SAND2008-7935P

Extensions to CANARY Water Quality Event Detection Software: Multivariate Pattern Matching

Eric Vugrin and Sean A. McKenna
Sandia National Laboratories
Albuquerque, New Mexico

November, 2008

Funding for this work was provided by PUB. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

The water quality clustering analysis consists of two major processes: analysis of historical water quality data to create a library of water quality change patterns and real-time comparison of water quality events with the template library. The following sections describe in detail the processes used to analyze the historical data and populate the pattern library in CANARY. The real-time analysis requires connection to the PUB SCADA system and this is connection is underway at this time.

1 Creation and Clustering of Water Quality Template Libraries

The analysis of the historical water quality data to create water quality template libraries is a multi-step process. The key steps in this process are:

1. Identification of water quality events in historical data;
2. Creating a template library that consists of coefficients for regression models of water quality signals immediately preceding events;
3. Clustering events to identify similar and repeated water quality changes;
4. Calculating cluster statistics.

The clustering approach developed here provides a summary of common water quality patterns against which any new water quality pattern that is seen can be quickly compared.

1.1 Identification of Water Quality Events

Before describing the process that CANARY uses to identify water quality events, it is necessary to give a very brief description about some of CANARY's objectives and outputs. CANARY requires a set of water quality signals for a particular location and period of time as input. The code analyzes the data, and the software uses the signal data from previous time steps to predict what the signal will be at the current time step. The prediction is compared with the actual signal data when it is observed to determine if the difference between the two values is "significant." Based upon the number of significant differences that occur in a moving window, CANARY calculates the probability that a water quality event has occurred. In this document, that probability is denoted as $P_C(t)$, and the dependence on time is explicitly noted. For a detailed description of this process, the reader is referred to Hart and McKenna (2008)

CANARY is being extended here to include a step in the process described above where significant differences between the predicted and observed water quality are assessed to determine if they are due to a change in the water quality that has been previously identified. The final goal is to be able to use CANARY with relatively sensitive parameter settings that will lead to increased probability of detecting anomalous water quality. The common disadvantage of increasing sensitivity is an increase in the number of false alarms. If the water quality patterns that create false alarms occur with some relatively common frequency, then it should be possible to identify these patterns, maintain them in a "pattern library" and assess any new water quality pattern that causes

an alarm against the existing patterns. If a match is found, then the current water quality is causing a false alarm and the alarm is cancelled.

The first step in the creation of water quality event template libraries is the identification of the events that will populate the library. To do so, the user is required to input a threshold probability, denoted P_{thresh}, in the input configuration file. CANARY compares the event probability with the threshold probability, and for the purposes of creating the template library, an event is defined to occur at the first time step in a continuous interval of time steps during which the event probability exceeds the threshold probability ($P_C(t) > P_{thresh}$). Table 1 contains hypothetical data to illustrate this process.

Table 1. An illustrative example to describe how events are identified: two events begin at time steps 3 and 7 and are colored red.

P_{thresh}	0.5											
$P_C(t)$	0	0.2	**0.6**	0.8	0.1	0.2	**0.6**	0.8	1	1	1	0
$P_C(t)>P_{thresh}$?	N	N	**Y**	Y	N	N	**Y**	Y	Y	Y	Y	N
Initiation of event?	N	N	**Y**	N	N	N	**Y**	N	N	N	N	N
Time Step	1	2	**3**	4	5	6	**7**	8	9	10	11	12

1.2 Creation of the Template Libraries

For each event CANARY fits a series of low order regression models to the water quality signals that are considered in the calculation of the event probability. For a particular signal, a regression model is determined for the data points that immediately precede the initiation of an event. CANARY removes any NaNs and zeros that denote missing or bad data from the dataset and then uses the MATLAB® function **`polyfit`** to perform the regression on the remaining data[1].

The CANARY user specifies in the configuration file the orders of the regression models and the numbers of data points to which the models are fit. The orders and number of data points may vary across different signal types, but they are constant across events. The regression coefficients for an event are stored in a matrix that is termed the template library. That is, the template library is an N_E by O_{Total} matrix, where N_E is the total number of events identified in the historical data and O_{Total} is the sum, over all of the water quality signals, of the orders of polynomial regression plus the number of signals considered (since a n^{th} order polynomial has n+1 coefficients). Figure 1 contains a flow chart of how the template library is created.

For example, we typically consider residual chlorine, pH, and conductivity signals when determining event probabilities. Empirical trials have determined that 3^{rd} to 5^{th} order

[1] For some signals, "2"s are removed from the regression data since SCADA systems sometimes report powers of 2 for signals (e.g., pH) when there are SCADA errors.

regression models typically work well when considering 90 data points. If each signal is fit with a 3rd order polynomial, then the first four entries of a row in the template library row contain regression coefficients for residual chlorine data, the fifth through eighth entries are regression coefficients for pH, and the last four row entries are regression coefficients for conductivity data. Thus, the template library would have twelve columns.

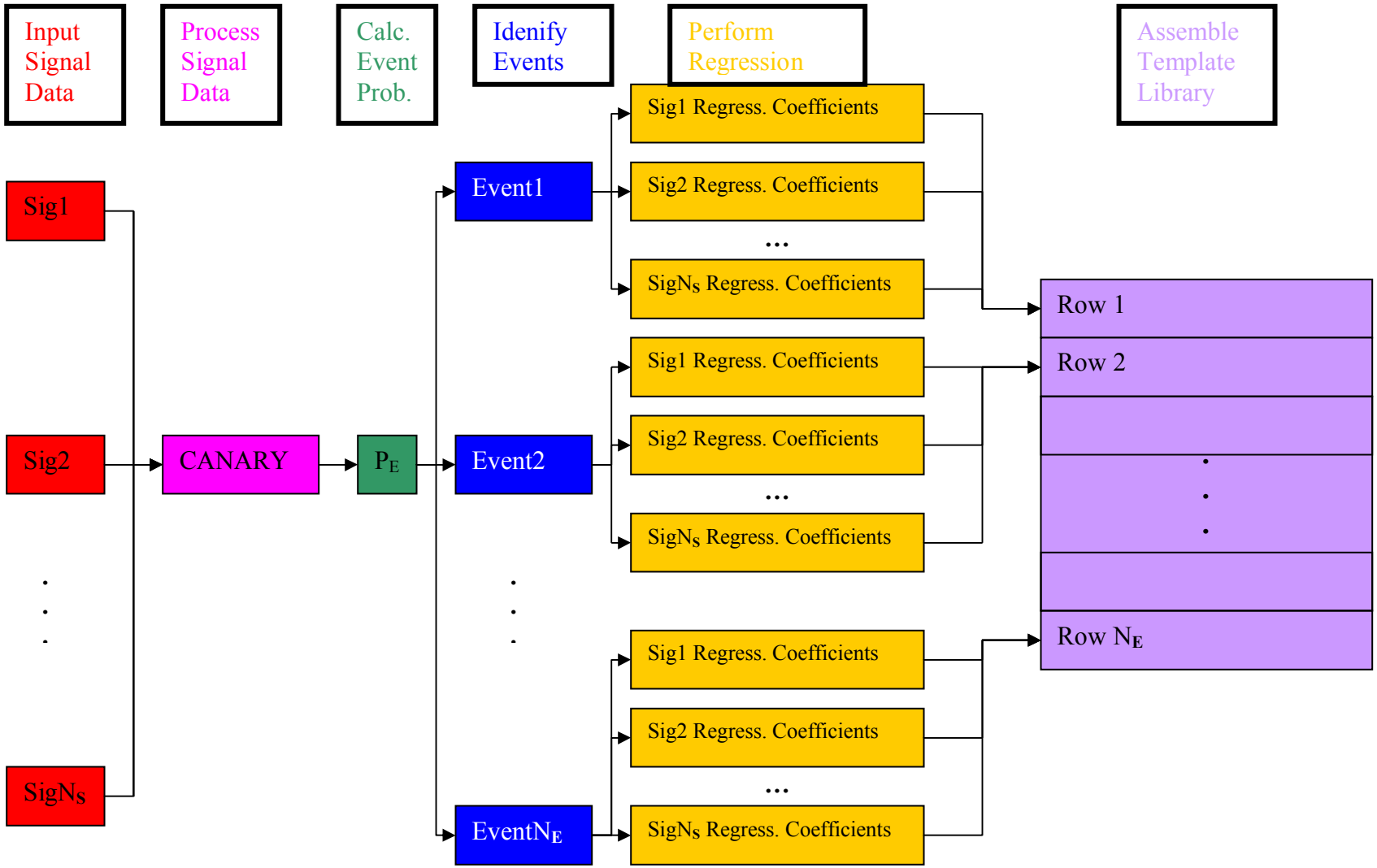

Figure 1. Flow Diagram for Creating the Water Quality Template Library.

1.3 Water Quality Change Clustering

Following the creation of the template library, the water quality change events are clustered. Recognizing that it is the pattern of the water quality change, not the actual water quality values during the change that must be identified, a trajectory clustering methodology has been implemented in CANARY. The algorithm simultaneously clusters the regression coefficients for all signals rather than the actual data values corresponding to the events. This section outlines the methods that CANARY uses to cluster the events in the water quality change template library.

CANARY uses the fuzzy c-means (FCM) algorithm to cluster the regression coefficients. The FCM algorithm is an iterative clustering algorithm developed by Dunn (1973) and further refined by Bezdek (1981). It is a "soft" clustering algorithm that permits events (or in this case, sets of regression coefficients) to belong to multiple clusters and, thus, differs from "hard" clustering techniques like the k-means algorithm (Hartigan and Wong 1978) that assigns events to a single cluster. For each event, the FCM algorithm calculates the degree to which each event belongs to each cluster.

The basis of the FCM algorithm is the minimization of the following objective function:

$$J = \sum_{i=1}^{N_E} \sum_{j=1}^{N_C} u_{ij}^m \left\| x_i - c_j \right\|^2 , \; 1 \le m < \infty ,$$

where

- N_E denotes the number of events being clustered;
- N_C denotes the number of clusters;
- x_i denotes the events that are being clustered;
- c_j denotes the cluster center for the j[th] cluster;
- u_{ij} is the degree of membership of x_i to cluster j. Note $0 \le u_{ij} \le 1$, and $\sum_{j=1}^{N_C} u_{ij} = 1$;
- $\| \; \|$ is a norm for measuring the distance of events from cluster centers; and
- m is a "fuzziness" parameter that can be adjusted to affect cluster membership. This parameter must be assigned a value greater than 1, and larger values lead to more overlap of the clusters.

As previously mentioned, the FCM algorithm is an iterative algorithm, and it is composed of the following steps:

1. Initialize the cluster membership matrix U^0, i.e., the matrix that contains u_{ij}.
2. At the k[th] step, calculate the cluster centers c_j^k using the cluster membership matrix U^K in the following equation

$$c_j^k = \frac{\sum_{i=1}^{N_E} \left(u_{ij}^k\right)^m x_i}{\sum_{i=1}^{N_E} \left(u_{ij}^k\right)^m}$$

3. Update the cluster membership matrix U^K with the following equation

$$u_{ij}^{k+1} = \frac{1}{\sum_{p=1}^{C_N} \left[\frac{\left\| x_i - c_j^k \right\|}{\left\| x_i - c_p^k \right\|} \right]^{\left(2/m-1\right)}}$$

4. Repeat steps 2 through 5 until $\left\| U^k - U^{k+1} \right\|_U < \varepsilon$ or $k > N_{term}$. The term ε is a positive constant used to establish convergence criteria for the FCM algorithm, and N_{term} is a positive integer that establishes additional termination criteria. The notation $\| \ \|_U$ is used to represent a matrix norm.

It is common practice to assign m a value of 2, and the CANARY implementation of the FCM algorithm follows this convention. Sensitivity analyses were conducted to determine other FCM parameter values. Table 2 lists parameter values that are assigned in CANARY's implementation of the FCM algorithm.

Table 2. Fuzzy C-Means clustering algorithm parameters in CANARY.

Parameter	m	ε	N_{term}	$\| \ \|_U$
Value	2	0.1	100	$\| \ \|_\infty$ for matrices

Several considerations had to be made when implementing the FCM algorithm in CANARY. The distance norm that was implemented is defined as follows:

$$\|v\| = \sqrt{\sum_{i=1}^{l} \left(\frac{v_i}{sd_i} \right)^2}$$

where

- l is the length of the vector;
- v_i denotes the ith element of the vector v;and
- sd_i denotes the standard deviation of all of the events' ith regression coefficients that are being clustered.

The norm is defined in this manner to equally weight the regression coefficients from all of the signals since the coefficients for all signals are clustered simultaneously. Often, specific conductivity values are 1 to 2 orders of magnitude larger than the other water quality signals, and if the standard Euclidian distance is used to define the norm in the

FCM algorithm, the clustering algorithm will more heavily weight the patterns in conductivity signals than patterns in the other signals. (The clustering methodology was tested on data in which residual chlorine values typically ranged between 1 and 3 mg/l, pH values varied between 7 and 9, and conductivity values were in ranges of 90 to 120 and 170 to 200 μs/cm.)

The FCM algorithm also requires an "initial guess" for the degree of cluster memberships (U^0 in Step 1 of the algorithm). It is common practice to assign random values to this matrix, but the efficiency of the algorithm may be sensitive to the initial guess. Thus, we implemented a different approach for assigning initial cluster membership values. To do this, the template library was initially clustered using MATLAB®'s hierarchical clustering function **`clusterdata`**. Hierarchical clustering is a "hard" clustering technique in which events are assigned to a single cluster. If an event was assigned to a particular cluster using the hierarchical clustering approach, the initial cluster membership degree for that event to the cluster was assigned a value of δ , and the degrees of membership for that event to all the other clusters were assigned a value equal to

$$\left(1-\delta\right) \Big/ \left(N_C - 1\right).$$

That is,

$$u_{ij} = \begin{cases} \delta, x_i \in \text{cluster j} \\ \left(\dfrac{1-\delta}{N_C - 1}\right), x_i \notin \text{cluster j} \end{cases}$$

The parameter δ is assigned a value of 0.8 in CANARY's FCM algorithm. This value was determined through trial and error.

Finally, the FCM algorithm requires that the analyst determine the number of clusters *a priori*. This can be difficult if the data are difficult to visualize or a large number of events are being clustered. At best, relying on the analyst's judgment is a subjective process. Thus, CANARY uses the PBM-index (Pakhira et al. 2003) to determine the optimal number of clusters. The PBM-index is defined as follows:

$$PBM\left(N_C\right) = \left(\frac{1}{N_C} \times \frac{E_1}{E_{N_C}} \times D_{N_C}\right)$$

where

- N_C is the number of clusters;
- $E_{N_C} = \sum_{i=1}^{N_C} E_i$;
- $E_i = \sum_{j=1}^{N_E} u_{ij} \left\| x_j - c_i \right\|, \ i = 1, N_C$; and
- $D_{N_C} = \max \left\| c_i - c_j \right\|$.

Note that x_i, c_j, u_{ij}, and $\| \ \|$ are defined in the same manner as they were in the FCM algorithm.

Pakhira et al. (2003) assert that the positive integer that maximizes the PBM-index is optimal in the sense that it minimizes the number of clusters while increasing compactness and separation between clusters. Hence, CANARY assigns the parameter representing the number of clusters in the FCM algorithm to the integer value between 2 and 10 (inclusive) that maximizes the PBM index. The upper bound on the number of clusters is arbitrarily set to 10 since most examples that have been analyzed optimize the PBM index with 3 to 6 clusters.

1.4 Calculating Cluster Statistics

In order to perform real-time comparison of water quality events with an existing template library, it is necessary calculate cluster statistics. We assume that the events in the clusters are normally distributed and use the following equations to calculate the cluster means and covariance matrices:

$$\mu_j = \frac{\sum_{i=1}^{N_E} \left(u_{ij}\right) x_i}{\sum_{i=1}^{N_E} u_{ij}}$$

$$COV_j = \frac{\sum_{i=1}^{N_E} \left(u_{ij}\right)\left(x_i - \mu_j\right)\left(x_i - \mu_j\right)^T}{\sum_{i=1}^{N_E} u_{ij}}$$

The subscript j denotes the cluster number.

2 Comparison of Incoming Data with the Template Library

Creation and clustering of a water quality change template library from historical data can be performed in an off-line mode. When CANARY has a real-time link to a SCADA water quality monitoring system, it can monitor incoming data and assess whether water quality changes are sufficiently similar to patterns already contained in the template library and, therefore, are unlikely to indicate serious problems (true alarms). Typically, water quality changes due to changes in operations of the utility are responsible for the most common patterns. Or, CANARY can assess if the changes are significantly different from the template library and merit further investigation. This section describes the process that CANARY uses to compare the real-time signals with the template library.

In its on-line mode, CANARY will be connected to a SCADA system that transmits water quality signals to the software. If the event probability calculated by CANARY exceeds the user-defined probability threshold described in Section 1.1, the software will perform polynomial regression fits to the same signals considered in the template library. This regression step must use all of the same parameters that were used to create the template library. The following calculations are then performed for each cluster:

$$xval_j = \left(x_{RT} - \mu_j\right)^T COV_j^{-1} \left(x_{RT} - \mu_j\right)$$

$$p_j = 1 - \left[\chi^2_{DOF}\right]^{-1} \left(xval_j\right)$$

where x_{RT} denotes the regression coefficients for the new event and $\left[\chi^2_{DOF}\right]^{-1}$ denotes the inverse cumulative distribution function (CDF) for the chi-squared distribution with degrees of freedom equal to the total number of regression coefficients. Under the assumption that the clusters are multivariate normal distributions, the term p_j denotes the percentile of each cluster's distribution to which x_{RT} corresponds. If no p_j values are less than a user defined tolerance level, a new cluster is added to the template library. This operation means that if the new event does not fall within a certain percentile of any cluster, then it is necessary to add a new cluster. The regression coefficients associated with the new event are the mean of the new cluster, and a user-specified covariance matrix is assigned to the cluster. If any p_j is less than the tolerance level, no new clusters are added to the template library. Rather, the regression coefficients corresponding to the new event are added to the library, and the FCM algorithm is re-run with on the entire supplemented library. Means and covariance matrices are then calculated for each cluster as described in Section 1.4.

3 Example Calculations

Several example calculations are provided in this section to demonstrate how the clustering and pattern matching tools within CANARY work. Data from the Chestnut II water works are used to demonstrate the pattern identification and matching process within CANARY.

Figure 2 shows the results of processing 6 months, January through June of 2008, of water quality data through the event detection process within CANARY. The basic output of CANARY is the probability of an event at each time step and this is shown in Figure 2. Parameters within CANARY control the calculation of this probability and for this analysis these parameters have been set to be conservative such that high probability of events occur frequently (Figure 2) leading to a large number of alarms.

Figure 3 shows the three surrogate parameters: total residual chlorine (TRC), pH and specific conductivity (CDTY) for the six-month period. The PUB threshold limits for the Chestnut II location are: TRC: 2.0 and 2.5 mg/l and pH 7.8 to 8.3. The major drops in TRC and associated rises in pH and CDTY are indicative of the periods when the Chestnut II plant is not producing output.

The red dots in Figure 3 indicate the time steps at which the probability of an event as calculated by CANARY first exceeds 0.75. There are a total of 59 red dots in Figure 3 and each red dot indicates a water quality event. The 90 time steps prior to the start of each event are then used in the pattern matching approach. An example of a single event and the patterns created by the data in the 90 time steps prior to the event are shown in the left side of Figure 4. These data are then fit with a relatively low order polynomial using a least squares regression model. In these examples a third-order polynomial was used. The regression models fit to the data are shown in the images on the right side of Figure 4.

Multivariate clustering as described above is applied to the coefficients of the regression models to classify the water quality data into groups of distinct patterns. The raw water quality data and the resulting regression models for four different sets of water quality patterns are shown in Figure 5 and Figure 6. The data and regression model patterns in these figures demonstrate several significant aspects of the pattern identification process. One of these aspects is that the regression models do not necessarily match every aspect of the raw water quality data. The regression models are smoothed representations of the water quality data. It is not necessary for the regression models to be completely accurate when compared to the observed data; it is only necessary to capture the differences in the water quality patterns in a consistent manner across all time steps in the data set.

Secondly, the absolute values of the water quality that end up in the same cluster can be quite variable. This result is due to the fractional degree of membership that is applied to each pattern in the cluster. Those water quality signal traces that are significantly

different from the other traces are still members of the same cluster, but will have a lower degree of membership than those traces near the mean trace of the cluster.

These results show an example of constructing water quality patterns from multivariate data using the Chestnut II water quality plant. We have examined water quality patterns at a number of monitoring stations within the PUB network and these results indicate that the most distinctive patterns occur at the outlets of the water works. The service reservoirs generally contain more stable water quality signals and have relatively fewer distinct patterns. Based on these results, we feel that the pattern matching algorithm will provide the most benefit to real-time water quality monitoring at the water works monitoring stations. However, additional data have been received from PUB that contain flow rate information for the service reservoirs and also water quality and flow rate information for the inlets to the service reservoirs. These data are being examined now and it is possible that distinct patterns in the service reservoir data may be found.

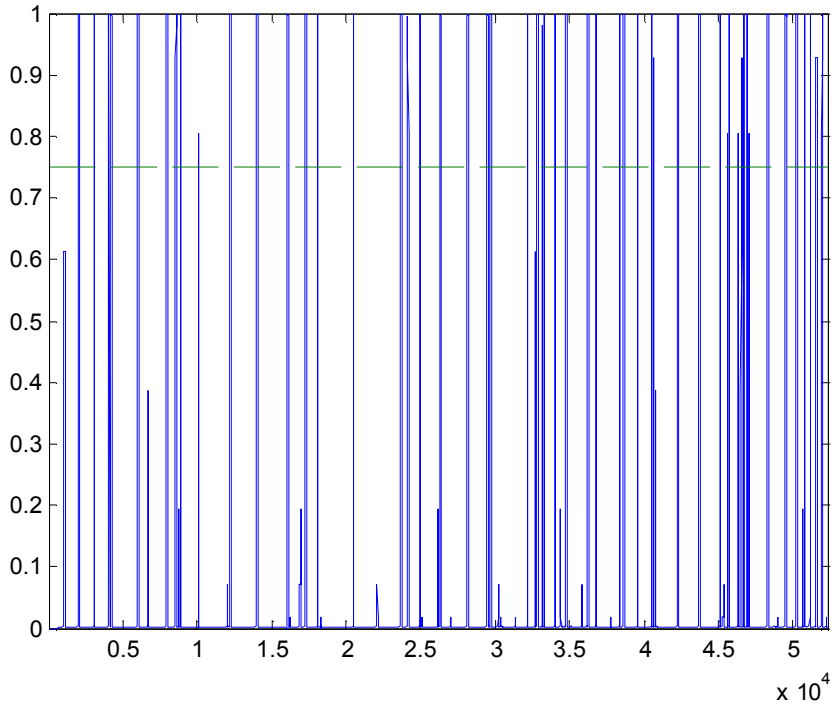

Figure 2: Probabilities from CANARY. When the blue line crosses the green probability threshold (0.75), an event is identified.

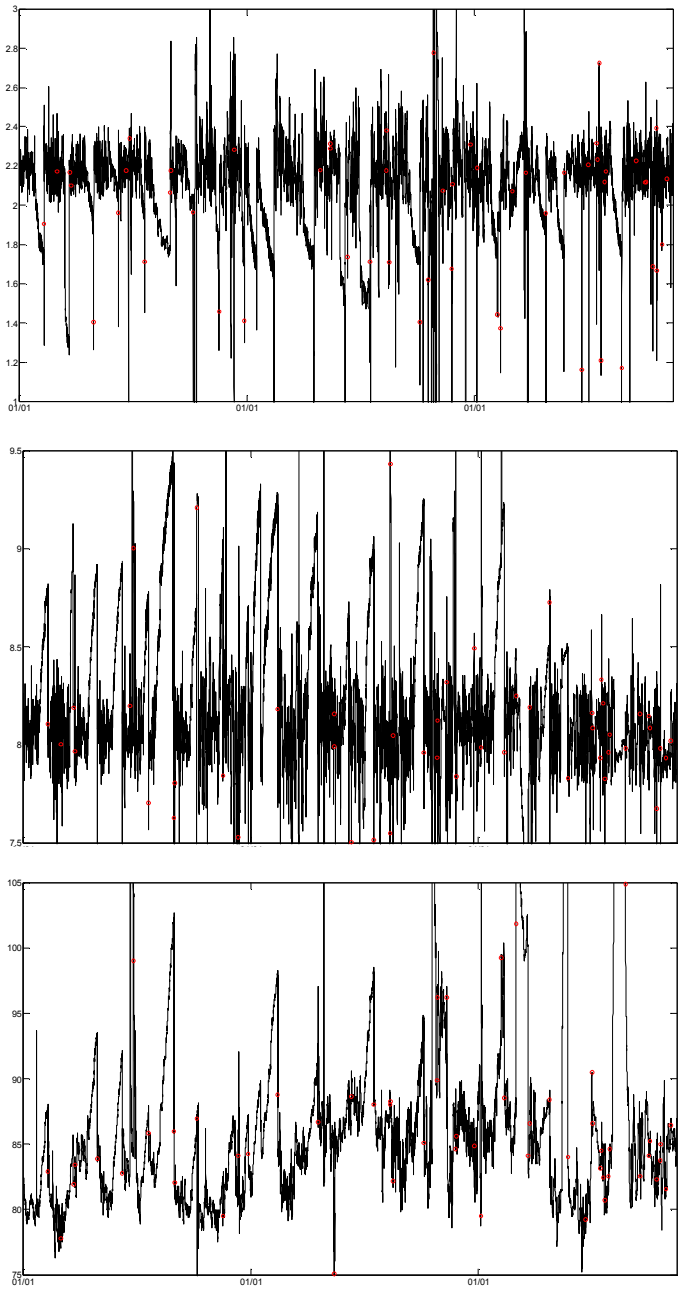

Figure 3. The timing of events that were identified in Figure 1 are used to identify signal data for analysis. TOP=Cl, MIDDLE= pH, BOTTOM= conductivity

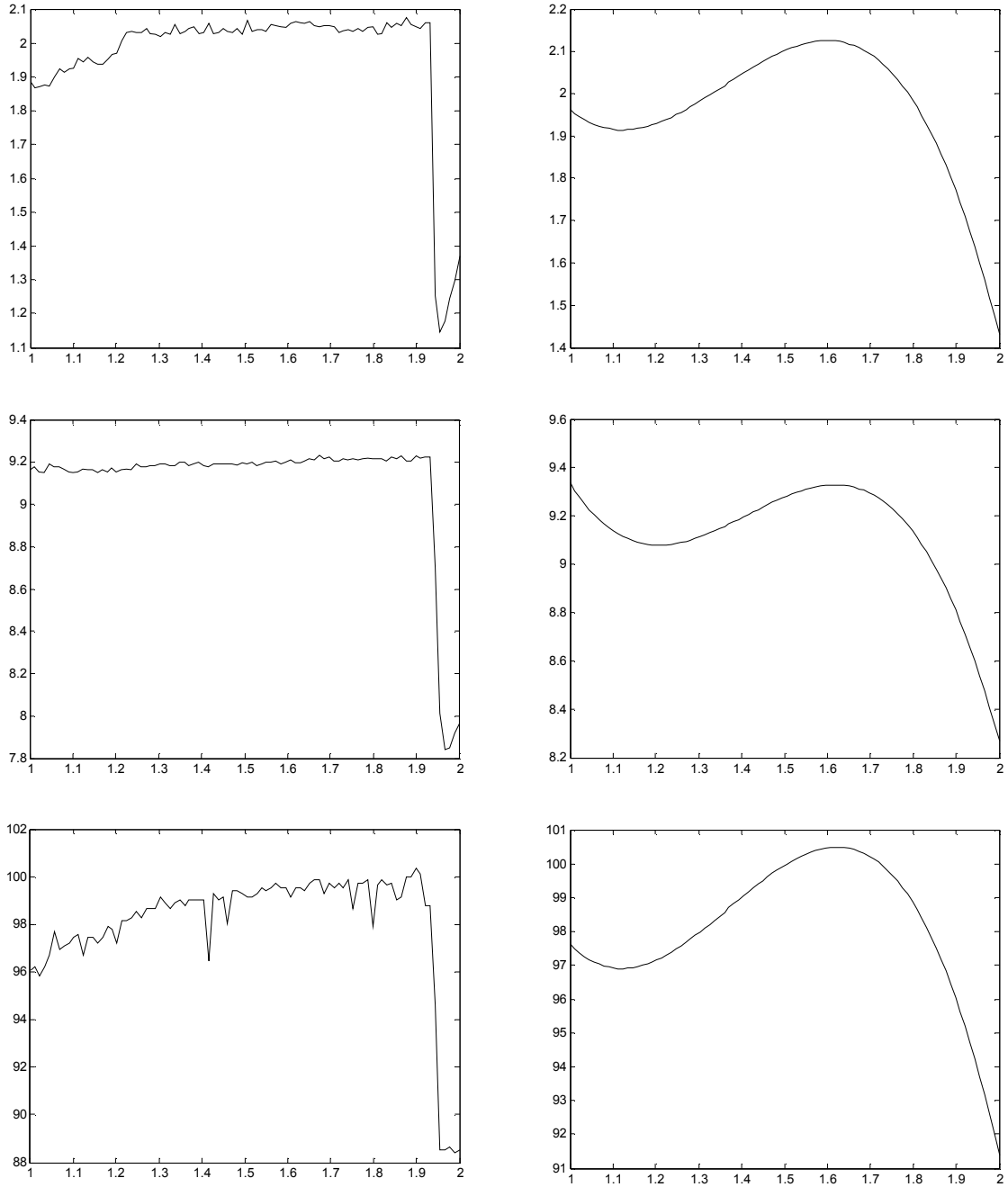

Figure 4. For each event, a regression polynomial is fit for each data signal. TOP=TRC, MIDDLE= pH, BOTTOM= CDTY.

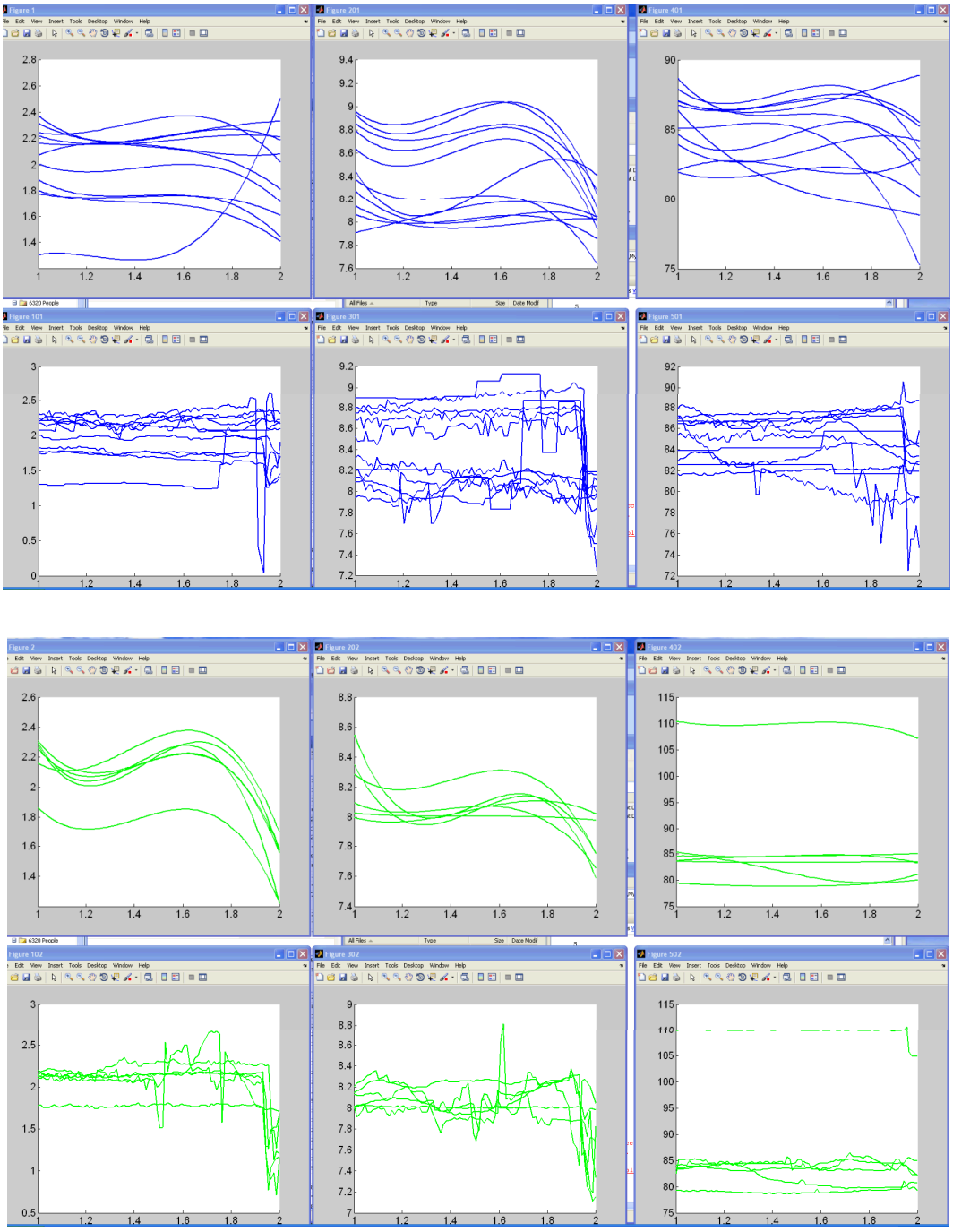

Figure 5. Two sets of water quality patterns. For each set, the upper images show the regression models fit to the water quality data and the lower images show the raw water quality data. TRC, ph and CDTY are shown from left to right.

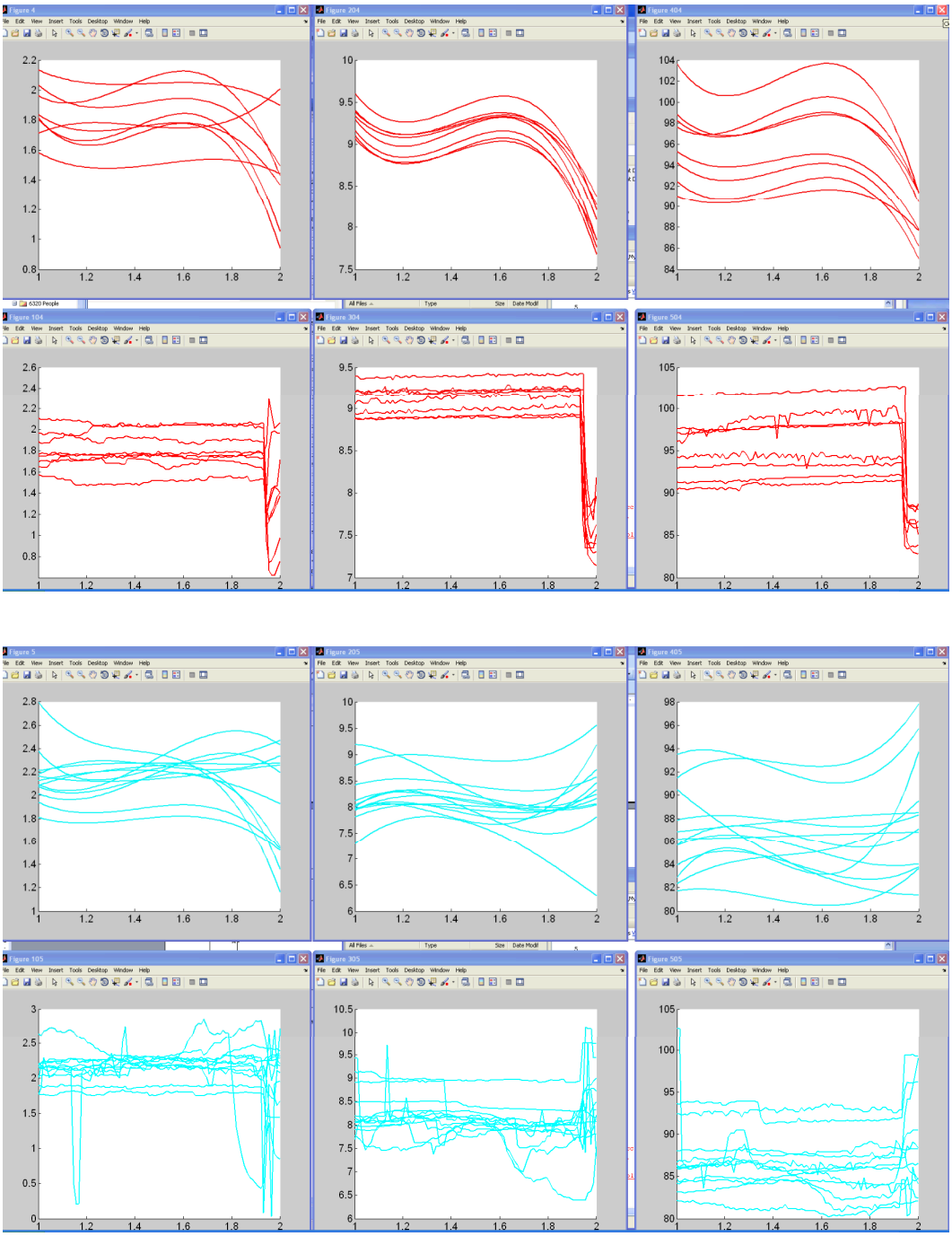

Figure 6. Two sets of water quality patterns. For each set, the upper images show the regression models fit to the water quality data and the lower images show the raw water quality data. TRC, ph and CDTY are shown from left to right.

4 References

J. C. Bezdek (1981). "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York.

J. C. Dunn (1973). "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics 3: 32-57.

Hart, D. B. and S. A. McKenna (2008). "User's Manual CANARY 3.0," U.S. Environmental Protection Agency, Washington, D. C.

Hartigan, J. A., and M. A. Wong (1978). "Algorithm AS 136: a K-means clustering algorithm." Applied Statistics, 28, 100–108.

Pakhira, M. K., S. Bandyopadhyay, and U. Maulik (2003). "Valididty Index for Crisp and Fuzzy Clusters," Pattern Recognition 37, 487-501.