Analysis of streaming GPS measurements of surface displacement through a web services environment

Robert Granat*, Galip Aydin†‡, Marlon Pierce‡, Zhigang Qi†‡
*Jet Propulsion Laboratory, Pasadena, CA 91109
Email: granat@jpl.nasa.gov
†Department of Computer Science
Indiana University, Bloomington IN 47404
Email: {gaydin, zqi}@cs.indiana.edu
‡Community Grids Laboratory
Indiana University, Bloomington, IN 47404
Email: mpierce@cs.indiana.edu

***Abstract*— We present a method for performing mode classification of real-time streams of GPS surface position data. Our approach has two parts: an algorithm for robust, unconstrained fitting of hidden Markov models (HMMs) to continuous-valued time series, and SensorGrid technology that manages data streams through a series of filters coupled with a publish/subscribe messaging system. The SensorGrid framework enables strong connections between data sources, the HMM time series analysis software, and users. We demonstrate our approach through a web portal environment through which users can easily access data from the SCIGN and SOPAC GPS networks in Southern California, apply the analysis method, and view results. Ongoing real-time mode classifications of streaming GPS data are displayed in a map-based visualization interface.**

I. Introduction

GPS measurements of crustal displacement are used in scientific investigation into the nature of earthquake fault behavior and the earthquake cycle. In order to detect pre- or post-seismic stress changes on or around earthquake faults, we wish to perform mode segmentation of GPS time series, as time series segmentation enables identification of subtle signals and mode transitions. This is useful for both interactive data exploration and automated detection schemes that operate on incoming data.

Our data is drawn from the Southern California Integrated GPS Network (SCIGN) [1] and Scripps Orbit and Permanent Array Center (SOPAC) [2] sensor networks; we use both the real-time (1Hz for most stations) three-dimensional position information and the daily averaged solutions. The observed GPS measurements are generated by an underlying system that is both noisy and poorly understood; the driving forces on the system derive not only from the physical processes of the solid earth but also from external factors, including atmospheric effects and human activity.

Fitting an hidden Markov model (HMM) to time series allows us to describe the statistics of the data in a simple way that ascribes discrete modes of behavior to the system. By matching incoming data against the statistics of previously learned modes, we can perform classification according to the best match. In addition, it is possible to perform signal detection across the entire sensor web by detecting simultaneous mode changes; a significant number of mode changes across the network or within a certain sub-network is an indication of an event that is occurring over a wide geographical area.

For many applications, reliable HMM fitting results are achieved by using a priori information to encode constraints that reduce the number of free parameters [3]–[9]. For GPS data, however, this information is not available as the underlying geophysical system is not well understood. As a result, we used the regularized deterministic annealing expectation-maximization (RDAEM) algorithm [10] to perform the fit. This method constructs high-quality, self-consistent model fits without using a priori information (although it does not exclude the use of such information where available), at the cost of some additional computation time.

To link this HMM technology to both the GPS data streams and users, we used the SensorGrid architecture. This provides a service-oriented approach for coupling real-time sensor messages with scientific applications in a Grid environment. Real-time data processing is supported by employing filters around a publish/subscribe messaging system.

We begin in Section II with a description of the RDAEM algorithm for fitting hidden Markov models (HMMs) and provide an example of its benefits as compared to standard methods on a sample GPS data set. We follow in Section III with a description of the SensorGrid architecture and how it is used to link the HMM software to data and users. In Section IV we show the results of the method as applied to real-time GPS streams and our map-based visualization interface, and in Section V we describe how the method can be applied to detect signals across an entire GPS network.

II. Fitting Hidden Markov Models

We start our description of the RDAEM algorithm for fitting HMMs by establishing some notation: a hidden Markov model λ with N states is composed of a vector of initial state probabilities $\pi = (\pi_1, \ldots, \pi_N)$, a matrix of state-to-state transition probabilities $A = (a_{11}, \ldots, a_{ij}, \ldots, a_{NN})$, and the observable output probability distributions $B = (b_1, \ldots, b_N)$.

The observable outputs can be either discrete or continuous. Here we are concerned with continuous valued outputs with probability distributions denoted by $b_i(y, \theta_i)$ where y is the real-valued observable output (scalar or vector) and the θ_is are the parameters describing the output probability distribution. For the normal distribution we have $b_i(y, \mu_i, \Sigma_i)$. An observation sequence O of length T is denoted $O_1 O_2 \cdots O_T$ and a state sequence Q of the model is denoted $q_1 q_2 \cdots q_T$.

A. Deterministic Annealing

Deterministic annealing is a technique based on statistical mechanics, used to mitigate the inherent sensitivity to initial conditions of the standard expectation-maximization (EM) method. It does this by using the principle of maximum entropy to specify an alternative posterior probability density for the hidden variables, allowing us to define a new effective cost function depending on a temperature parameter. This new cost function is analogous to the thermodynamic free energy. Maximization of the likelihood of the HMM at a given temperature is achieved via minimization of this cost function. Deterministic annealing attempts to avoid the standard problems associated with simulated annealing [11], such as relaxation from initial conditions and false local minima in the free energy, through deterministic optimization of the cost function at each temperature. In theory, it provides similarly global solutions with less variance, while reducing average computational cost.

Application of the deterministic annealing method to HMM optimization was proposed both by Rose and Rao [12] and by Granat and Donnellan [13]. Our deterministic annealing approach is similar to that used by Rose and Rao but differs in two important respects. First, since we address problems in which labeled training data is not available, it is not a supervised training method, and thus optimizes the likelihood rather than the minimum classification error. Second, it employs EM rather than gradient descent at each temperature. Our method is described in full in [10] but can be summarized as follows: at each computational temperature we follow an EM-like procedure to minimize the free energy, on the kth iteration of which we optimize over the function

$$\begin{aligned} U(\gamma, \lambda|\lambda^{(k)}) = &\sum_{i=1}^{N} \tau_{i1}^{(k)}(\gamma) \log \pi_i \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1} \tau_{ijt}^{(k)}(\gamma) \log a_{ij} \\ &+ \sum_{i=1}^{N}\sum_{t=1}^{T} \tau_{it}^{(k)}(\gamma) \log b_i(O_t), \end{aligned} \quad (1)$$

where $1/\gamma$ is the computational temperature and

$$\tau_{it}(\gamma) = \frac{\alpha_t(i,\gamma)\beta_t(i,\gamma)}{\sum_{i=1}^{N} \alpha_t(i,\gamma)\beta_t(i,\gamma)}, \quad (2)$$

and

$$\tau_{ijt}(\gamma) = \frac{\alpha_t(i,\gamma) a_{ij}^{\gamma} b_j^{\gamma}(O_{t+1}) \beta_{t+1}(j,\gamma)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i,\gamma) a_{ij}^{\gamma} b_j^{\gamma}(O_{t+1}) \beta_{t+1}(j,\gamma)}. \quad (3)$$

The modified forward and backward variables $\alpha(\gamma)$ and $\beta(\gamma)$ are calculated by a modification of the standard iteratitive procedure:

1) Initialization:

$$\alpha_t(i,\gamma) = \pi_i^{\gamma} b_i^{\gamma}(O_1), \ i = 1, \ldots, N. \quad (4)$$

$$\beta_T(i,\gamma) = 1, \ i = 1, \ldots, N. \quad (5)$$

2) Induction:

$$\begin{aligned} \alpha_{t+1}(j,\gamma) = &\left[\sum_{i=1}^{N} \alpha_t(i,\gamma) a_{ij}^{\gamma}\right] b_j^{\gamma}(O_{t+1}), \\ &t = 1, \ldots, T-1, \\ &j = 1, \ldots, N. \end{aligned} \quad (6)$$

$$\begin{aligned} \beta_t(i,\gamma) = &\sum_{j=1}^{N} a_{ij}^{\gamma} b_j^{\gamma}(O_{t+1}) \beta_{t+1}(j,\gamma), \\ &t = T-1, \ldots, 1, \\ &i = 1, \ldots, N. \end{aligned} \quad (7)$$

B. Regularized Deterministic Annealing

Previous approaches to regularizing the HMM optimization process have been based on the use of statistical priors to modify the HMM objective function. These priors manifest themselves in the calculations as regularization terms added to the so-called Q-function maximized during the M-step of the EM algorithm. Examples of such include Dirichlet-type priors used to prevent overtraining of discrete output distributions on limited data sets [14] and maximum entropy priors used for pruning the HMM structure [15]. In addition, statistical priors have been used more generally to provide upper bounds on the objective function of models with continuous output distributions. For instance, the conjugate prior for Gaussian output distributions described by Ormoneit and Tresp [16] wards against solutions with infinitely narrow Gaussian outputs.

Our regularized DAEM method is designed to discourage HMM local optima characterized by component states with identical or nearly identical forms. The experimental observation that local maxima solutions of this type are produced quite frequently by the DAEM method is in agreement with previous investigations into Naive Bayes networks [17].

Recall that for an HMM, the Q-function is

$$\begin{aligned} Q(\lambda, \lambda^{(k)}) = &\sum_{i=1}^{N} \tau_{i1}^{(k)} \log \pi_i \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1} \tau_{ijt}^{(k)} \log a_{ij} \\ &+ \sum_{i=1}^{N}\sum_{t=1}^{T} \tau_{it}^{(k)} \log b_i(O_t). \end{aligned}$$

Since this is separable in π, A, and B, we can divide it into the sum of three functions: $Q_1(\pi)$, $Q_2(A)$, and $Q_3(B)$. Although other priors are possible [10], for conceptual and computational simplicity we chose to use an improper prior on

the HMM likelihood, based on the squared Euclidean distance between the means of the state output distributions:

$$P_3(B) = \prod_{i=1}^{N}\prod_{j=1}^{N} \exp\Big(\frac{\omega_{Q_3}}{2}(\mu_i - \mu_j)^T(\mu_i - \mu_j)\Big), \quad (8)$$

where $\omega_{Q_3} > 0$ is a weighting term. This prior rewards solutions with widely spaced output distributions. From it we derive the modified Q-function with

$$\begin{aligned} Q_3' = \sum_{i=1}^{N}\sum_{t=1}^{T} \tau_{it}^{(k)} \Bigg(& \log n - \frac{1}{2}\log\det(\Sigma_i) \\ & - \frac{1}{2}(m_i - \mu_i)^T\Sigma_i^{-1}(m_i - \mu_i) \\ & - \frac{1}{2}(O_t - m_i)^T\Sigma_i^{-1}(O_t - m_i) \\ & + \frac{\omega_{Q_3}}{2}\sum_{j=1}^{N}(\mu_i - \mu_j)^T(\mu_i - \mu_j)\Bigg) \end{aligned} \quad (9)$$

where $m_i = \sum_{t=1}^{T}\tau_{it}^{(k)}O_t / \sum_{t=1}^{T}\tau_{it}^{(k)}$. Note that for the ease of subsequent manipulation and computation, we do not properly account for the independence of the prior from the hidden state variable. In theory, systems with highly skewed populations of observations from each of the output distributions may be misrepresented in this regularization scheme. In practice, we have observed no evidence of any serious consequence; nevertheless we keep in mind that the regularized Q-function (9) is only an approximation.

To derive the revised EM update rule we first take the vector derivative of Q_3' in the means:

$$\frac{\partial Q_3'}{\partial \mu_i^T} = \Sigma_i^{-1}(m_i - \mu_i) - \omega_{Q_3}\sum_{j=1}^{N}\mu_j + N\omega_{Q_3}\mu_i. \quad (10)$$

Setting the derivative to zero we have

$$\Sigma_i^{-1}m_i + (N\omega_{Q_3}I - \Sigma_i^{-1})\mu_i = \omega_{Q_3}\sum_{j=1}^{N}\mu_j, \quad (11)$$

for $i = 1, \ldots, N$, which leads us to the system of equations

$$\begin{aligned} \Bigg(\begin{bmatrix} \Sigma_1^{-1} & & \\ & \ddots & \\ & & \Sigma_N^{-1} \end{bmatrix} & \\ + \omega_{Q_3} \begin{bmatrix} I_{D\times D} & \cdots & I_{D\times D} \\ \vdots & \ddots & \vdots \\ I_{D\times D} & \cdots & I_{D\times D} \end{bmatrix} \Bigg) U & \\ - N\omega_{Q_3}U = & \\ \begin{bmatrix} \Sigma_1^{-1} & & \\ & \ddots & \\ & & \Sigma_N^{-1} \end{bmatrix} M, & \end{aligned} \quad (12)$$

where $M = [m_1; \cdots; m_N]$, $U = [\mu_1; \cdots; \mu_N]$, and $I_{D\times D}$ is a $D \times D$ identity matrix. This system can be solved by any standard linear method given the inverse covariances Σ_i^{-1}.

We evaluate the conditions on the weighting term ω_{Q_3} under which this solution is the global maximum by calculating the Hessian:

$$\frac{\partial^2 Q_3'}{\partial \mu_i^T \partial \mu_j} = \begin{cases} \omega_{Q_3}(N-1)I - \Sigma_i^{-1} & \text{if } i = j \\ -\omega_{Q_3} & \text{otherwise} \end{cases} \quad (13)$$

If the Q-function is concave, then the Hessian H is negative definite, and so for all nonzero column vectors x composed of stacked $N \times 1$ vectors x_i,

$$\begin{aligned} x^T H x < 0 & \\ \sum_{i=1}^{N} x_i^T\Big(N\omega_{Q_3}I - \Sigma_i^{-1}\Big)x_i - \omega_{Q_3}\sum_{i=1}^{N}x_i^T\sum_{j=1}^{N}x_j < 0 & \\ 2N\omega_{Q_3}\sum_{i=1}^{N}x_i^Tx_i - \sum_{i=1}^{N}x_i^T\Sigma_i^{-1}x_i \le 0 & \\ 2N\omega_{Q_3}\sum_{i=1}^{N}x_i^Tx_i - \sum_{i=1}^{N}x_i^Tx_i||\Sigma_i^{-1}|| \le 0 & \\ \omega_{Q_3} \le ||\Sigma_i^{-1}||/2N. & \end{aligned} \quad (14)$$

This gives us a condition on ω_{Q_3} for the Q-function to have a global maxima in the means. Since the means are unbounded, the solution does not lie on a boundary and we can have confidence in the validity of using a positive regularization term for a maximization problem.

To find the maximum in the covariances, we take the derivatives of (9) in the components of Σ_i^{-1} and set them equal to zero, which gives us

$$\Sigma_i = \sum_{t=1}^{T}\tau_{it}^{(k)}(O_t - \mu_i)(O_t - \mu_i)^T / \sum_{t=1}^{T}\tau_{it}^{(k)}. \quad (15)$$

This is a global maximum since the Q-function is concave as a function of the covariances Σ_i. To find the means and covariances we need to solve equations (12) and (15) simultaneously. This can be done by using the approximation $\Sigma_i = S_i$, where

$$S_i = \sum_{t=1}^{T}\tau_{it}^{(k)}(O_t - m_i)(O_t - m_i)^T / \sum_{t=1}^{T}\tau_{it}^{(k)}, \quad (16)$$

in equation (12) as an initial guess and then iterating between equations (12) and (15) until the solution converges. In practice, it is usually sufficient merely to approximate Σ_i as S_i when calculating the means without any attempt at iterative convergence whatsoever. This iterative rather than direct maximization forces us to characterize the method as a generalized EM algorithm rather than a pure EM approach.

C. RDAEM Fitting Results

Although a full analysis of the performance of the RDAEM algorithm is beyond the scope of this work, we provide a brief example of the benefits of this approach over standard methods. In this example, we test the following model-fitting algorithms: (1) the baseline EM algorithm (`em_basic`), (2) basic EM wrapped inside a loop that runs the algorithm

ten times and then picks as the final solution the one with the highest likelihood (em_ntries), (3) the DAEM algorithm with starting computational temperature $\gamma_{min} = 0$ and annealing schedule $\gamma_{new} = \gamma_{old} + 0.01$ (daem_step), (4) the DAEM algorithm with $\gamma_{min} = 0.1$ and annealing schedule $\gamma_{new} = 1.1\gamma_{old}$ (daem_geometric), (5) the RDAEM algorithm with ω_{Q_3} set to the maximum at each iteration (rdaem), (6) the RDAEM algorithm with $\omega_{Q_3} = 1$ (rdaem_small), (7) the RDAEM algorithm with ω_{Q_3} set to the maximum at each iteration, wrapped in an outer loop as per em_ntries (rdaem_ntries), and (8) the RDAEM algorithm algorithm with $\omega_{Q_3} = 1$, likewise wrapped in an outer loop as per em_ntries (rdaem_ntries_small). Methods 5-8 used the same geometric annealing schedule as in daem_geometric. In all cases, an HMM with Gaussian output distributions was used to analyze the data, and a conjugate prior of the type described in [16] with weight factor 10^{-6} was used to bound the output distributions above. To avoid unit and scale disparity effects, each dimension of the input was shifted and normalized to lie between 0 to 1.

Figure 1 presents a summary of the results for an experiment performed on approximately two years of daily GPS solutions collected by a station in Claremont, California. On the left are two graphs showing the mean log likelihood of the solutions found by different methods, each averaged across 100 experimental trials; the results for the methods em_basic and em_ntries are plotted in both upper and lower charts for reference. On the right side are two graphs showing the number of different solutions found by the different methods, out of 100 trials. We see that rdaem_ntries_small provides the best performance overall, losing nothing in terms of log likelihood to any other method while providing considerably improved solution stability (fewer distinct solutions). Although em_ntries is a close match in log likelihood, the stability advantage of rdaem_ntries_small over em_ntries is large, up to four times as much for some model sizes.

III. SensorGrid

SensorGrid is our research project that aims to build the distributed computing infrastructure (i.e. "Grid") to support real-time streaming data. SensorGrid is also the name for our software, built using topic-based publish/subscribe and Web Service concepts. The architecture is general but we describe here its specific application to managing GPS data streams. In particular, we focus on the integration of an implementation of the RDAEM algorithm for real-time state change detection in GPS signals.

This architecture consists of three major components in addition to the actual sensor nodes: individual filters that process the real-time data streams, information services that provide the metadata about the filters or filter chains, and a Grid messaging system, which provides supports in areas like fault tolerance, notification, and recovery. The filters are run in specific orders to achieve particular processing goals. This corresponds to different workflow scenarios in different Grid domains. Filters are exposed as Web Services to allow remote composition of filter chains.

A. Streaming Message Support

In a real-time data Grid the top priority is to be able to provide access and process the continuously streaming data. Any kind of interruption in this process will result in data loss. To prevent data loss or similar problems, the messages should be reliably transferred between the services. Traditional Web Services are implemented on top of the already existing Web infrastructure, making them universally accessible. However, HTTP is not appropriate for high-rate data flow requirements. For instance, if we want to disseminate real-time messages from GPS sensors that output measurements once per second, the latency associated with HTTP would cause data loss, especially for long distances. This is also true for large scale data transfers. Furthermore, since the sensors are always alive and continuously produce data the reliability and fault-tolerance of the messaging architecture is extremely important. Therefore, a better messaging solution is required for connecting the filters and the sources.

To address this problem, we employed the NaradaBrokering [18], [19] distributed messaging infrastructure. NaradaBrokering provides two related capabilities. First, it provides a message oriented middleware (MoM) which facilitates communications between entities (including clients, resources, services and proxies) through the exchange of messages. Second, it provides a notification framework by efficiently routing messages from the originators to only the registered consumers of the message in question. More details on the capabilities and advantages of NaradaBrokering can be found in [18].

B. Real-Time Data Grid Implementation for GPS Networks

The Scripps Orbit and Permanent Array Center (SOPAC) provides continuous, publicly available position and error data for stations in its GPS networks. Raw data from the GPS stations are collected by a Common Link proxy (RTD server) and archived in RINEX files. To receive station positions, clients are expected to open a socket connection to the RTD server. After the RTD server receives raw data from the stations it applies filters and for each network generates a message. This message contains a collection of instantaneous position information for every individual station, along with other information such as the quality of the measurements and their variances. For each GPS network, the RTD server broadcasts one position message per second through a port in a binary format known as RYO.

We used NaradaBrokering to provide these RTD server broadcasts as real-time streaming position information to client applications. The core of the system is an assortment of filter chains that convert or otherwise process the incoming data streams. These filters serve as both subscribers (data sinks) and publishers (data sources). A basic filter consists of three parts: a NaradaBrokering subscriber, a publisher, and a data processing unit. An abstract filter interface we developed provides subscriber and publisher capabilities. Typically

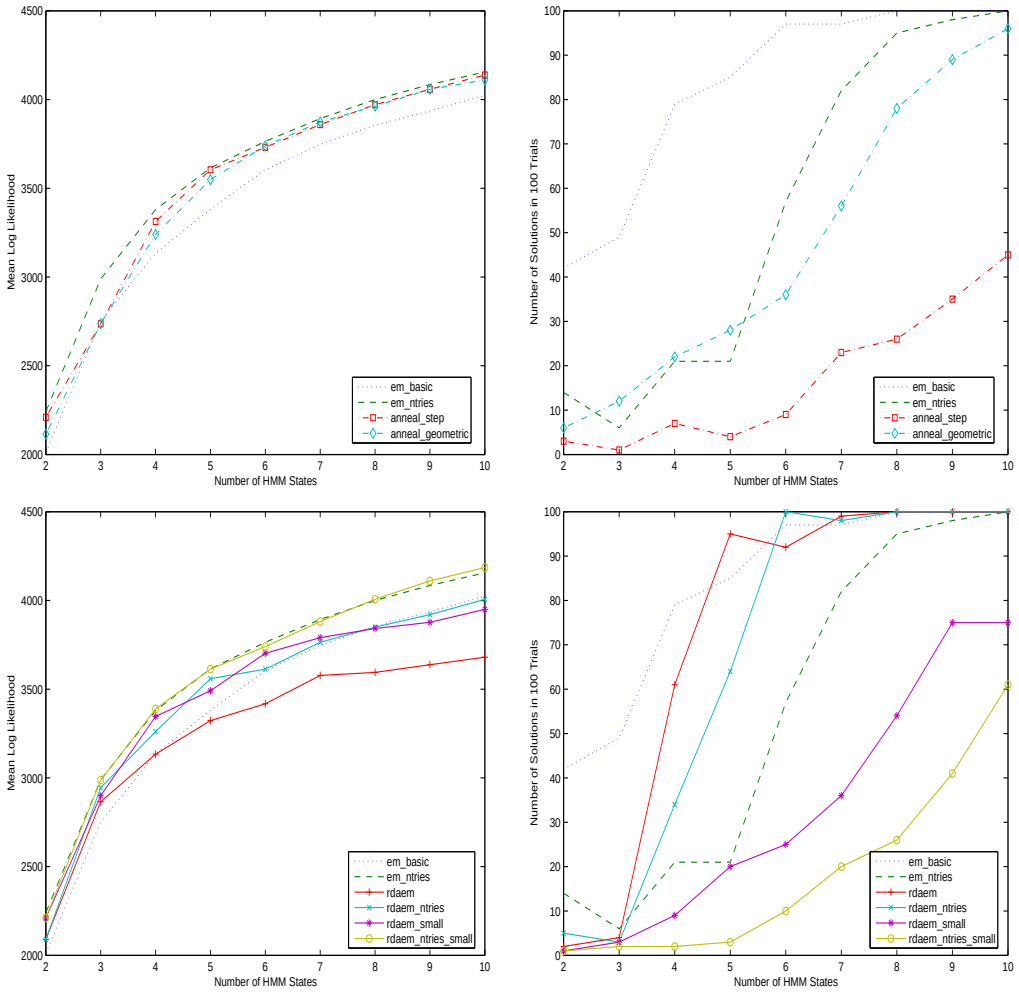


Fig. 1. A summary of the results of various HMM fitting algorithms applied to daily GPS solutions from a station in Claremont, California. Upper Left: Both annealing methods are superior to `em_basic` in log likelihood terms, but fall slightly behind `em_ntries`. Lower Left: The RDAEM methods do poorly in log likelihood terms when ω_{Q_3} is set to the maximum at each iteration, but equal or outperform their respective vanilla EM equivalents when $\omega_{Q_3} = 1$ is used. Upper Right: The faster annealing schedule used by `anneal_geometric` resulted in less stable solutions than the slower one used by `anneal_step`, but still matched `em_ntries`. Lower Right: `rdaem_ntries_small` outperforms all others in terms of solution stability by a wide margin, while `rdaem_small` still provided better stability than `em_ntries` at significantly lower computational cost.

a filter subscribes to a specified NaradaBrokering topic to receive streaming messages, processes the received data, and publishes the results to another topic. However, outputs need not be always published; for instance, a Database Filter may only receive the station positions to insert into a database. Filters can be connected in parallel or serial for realizing more complex tasks.

C. Coupling HMM Analysis with Streaming Data

The GPS data passes through a number of filters before being presented to the HMM-based analysis application. Since the data stream provided by RTD server is in a binary format, the first filters act to decode it and present it in different formats. Once we receive the original binary data, we immediately publish them to a NaradaBrokering topic (null filter). A filter that converts the binary message to ASCII subscribes to this topic and publishes the output message to another topic. Another filter application subscribes to the ASCII message topic and, using a GML schema for GPS position messages we developed, publishes the GML representation to a different topic. This approach allows us to keep the original data intact while making different message formats available to multiple clients. A diagram of the GPS filter chain architecture can be seen in Figure 2, showing how the data stream is passed through the various format conversion filters before being sent to the HMM analysis application (RDAHMM), which is treated by the architecture as just another filter.

The RDAHMM software operates in two modes. In training mode, the program finds optimal HMM parameters to fit the model to the data using the RDAEM algorithm described in Section II. In evaluation mode, the program uses a pre-existing HMM to classify observations in a time series according to the calculated state sequence. To classify modes in an incoming

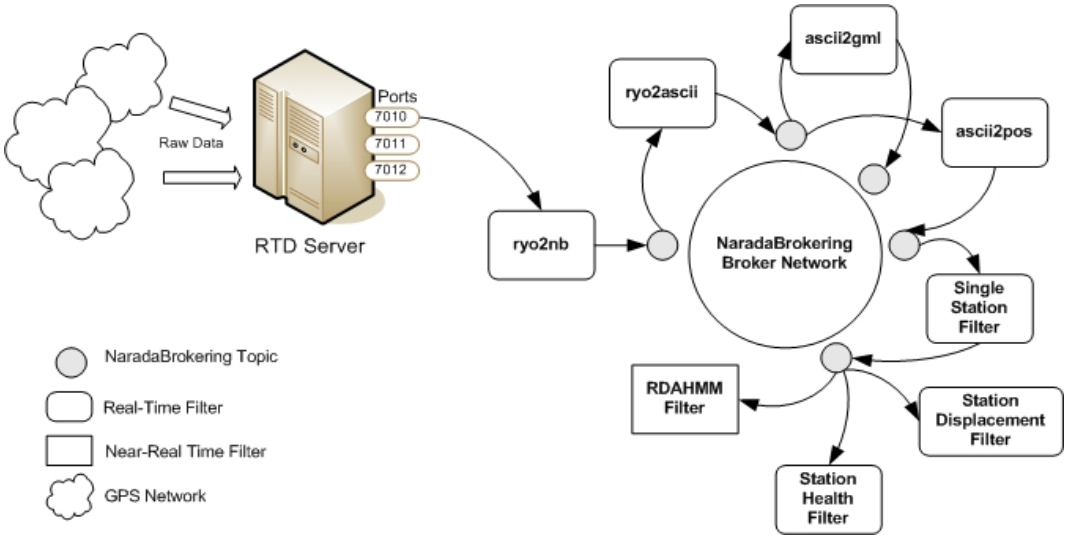


Fig. 2. The filter chain architecture used to process streaming GPS data. Several format conversion filters process the data before it is sent to the RDAHMM software that performs the HMM-based time series analysis.

data stream, we first use the RDAHMM software in training mode to fit an HMM to an initial body of data from that stream. We then run that model continuously in evaluation mode on successive time windows of data from that stream. Although the current version of the software is not completely real-time, we can run it near real-time by keeping the time window for evaluation relatively small. To do this, the RDAHMM filter listens to the ASCII position topic and accumulates a certain number of messages. As soon as the limit number is reached, the RDAHMM filter invokes the RDAHMM application in evaluation mode on the accumulated data.

IV. Single Station Real-Time Analysis

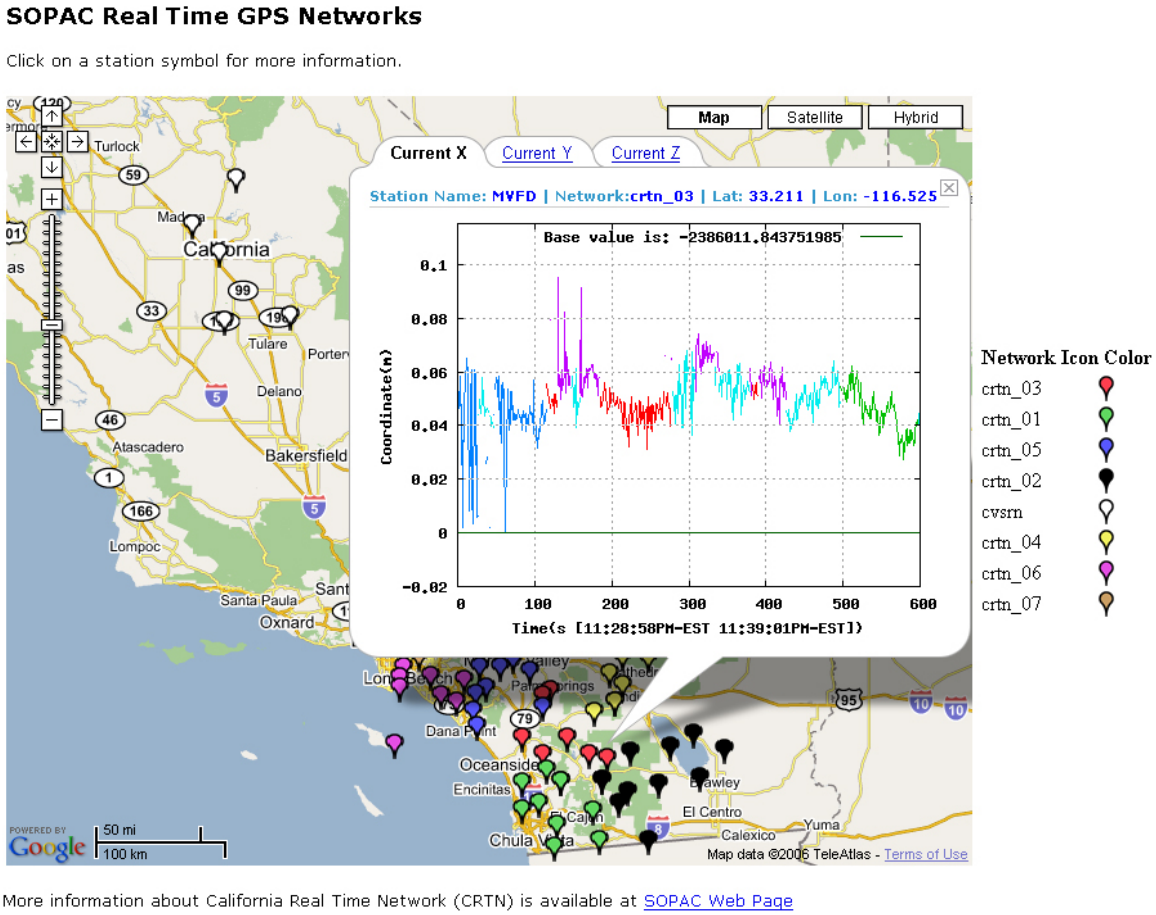
We have developed a visualization interface for analysis of the SOPAC GPS networks using Google Maps. The locations of 85 GPS stations belonging to eight networks are displayed on a map of Southern California; when a user selects one of the stations, a window displays the most recent data from that station, classified according to the trained HMM for that station. An example of this interface in use can be seen in Figure 3. A data accumulation filter was deployed to listen to the ASCII position topic of each network and pick up the individual station positions. This filter was set to collect as many as 600 data points (roughly equal to 10 minutes worth of data); these points are then sent to the RDAHMM filter. When the RDAHMM analysis completes, it triggers the visualization filter to display the classified accumulated data. The analysis was performed for the X, Y and Z positions of each station; the user can see the Y and Z axis position results by clicking on the corresponding tab. These output plots are also available via Web Service invocations and so can be integrated into other user interfaces.

V. Multiple Station Analysis

In the preceding sections we have concentrated on data flow and analysis for individual GPS stations. In this section we extend the techniques used to analyze data from a single station to the analysis of entire networks or subnetworks. We are interested in detecting geophysical events with geographically disperse signatures, including not only earthquakes but also aseismic events linked to crustal block motion or stress transfer between earthquake faults. These types of events have been observed in a few instances [20]–[26], but detections remain rare due to the subtlety of the signals. We hope to observe evidence of these types of aseismic events in the GPS data.

In our test of this approach, we used daily displacement solutions for all 127 available SCIGN stations in a 820 day window beginning Jan 1st, 1998. When GPS displacement values for a given station were not available on a particular day due to signal dropout or incomplete installation, we assumed a zero displacement measurement for that day. We note that since actual measurements are almost never of zero displacement, this in effect adds an additional "dropout" class to the data. Our next step was to train a separate hidden Markov model on each of these GPS signals. Since the GPS signals had similar statistics to one another, we could use the results of our experiments on the Claremont, California GPS station in II to estimate the model size. We observed that there were very few maxima for up to five states; adding an additional state to account for the dropout class, we used a six state model. Once these models were trained, they provided us with the state sequences for each GPS time series. We suspected that interesting geophysical events would manifest themselves as changes in the signals across multiple GPS stations, thus we looked for correlations in state changes across the network.

Figure 4 shows the number of same-day state changes across all stations using classifications given by six-state models. There are a number of strong peaks indicating correlated state changes; of note is the strong peak on day 652, which corresponds to the Hector Mine earthquake. We observe that there



More information about California Real Time Network (CRTN) is available at SOPAC Web Page

Fig. 3. A screenshot of the visualization interface to HMM-based analysis of streaming GPS data. A user selects a station from a Google Map interface; color-classified observations from that station from the last several minutes are displayed in a window.

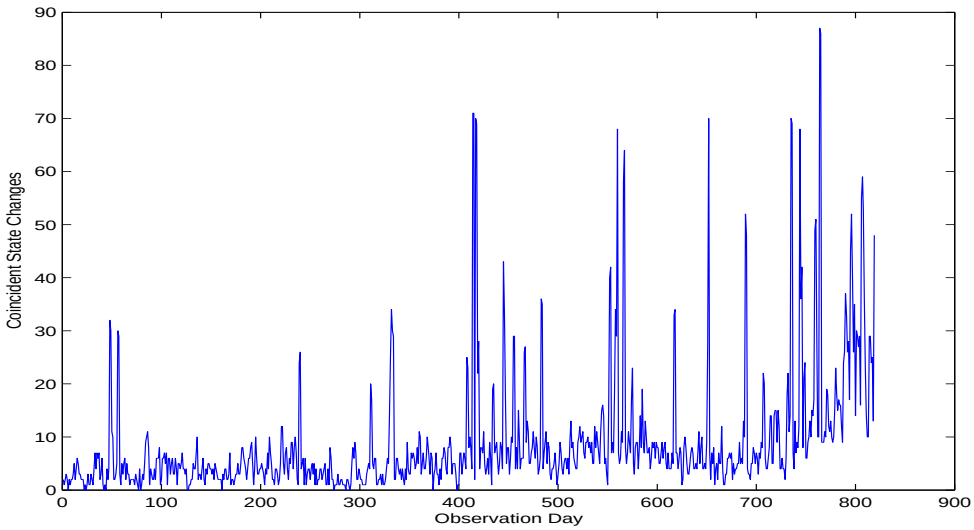


Fig. 4. Coincident state changes for six-state HMMs trained on daily position solutions from each of 127 SCIGN GPS stations.

is an increasing trend in the average number of coincident state transitions; this is because of the increasing number of stations installed and activated during the observation period. When we compare these network change correlations with the seismic record, we find that the only large event during that time period was the Hector Mine earthquake; other correlation peaks are not strongly associated with seismic events, and thus indicate aseismc events. Although the exact nature of

these events remains unclear and demands follow-on study, we can infer from the timing of the correlation peaks that some events most likely produce transitory effects, while others are more lasting. We expect events that produce transitory effects to be characterized by closely-paired correlation peaks as the regional signal departs from and returns to the baseline. Long-lasting effects, by the same logic, should only manifest single peaks. Of course, some double peaks may form by coincidence, but this observation provides a starting point for further investigation.

VI. Conclusions

We have developed and implemented a method for performing robust mode classification analysis of streaming, real-time GPS data. We linked streaming data from GPS networks in Southern California to software that performs reliable fitting of hidden Markov models even in the absence of a priori information. A web portal environment provided a visualization interface by which users could access data and HMM classification results. In addition, we demonstrated that using the classification results from multiple GPS stations can enable detection of geographically distributed signals.

There are several directions for future work in this area. We intend to improve the RDAEM HMM fitting algorithm to reduce computational cost and enable true real-time analysis. Follow-up work on regional signal detection through GPS networks is necessary to determine the origins of regional signals. In addition, we would like to add the regional signal detection capability as a service available through SensorGrid and extend its application beyond daily position time series to real-time station data. This will enable identification of unusual events happening on time scales of seconds to minutes, rather than hours to days. In addition, this technology makes possible data fusion between 1Hz seismograph and GPS data, bridging the frequency gap between these two sources. Improving the scalability and performance of the SensorGrid system is another major goal; preliminary studies indicate that the system (Figure 2) should scale to hundreds of GPS stations for a single broker.

References

[1] K. W. Hudnut, Y. Bock, J. E. Galetzka, F. H. Webb, and W. H. Young, "The Southern California integrated GPS network (SCIGN)," in *Proceedings of the International Workshop on Seismotectonics at the Subduction Zone*, Y. Fujinawa, Ed., NIED, Tsukuba, Japan, 1999, pp. 175–196.

[2] P. Jamason, Y. Bock, P. Fang, B. Gilmore, D. Malveaux, L. Prawirodirdjo, and M. Scharber, "SOPAC web site (http://sopac.ucsd.edu)," *GPS Solutions*, vol. 8, no. 4, pp. 272–277, December 2004.

[3] B.H. Juang and L.R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Transactions On Acoustics Speech And Signal Processing*, vol. 33, no. 6, pp. 1404–1413, 1985.

[4] A. Farago and G. Lugosi, "An algorithm to find the global optimum of left-to-right hidden Markov model parameters," *Problems Of Control And Information Theory-Problemy Upravleniya I Teorii Informatsii*, vol. 18, no. 6, pp. 435–444, 1989.

[5] G. McGuire, F. Wright, and M.J. Prentice, "A Bayesian model for detecting past recombination events in DNA multiple alignments," *Journal Of Computational Biology*, vol. 7, no. 1-2, pp. 159–170, 2000.

[6] Y. Ephraim, A. Dembo, and L.R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," *IEEE Transactions On Information Theory*, vol. 35, no. 5, pp. 1001–1013, 1989.

[7] J.R. Bellegarda and L.R. Nahamoo, "Tied mixture continuous parameter modeling for speech recognition," *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, vol. 38, no. 12, pp. 2033–2045, 1990.

[8] S.J. Young and P.C. Woodland, "State clustering in hidden Markov model-based continuous speech recognition," *Computer Speech And Language*, vol. 8, no. 4, pp. 369–383, 1994.

[9] E. Bocchieri and B.K.W. Mak, "Subspace distribution clustering hidden Markov model," *IEEE Trans. on Speech and Audio Proc.*, vol. 9, no. 3, pp. 264–275, 2001.

[10] R. A. Granat, *Regularized Deterministic Annealing EM for Hidden Markov Models*, Ph.D. thesis, University of California, Los Angeles, 2004.

[11] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[12] K. Rose and A. V. Rao, "Deterministically annealed design of hidden Markov model speech recognizers," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 2, pp. 111–126, 2001.

[13] R. Granat and A. Donnellan, "Deterministic annnealing hidden Markov models for geophysical data exploration," *Proc. 3rd Workshop on Mining Scientific Data*, pp. 1–8, 2001.

[14] A. Krogh, M. Brown, I.S. Mian, K. Sjolander, and D. Haussler, "Hidden Markov models in computational biology - applications to protein modeling," *Journal Of Molecular Biology*, vol. 235, no. 5, pp. 1501–1531, 1994.

[15] M. Brand, "Structure learning in conditional probability models via an entropic prior and parameter extinction," *Neural Computation*, vol. 11, pp. 1155–1182, 1999.

[16] D. Ormoneit and V. Tresp, "Improved Gaussian mixture density estimates using Bayesian penalty terms and network averaging," *Advances in Neural Information Processing Systems 8*, pp. 542–548, 1996.

[17] M. Whiley and D. M. Titterington, "Applying the deterministic annealing expectation maximization algorithm to naive Bayesian networks," *Univ. Glasgow Tech. Report*, 2002.

[18] Shrideep Pallickara and Geoffrey Fox, "NaradaBrokering: A distributed middleware framework and architecture for enabling durable peer-to-peer grids," in *Middleware*, 2003, pp. 41–61.

[19] Geoffrey Fox, Sang Lim, Shrideep Pallickara, and Marlon E. Pierce, "Message-based cellular peer-to-peer grids: Foundations for secure federation and autonomic services," *Future Generation Comp. Syst.*, vol. 21, no. 3, pp. 401–415, 2005.

[20] T.I. Melbourne and F.H. Webb, "Slow but not quite silent," *Science*, vol. 300, no. 5627, pp. 1886–1887, 2003.

[21] G. Rogers and H. Dragert, "Episodic tremor and slip on the Cascadia subduction zone: The chatter of silent slip," *Science*, vol. 300, no. 5627, pp. 1942–1943, 2003.

[22] T.I. Melbourne and F.H. Webb, "Precursory transient slip during the 2001 m_w=8.4 Peru earthquake sequence from continuous gps," *Geophysical Research Letters*, vol. 29, no. 21, pp. art. no.–2032, 2002.

[23] T.I. Melbourne, F.H. Webb, J.M. Stock, and C. Reigber, "Rapid postseismic transients in subduction zones from continuous gps," *Journal Of Geophysical Research – Solid Earth*, vol. 107, no. B10, pp. art. no.–2241, 2002.

[24] M.M. Miller, T. Melbourne, D.J. Johnson, and W.Q. Sumner, "Periodic slow earthquakes from the cascadia subduction zone," *Science*, vol. 295, no. 5564, pp. 2423–2423, 2002.

[25] H. Hirose, K. Hirahara, F. Kimata, N. Fujii, and S. Miyazaki, "A slow thrust slip event following the two 1996 Hyuganada earthquakes beneath the Bungo Channel, southwest Japan," *Geophysical Research Letters*, vol. 26, no. 21, pp. 3237–3240, 1999.

[26] K. Heki, S. Miyazaki, and H. Tsuji, "Silent fault slip following an interplate thrust earthquake at the Japan Trench," *Nature*, vol. 386, no. 6625, pp. 595–598, 1997.