper we modify the Pattern Informatics method of earthquake forecasting to take advantage of the wave-like properties of seismic stresses and utilize the Hilbert transform to create complex eigenvectors out of measured time series. We show that Pattern Informatics analyses using complex eigenvectors create short-term forecast hot-spot maps that differ from hot-spot maps created using only real-valued data and suggest methods of analyzing the differences and calculating the information gain.
© 2005 Elsevier B.V. All rights reserved.

1. Introduction

Principal component analysis (PCA) is a mathematical procedure that transforms a set of correlated variables into a smaller set of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component attempts to account for the remaining variability. Savage (1988) introduced PCA to the seismic community by using it to decompose time series data into a complete set of orthonormal subspaces that isolate spatial and temporal eigensources.

Complex principal component analysis is an extension of classical principal component analysis in which the spatial basis vectors represent the eigenfunctions of a complex correlation matrix. It is closely related to principal oscillation pattern (POP) analysis, in which the oscillating basis pattern states are the eigenfunctions of a deterministic feedback matrix (Penland, 1989) (both techniques empirically identify time-dependent spatial patterns in a multivariate time series of geophys-

---

* Corresponding author. Center for Computational Science and Engineering, University of California, Davis, California, USA.

E-mail addresses: holliday@cse.ucdavis.edu (J.R. Holliday), jbrundle@ucdavis.edu (J.B. Rundle), ktiampo@uwo.ca (K.F. Tiampo), klein@buphyc.bu.edu (W. Klein), donnellan@jpl.nasa.gov (A. Donnellan).

ical or other data). POP analysis has been shown to be reasonably successful in forecasting El Niño-Southern Oscillation (ENSO) events up to a year in advance (Wu et al., 1994). In complex PCA, a real-valued time series is analytically continued into the complex-valued domain by means of a Hilbert transform (Horel, 1984), then the $N \times N$ complex correlation matrix is formed via cross-correlation of the N independent time series. These methods have been applied extensively in the atmospheric and ocean sciences (Penland, 1989; Burger, 1993; Zhang et al., 1997; Egger, 1999; Kim and North, 1999).

The primary benefit of complex PCA compared to other analysis procedures is that it allows propagating features within the time series to be detected and dissected in terms of their spatial and temporal behavior (Horel, 1984). In particular, localized propagating phenomena, if they exist, can be easily detected. Classical PCA, for example, allows only the detection of standing oscillations.

Recently it has been shown that the same kinds of real-valued PCA analysis can be applied to earthquake fault systems for forecasting and prediction (Rundle et al., 2000b; Tiampo et al., 2002a,b). It is known that earthquakes recur in complex cycles, similar to ENSO events, albeit with the larger earthquake events having substantially longer time scales (Scholz, 1990) than those that apply to ENSO—typically a decade or less. In addition, theoretical analysis (Klein, 2004) indicates that earthquake stresses may obey a wave-like equation, having solutions with inverse frequencies for a given fault similar to those that characterize the time intervals between the largest events on the fault. It is of considerable interest to apply complex PCA and POP analysis to develop earthquake forecast algorithms, taking account of the complex cyclic and quasi-periodic nature of these events.

A problem with this approach is that earthquake event time series are typically not continuous and differentiable, but instead are point processes, both in space and in time. In addition, high quality measurements of earthquakes have only been comprehensively observed with instruments for a few decades, so the complete (high-density) time series that are available are relatively short compared to the recurrence periods for large earthquakes of hundreds of years and longer. The Pattern Informatics (PI) method for earthquake forecasting is well suited for these types of impulsive time series and performs very well with data sets much shorter than the recurrence periods for large earthquakes events (Holliday et al., 2005). As such, it is an ideal candidate for modification to use complex

eigenfunctions and eigenvectors. Assuming that seismic phenomena are analytic, causality considerations allow us to apply the Cauchy Riemann dispersion relations (Arfken and Weber, 2001) and analytically continue the measured time series from the real axis into the entire upper half-plane of complex space. In this new space we propose to utilize the PI method.

2. Modified method

Our modified PI method is based on the idea that the future time evolution of seismicity can be described by pure phase dynamics (Mori and Kuramoto, 1998; Rundle et al., 2000a,b), hence a complex seismic phase function $\hat{S}(\mathbf{x}_i, t_b, t)$ is constructed and allowed to rotate in its Hilbert space. This modified representation of the input data serves two purposes. First, a complex Hilbert space allows detection both of standing oscillations and traveling waves (Horel, 1984). This is important for identifying the quasi-periodic nature of seismicity. Second, the construction allows for interference between the real and imaginary parts of the phase function. This interference helps correlate geographic locations that are spatially separated.

To create our phase function, the geographic area of interest is partitioned into N square bins centered on a point $\mathbf{x}_i$ and with an edge length δx determined by the nature of the physical system. For our analysis we chose $\delta x = 0.1 \approx 11$ km, corresponding to the linear size of a magnitude $M \sim 6$ earthquake. Within each box, a time series $\psi_{obs}(\mathbf{x}_i, t)$ is defined by counting how many earthquakes with magnitude greater than M_{min} occurred during the time period t to $t + \delta t$. These time series are interpreted as the real-valued portion of an analytic signal, and thus the entire signal is recreated by combining ψ_{obs} with its Hilbert transform:

$$\psi_{obs}(\mathbf{x}_i, t) \rightarrow \Psi(\mathbf{x}_i, t) \equiv \psi_{obs} + \tilde{\psi}_{obs}, \quad (1)$$

where $\tilde{\psi}_{obs}(\mathbf{x}_i, t) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{\psi(\mathbf{x}_i, \tau) d\tau}{t - \tau}$ and Cauchy principal value integration is specified (Bracewell, 1999). Next, the activity rate function $S(\mathbf{x}_i, t_b, T)$ is defined as the average rate of occurrence of earthquakes in box i over the period t_b to T:

$$S(\mathbf{x}_i, t_b, T) = \frac{\sum_{t=t_b}^{T} \Psi(\mathbf{x}_i, t)}{T - t_b}. \quad (2)$$

If t_b is held to be a fixed time, $S(\mathbf{x}_i, t_b, T)$ can be interpreted as the ith component of a general, time-dependent vector evolving in an N-dimensional space

(Tiampo et al., 2002b). Furthermore, it can be shown that this N-dimensional correlation space is defined by the eigenvectors of an $N \times N$ correlation matrix (Rundle et al., 2000a,b). In order to remove the final free parameter in the system—the choice of base year— changes in the activity rate function are then averaged over all possible base-time periods:

$$\underline{S}(\mathbf{x}_i, t_0, T) = \frac{\sum_{t_b=t_0}^{T} S(\mathbf{x}_i, t_b, T)}{T - t_0}. \qquad (3)$$

The base averaged activity rate function is then normalized by subtracting the spatial mean over all boxes and scaling to give a unit-norm:

$$\hat{\underline{S}}(\mathbf{x}_i, t_0, T)$$

$$= \frac{\underline{S}(\mathbf{x}_i, t_0, T) - \frac{1}{N} \sum_{j=1}^{N} \underline{S}(\mathbf{x}_j, t_0, T)}{\sqrt{\sum_{j=1}^{N} \left[\underline{S}(\mathbf{x}_j, t_0, T) - \frac{1}{N} \sum_{k=1}^{N} \underline{S}(\mathbf{x}_k, t_0, T) \right]^2}}. $$
$$(4)$$

The requirement that the rate functions have a constant norm helps remove random fluctuations from the system. Following the assumption of pure phase dynamics (Rundle et al., 2000a,b), the important changes in seismicity will be given by the change in the normalized base averaged activity rate function from the time period t_1 to t_2:

$$\Delta \hat{\underline{S}}(\mathbf{x}_i, t_0, t_1, t_2) = \hat{\underline{S}}(\mathbf{x}_i, t_0, t_2) - \hat{\underline{S}}(\mathbf{x}_i, t_0, t_1). \qquad (5)$$

This is simply a pure rotation of the N-dimensional unit vector $\hat{\underline{S}}(\mathbf{x}_i, t_0, T)$ through time. Finally, the probability of change of activity in a given box is deduced from the square of its base averaged, mean normalized change in activity rate:

$$P(\mathbf{x}_i, t_0, t_1, t_2) = \left[\Delta \hat{\underline{S}}(\mathbf{x}_i, t_0, t_1, t_2) \right]^*$$
$$\times \left[\Delta \hat{\underline{S}}(\mathbf{x}_i, t_0, t_1, t_2) \right], \qquad (6)$$

where multiplication and complex conjugation are indicated. In phase dynamical systems, probabilities are always related to the square of the associated vector phase function (Mori and Kuramoto, 1998; Rundle et al., 2000b). This probability function is often given relative to the background by subtracting off its spatial mean:

$$P'(\mathbf{x}_i, t_0, t_1, t_2) = P(\mathbf{x}_i, t_0, t_1, t_2) - \mu, \qquad (7)$$

where $\mu = \frac{1}{N} \sum_{j=1}^{N} P(\mathbf{x}_j, t_0, t_1, t_2)$ and P' indicates the probability of change in activity is measured relative to the background.

3. Application of the method

As an application of the modified PI method, we created a short-term forecast seismic hot-spot map for Southern California over the time period 1 August 2004 to 31 July 2009. The result is shown in Fig. 1A. Also presented in Fig. 1 is the same forecast map created with real-valued eigenvectors (Fig. 1B) and a difference map between the two methods (Fig. 1C).

Two data sets were employed in this analysis, the first being the entire historic seismic catalog from 1 January 1932 through 31 July 2004, obtained from the Southern California Earthquake Data Center (SCEDC) on-line searchable database,[1] with all non-local and blast events specifically removed. The relevant data consists of location, in East longitude and North latitude, and the date the event occurred. Seismic events between $-122°$ and $-155°$ longitude and between $32°$ and $35°$ latitude (any depth and quality) and with magnitude greater than or equal to $M_{min} = 3.0$ were selected. Data from the time period 1977–1980 is currently missing from the database but can be found at the older Southern California Seismic Network (SCSN) archives.[2]

The second source of data employed in this analysis was acquired from the Northern California Earthquake Data Center (NCEDC) on-line searchable database,[3] with all non-local and blast events again specifically removed. When incorporating this catalog, seismic events between $-122°$ and $-155°$ longitude and between $35°$ and $37°$ latitude (any depth and quality) and with magnitude greater than or equal to $M_{min} = 3.0$ were selected. The necessity of combining catalogs arises from the fact that the SCEDC catalog is not complete in its network coverage above the joining mark. Most notably, it does not contain earthquakes from the San Simeon region (location of the $M = 6.5$, 2003 event).

As can be seen in Fig. 1, the map created using complex eigenvectors is similar to the map created using real-valued eigenvectors. Important differences, however, are present. Most prominent are the increased emphasis of forecasted activity surrounding $36°$ North latitude, $-117.9°$ East longitude and the decreased

---

[1] http://www.data.scec.org/catalog_search/index.html.
[2] http://www.data.scec.org/ftp/catalogs/SCSN/.
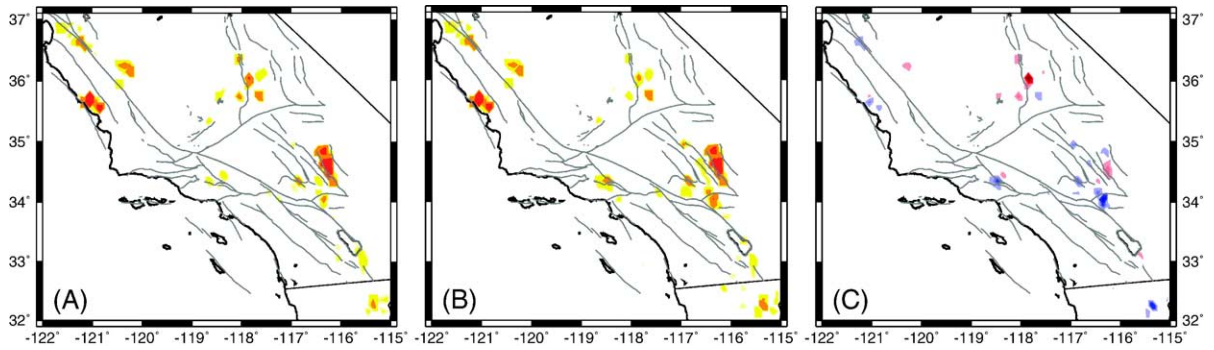[3] http://quake.geo.berkeley.edu/ncedc/catalog-search.html.

Fig. 1. Logarithmic seismic hot-spot map for large earthquake events with $M \geq 5$ for the forecasted time period 1 August 2004 to 31 July 2009 using (A) complex eigenvectors and (B) real-valued eigenvectors. Data from the SCEDC catalog was used below 35° North latitude while data from the NCEDC catalog was used above 35° North latitude. (C) Difference map plotted with a linear scale.

emphasis of forecasted activity southwest of the 1999 Hector Mine events. While future monitoring of these areas will be necessary to help determine the accuracy and reliability of complex PI analysis, certain measurements can be performed to estimate the information gain.

3.1. Entropy

Using methods from information theory (Cover and Thomas, 1991), we can calculate the entropy, H, of our two hot-spot maps. Entropy can be considered a measure of disorder (e.g., randomness) or "surprise", hence maps with lower entropy contain more useful information than maps with higher entropy. We define entropy as

$$H(z) = - \sum_{i=1}^{N} p(\mathbf{x}_i; z) \log\ p(\mathbf{x}_i; z), \qquad (8)$$

where

$$p(\mathbf{x}_i; z) = \begin{cases} P(\mathbf{x}_i, t_0, t_1, t_2) & P(\mathbf{x}_i, t_0, t_1, t_2) \geq z \\ 0 & P(\mathbf{x}_i, t_0, t_1, t_2) < z \end{cases}, \qquad (9)$$

and the probabilities are scaled such that $\sum_{i=1}^{N} p(\mathbf{x}_i) = 1$. This definitions allows a measurement of entropy as a function of some cut-off threshold.

Performing this calculation on the two maps indicates that the complex PI analysis does indeed yield more useful information (lower H-value) than the original analysis, but only when the threshold value is non-zero. With complex PCA calculations, sudden transitions and noisy spikes are emphasized (Horel, 1984). Since seismic time series data can be approximated by chains of delta functions, we expect that calculations in the complex domain would contain more low-level

noise. A small, non-zero threshold allows us to measure the entropy above and relative to this noise.

3.2. ROC analysis

A second measure for the accuracy of the hot-spot maps can be inferred from relative operating characteristic (ROC) diagrams. ROC curves are essentially signal detection curves for binary forecasts obtained by plotting the hit rate (y-axis) against the false alarm rate (x-axis) over a range of different thresholds (Joliffee and Stephenson, 2003). Originally adopted for verifying tornado forecasts (Murphy and Winkler, 1987), ROC frameworks have recently become popular in the seismic community as well (Molchan, 1997).

While only one year has passed since the onset of the hot-spot forecasts given in Fig. 1, we can create ROC diagrams for the two maps by considering a "hit" to be any box i with $P(\mathbf{x}_i, t_0, t_1, t_2) \geq z$, for some threshold z, that contains a future large earthquake. Similarly we consider a "false alarm" to be any box j with $P(\mathbf{x}_j, t_0, t_1, t_2) \geq z$, for some threshold z, that does not contain a future large earthquake. Since a successful forecast will maximize the hit rate while minimizing the false alarm rate, a measure of the forecast accuracy is given by the area, A_{ROC} under the ROC curve. It can be shown that $A_{ROC} \rightarrow 1$ for a perfect forecast and $A_{ROC} \rightarrow 0.5$ for a forecast consisting of randomly distributed alarms.

Performing this calculation on the two maps again indicates that the complex PI analysis is better correlated with future large events (higher A_{ROC}-value) than the original analysis. It is important to consider, however, that this analysis is only using one year of observed future seismicity. A full analysis should be performed at the end of the forecast interval.

4. Conclusion

Complex PCA is a useful tool and is ideally suited for many applications. There are, however, situations where the results of complex PCA are difficult to interpret such as when both amplitude and phase relationships must be considered. For these types of systems, the existence of phase information by itself suggests the need for an analysis in the full complex domain.

The theoretical evidence that earthquake stress fields are wave-like in nature indicates that seismicity is better studied using complex time series. Due to its ability to create seismic hot-spot forecast maps using relatively short time series data and its handling of impulsive data sets, the PI method is naturally extended to this complex domain.

In our five year seismic hot-spot forecast for southern California, the map created using complex eigenvectors has subtle differences with the map created using real-valued eigenvectors. These differences result in more useful information (i.e., a reduction in the map entropy) and in better apparent correlation with future large earthquakes. Future monitoring and testing, however, will be necessary to conclusively determine the accuracy and reliability of complex PI analysis.

Acknowledgments

References

Arfken, G.B., Weber, H.J., 2001. Mathematical Methods for Physicists, 5th edition. Academic Press.

Bracewell, R.N., 1999. The Fourier Transform and its Applications. McGraw-Hill, New York.

Burger, G., 1993. Complex principal oscillation pattern analysis. J. Climate 6 (10), 1972–1986 (Oct.).

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. Wiley-Interscience, New York.

Egger, J., 1999. POPs and MOPs principal and main oscillation patterns. Clim. Dyn. 15 (7), 561–568 (July).

Holliday, J.R., Rundle, J.B., Tiampo, K.F., Klein, W., Donnellan, A., in print. Systematic Procedural and Sensitivity Analysis of the Pattern Informatics Method for Forecasting Large ($m \geq 5$) Earthquake Events in Southern California.

Horel, J.D., 1984. Complex principal component analysis: theory and examples. J. Appl. Meteorol. 23, 1660–1673.

Joliffee, I.T., Stephenson, D.B., 2003. Forecast Verification. John Wiley.

Kim, K., North, G.R., 1999. EOF-based linear prediction algorithm: examples. J. Climate 12 (7), 2076–2092 (July).

Klein, W., 2004. Stress field evolution near the spinodal, unpublished.

Molchan, G.M., 1997. Earthquake predictions as a decision-making problem. Pure Appl. Geophys. 149, 233–247.

Mori, H., Kuramoto, Y., 1998. Dissipative Structures and Chaos. Springer-Verlag, Berlin.

Murphy, A.H., Winkler, R.L., 1987. A general framework for forecast verification. Mon. Weather Rev. 115, 1330–1338.

Penland, C., 1989. Random forcing and forecasting using principal oscillation pattern analysis. Mon. Weather Rev. 117, 2165–2185.

Rundle, J.B., Klein, W., Gross, S.J., Tiampo, K.F., 2000a. Dynamics of seismicity patterns in systems of earthquake faults. In: Rundle, J.B., Turcotte, D.L., Klein, W. (Eds.), Geocomplexity and the Physics of Earthquakes. Vol. of 120 of Geophys. Monogr. Ser. AGU, Washington, D.C., pp. 127–146.

Rundle, J.B., Klein, W., Tiampo, K.F., Gross, S.J., 2000b. Linear pattern dynamics in nonlinear threshold systems. Phys. Rev., E 61, 2418–2432.

Savage, J.C., 1988. Principal component analysis of geodetically measured deformation in Long Valley caldera, eastern California, 1983–1987. J. Geophys. Res. 93, 13297–13305.

Scholz, C.H., 1990. Geophysics—earthquakes as chaos. Nature 348, 197–198.

Tiampo, K.F., Rundle, J.B., McGinnis, S., Gross, S.J., Klein, W., 2002a. Eigenpatterns in southern California seismicity. J. Geophys. Res. 107 (B12), 2354.

Tiampo, K.F., Rundle, J.B., McGinnis, S., Klein, W., 2002b. Pattern dynamics and forecast methods in seismically active regions. Pure Appl. Geophys. 159, 2429–2467.

Wu, D.H., Anderson, D.L.T., Davey, M.K., 1994. ENSO prediction experiments using a simple ocean–atmosphere model. Tellus Ser. A—Dynamic Meteorology and Oceanography, vol. 46, pp. 465–480.

Zhang, Y., Dymnikov, V., Wallace, J.M., 1997. Sensitiviy test of POP system matrices—an application of spectral portrait of a nonsymmetric matrix. J. Climate 10 (7), 1753–1758 (July).