Report on 2010 SCEC Grant
Data adapted algorithms for automatic transient detection in GPS time series

Principal Investigators:
Roger Ghanem and Thorsten Becker

University of Southern California, Los Angeles

March, 2011

Summary

We study continuous geodetic network data, such as those recorded by EarthScope PBO GPS location solutions, in the context of crustal deformation and fault loading. Two different transient detection algorithms are analyzed, in particular regarding their ability to be fully adapted to the crustal deformation time series. In one case, the empirical mode decomposition (EMD) method is used to separate the nonlinear and non-stationary time series into distinct oscillating modes called the intrinsic mode functions (IMFs). Then, following the enhanced Hilbert-Huang transform methodology, the IMFs are fitted to a time-dependent, vector autoregressive, moving-average model (VARMA) prior to being analyzed in terms of their frequency distributions. Change occurs when the statistical content of the VARMA coefficients changes, which allows us to segment the time series into distinct states. In a second approach, each time series is modeled as an autoregressive (AR) processes with time-varying order and time-varying coefficients such that our problem is reduced to an optimal filtering exercise; i.e. the sequential estimation in time of the unknown hidden sates, which consist of the AR order as well as the real and complex reciprocal poles of the characteristic polynomial of the AR process. The detection of transient signals within the data set provides some insight regarding the nature and the scale of the different dynamics driving the system.

The grant mainly provided funding for USC engineering graduate student Maud Comboul. Initial results were presented at the 2010 SCEC annual meeting, where we also participated in the Transient Detection Workshop (September 2010).

Current Results

For the purposes of the analysis, we assume that the location data collected based on the GPS devices is the only available representation of the real, dynamic system. This study then attempts to understand and predict the underlying mechanical drivers from the GPS data, without any a priori physical model. We must, however, carefully interpret those measurements, inasmuch as the observed displacement fields emerge from a conjunction of processes acting at multiple scales. This complexity characterizing the structural earth system is reflected in inherently non-stationary and nonlinear time series. Furthermore, the network of GPS stations is rapidly increasing and the data recorded on a daily basis is vast, therefore, it is also crucial to develop the ability to analyze the incoming measurements efficiently and automatically. The present work explores data-adapted algorithms, which could process and analyze an entire time series data set without supervision. In the following sections, we briefly describe the two different approaches we have looked at so far. Those are complementary to more established approaches used or further developed by the SCEC community.

Empirical Mode Decomposition approach

Most data processing techniques require linearity or stationary assumptions in order to be successful; however the real world is usually neither one. In our case, the main assumption would be that crustal velocities are constant and location time-series linear. This tectonic background motion is clearly interrupted by transient behavior caused by perturbations, both physical, related to the earthquake cycle, and based on other deterministic, but less interesting (e.g. hydrologic) sources, or noise (Figure 1). Another characteristic, as one observes several crustal deformation records from the GPS network, is the non-uniform nature of the time scale lengths associated with transient events. The Empirical Mode Decomposition (EMD) and the Ensemble Empirical Mode Decomposition (EEMD) methods proposed by Huang et al. (1998) to analyze data, do not assume any a priori representation such as basis functions for the data, they only suppose that the data is the result of superposed oscillatory modes having various frequencies. The decomposition provides an adaptive framework to study data generated from nonlinear and non-stationary processes. Components obtained from the EMD are called intrinsic mode functions (IMF). The procedure to decompose the signal into distinct IMFs involves an iterative algorithm called the *sifting process*. We use the EEMD over the EMD to reduce *mode mixing*, i.e. when different oscillating modes remain in a single IMF, which often happens with signal showing *intermittent* characteristics.

Given a time series $x(t)$, we applied the sifting algorithm on an ensemble of N time series

$$(y_i(t) = x(t) + \mu_i(t))_{i=1,\ldots,N} \tag{1}$$

consisting of the original signal perturbed with normal white noise μ_i, and took the average of the IMF ensemble. The EEMD generates multiple mono-component time series (IMFs) from a scalar nonlinear and non-stationary signal. Figure 2 demonstrates the oscillating modes from fastest to slowest. The IMFs are narrow band signals with the interesting property of having the same number of extrema and zero crossings. In addition, prediction of the time series forward in time is reduced to finding the locations and the values of the next extrema.

Next, the IMFs of the crustal displacements time series are fitted with a time varying VARMA process (Yinfeng et al., 2006). In its general form, the $ARMA(p, q)$ is the combination of an autoregressive process of order p with a moving average part of order q. Those models are widely used to fit weakly stationary time series. However, in this work, the coefficients of the model evolve with time and adapt to the non-stationarities of the signal. Assuming the VARMA coefficients evolve according to a random walk in time, the Kalman Filter was used to update the coefficients in time. The state vector composed of the VARMA coefficients is a Markov process driven by a random walk in time. The first np rows of the vector correspond to the Autoregressive coefficients while the remaining nq rows are associated with the Moving Average coefficients. In this study, we compute the joint probability distribution functions (pdf) of subsets of coefficients and observe their evolution in time. For example, results from the technique applied to the north component of the crustal displacements recorded from EarthScope PBO at an example station location (Figure 1) are shown here. To generate Figure 3, we computed the joint pdf of the AR coefficients corresponding to the third IMF over twenty five-day time windows. Three peaks clearly stand out from the picture, which suggests that the system undergoes three states during that time. In the bottom picture, we show the segmented time series where the color changes when a change in the successive pdfs is observed. The L_2-norm was used in this specific case for change detection; let $f_t(X)$ be the pdf of the VARMA coefficients X at time t, then change occurs between time $t-1$ and t if $\|f_t(X) - f_{t-1}(X)\| > \epsilon$ where ϵ is a small chosen threshold. This process clearly identifies three different states of the time series.

It remains to be evaluated if this segmentation is robust, and if it might allow an association of timeseries segments with long-term trends, precursory accelerated or decelerated slip, and post-seismic transients. We are evaluating synthetic models with predefined changes in the time series at present to fine tune the algorithm.

Reciprocal pole representation of TVAR model

This method investigates a time series representation based on the observation that the power spectra of the signal changes with time. Time series are modeled using a popular stochastic representation, namely the Autoregressive (AR) process. However, we let the order of the AR processes as well as the corresponding coefficients change in time in order to adapt to dynamical changes in the signal. Let

$$y_t = \sum_{i=1}^{p_t} \Phi_i(t) y_{t-i} + \sigma_t v_t$$

be the AR representation of a time series $\{y_t; t = 1, \ldots, T\}$, then the reciprocal poles of the AR characteristic polynomial $\Phi_t(u) = 1 - \Phi_1 u - \ldots - \Phi_{p_t} u^p$ are the eigenvalues of the matrix

$$\mathbf{G}(\mathbf{\Phi(t)}) = \begin{pmatrix} \Phi_{1:p_t-1}(t) & \Phi_{p_t}(t) \\ \mathbf{I_{p_t-1}} & \mathbf{0_{p_t-1}}. \end{pmatrix}$$

If we denote by $R_t = (r_t, c_t)$ such that $p_t = r_t + 2c_t$ the number of real poles r_t and complex poles c_t of the characteristic polynomial $\Phi_t(t)$ at time t, then we can recast it in terms of its poles:

$$\Phi_t(u) = \prod_{j=1}^{c_t} \left(1 - \rho_j e^{-2\pi i \theta_j}\right)\left(1 - \rho_j e^{2\pi i \theta_j}\right) \prod_{j=c_t+1}^{c_t+r_t} (1 - \rho_j u)$$

where $\{\alpha_j = (\rho_j, \theta_j); j = 1, \ldots, c_t\}$ are the modulus and frequency of the j^{th} complex root and $\{\alpha_j = \rho_j; j = c_t + 1, \ldots, c_t + r_t\}$ is the modulus of the j^{th} real root. The algorithm sequentially learns the hidden state $X_t = (R_t, \alpha_t, \sigma_t)$ representing the poles of the characteristic polynomial $\Phi_t(u)$ associated with the time varying AR model and the noise variance of the observation model: $y_t = \sum_{i=1}^{p_t} \Phi_i(t) y_{t-i} + \sigma_t v_t$.

The optimal filtering problem consists of the estimation of the posterior distributions $p(X_{0:t}|y_{0:t})$ for $t = 1, \ldots, T$, which can be recursively obtained from $p(X_{0:t-1}|y_{0:t-1})$ with Bayes formula. Because R_t is a Markov chain taking discrete values and characterized by transition probabilities and (α_t, σ_t) is a Markov process with state transition functions, we use a particle filter algorithm, which approximates the wanted conditional distributions with discrete probability measure defines by weighted particles. For efficiency purposes, we combine importance sampling and resampling techniques with the unscented Kalman filter to avoid degeneracy of the particle distribution in time (Andrieu et al., 2003). We tried the algorithm on synthetic data presenting a transient in its power spectra. Despite of the relatively high computational costs, the algorithm appears to correctly identify the location and duration of the transient signal. We are working on a comparative analysis. Once both algorithms are more extensively tested, we intend to partake in the actual SCEC transient detection exercise.

Conclusion

For the sake of analysis, we assume that the crustal deformation time series that are available from the GPS network are the only representations of reality that we are confident about, minus outliers and errors such as due to faulty stations, and harmonic signals that are of lesser interest, such as those due to climatic changes. Data analysis is therefore an obligatory step to eventually discover the fundamental mechanical drivers of the system. Because the processes affecting the measurements act at multiple scales (Figure 1), we use the EMD in one case and an optimal filtering technique in another case to distinguish the different frequency and amplitude modes in each time series. The expansion of the analysis to the full GPS network data could potentially allow us to define different states for the underlying physical process as well as identify the current state of the system with the corresponding transition probabilities of the system jumping into a different state.

References

Andrieu, C., Davy, M., Doucet, A., and Stat. Group, Univ. of Bristol, U. (2003). Efficient particle filtering for jump Markov systems. Application to time-varying autoregressions. *Signal Processing*, 51:1762–1770.

Huang, N., Shen, Z., Long, S., Wu, M., Shih, H., Zheng, Q., Yen, N., Tung, C., and Liu, H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. A.*, 454:903–995.

Yinfeng, D., Shuyan, J., and Yingmin, L. (2006). An enhanced Hilbert-Huang transform method for analysis of earthquake ground motions. *ICEE 2006: 4th International Conference on Earthquake Engineering Taipei, Taiwan*.

Figures

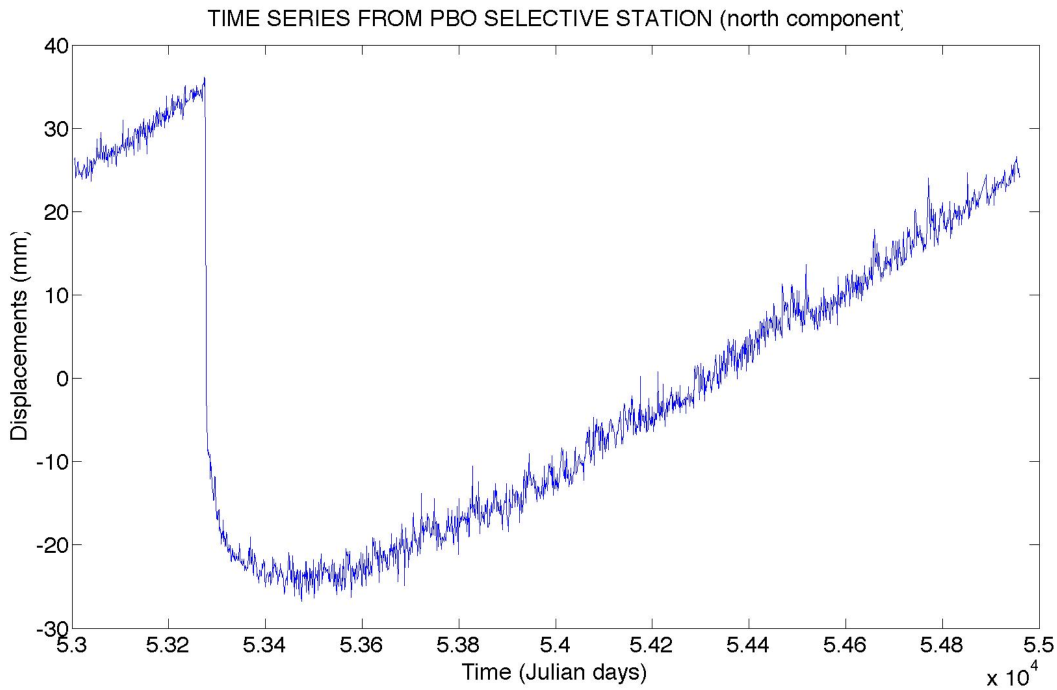

Figure 1: Original time series from PBO daily solutions for an example station location (north direction).

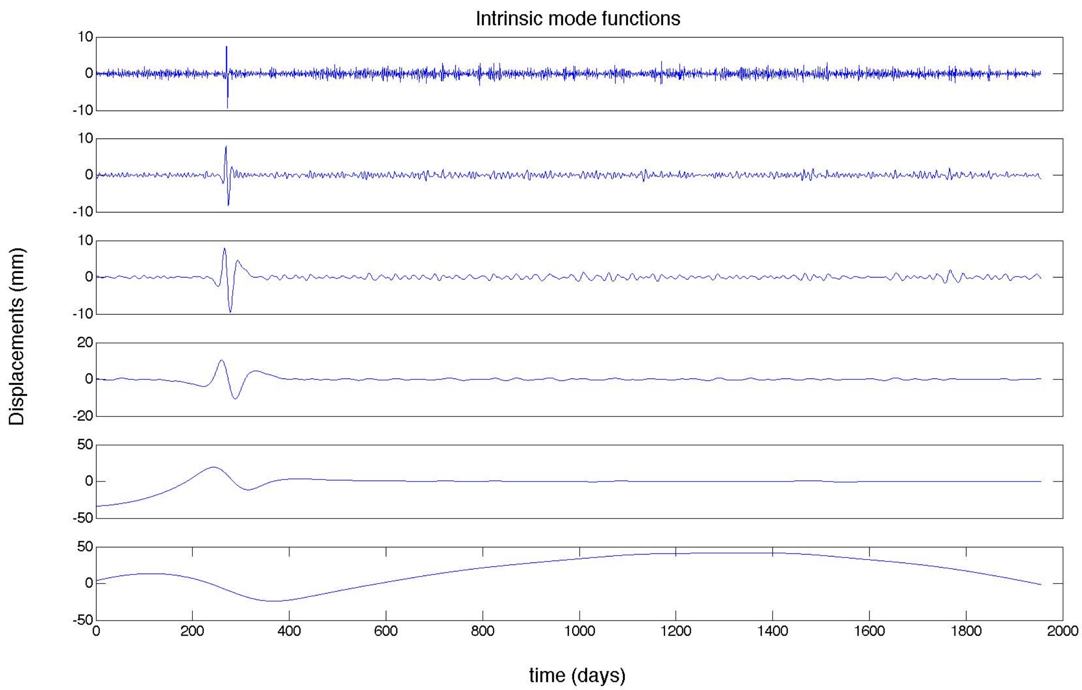

Figure 2: Resulting IMF from EMD analysis performed on PBO daily solutions for an example station location (Horizontal axis: Julian days, Vertical axis: mm), as in Figure 1.

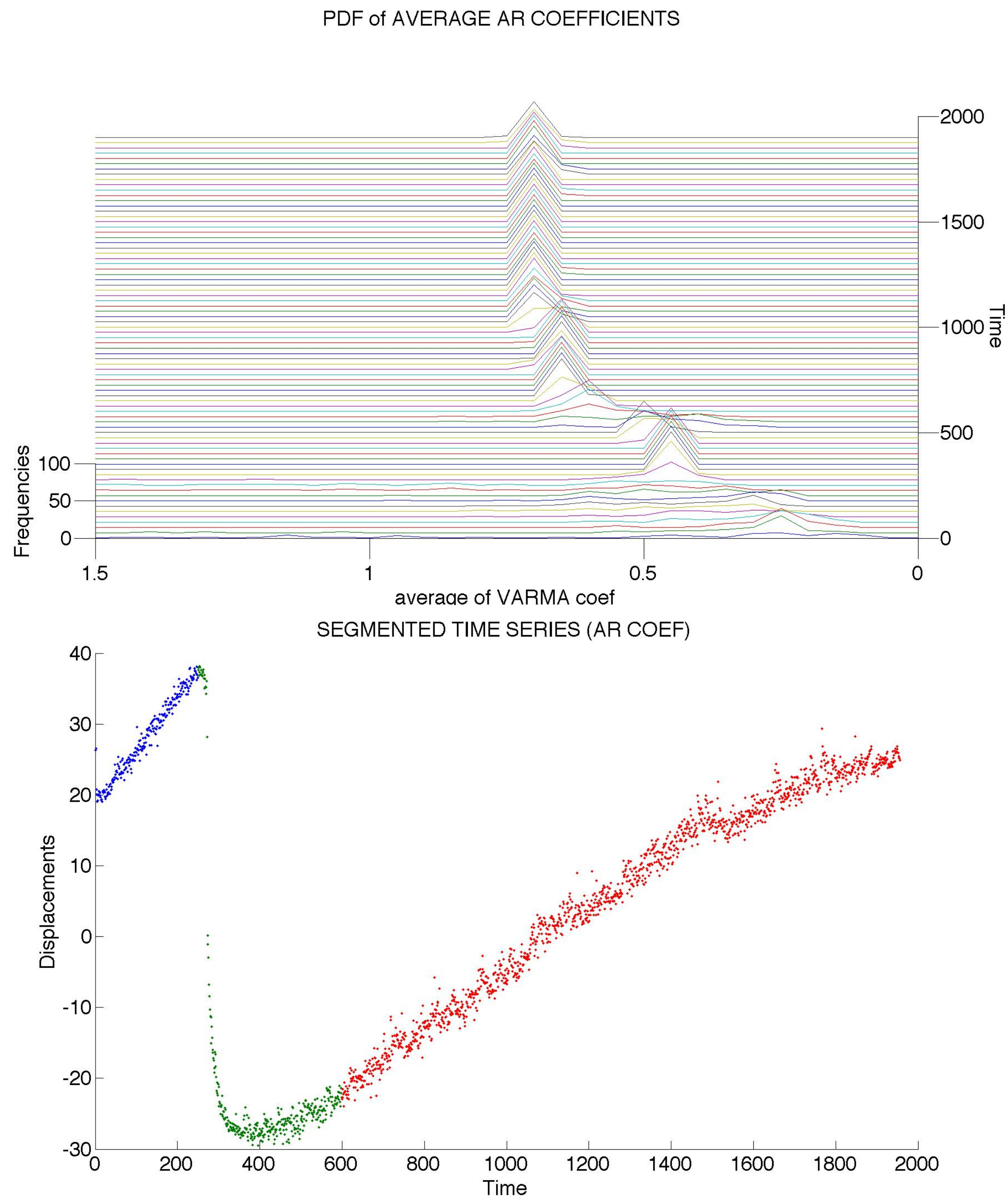

Figure 3: Top: time evolution of the autoregressive coefficient distribution for the third IMF. Bottom: segmented time series with respect to changes in the AR coefficients distribution of the third IMF, based on Figure 2.