Rare Event Detection Algorithm

- Applicability To Water Quality Monitoring

Michael J. Ungs

Objectives And Technical Approach

- SCADA systems record large amounts of water quality data and report anomalies to system operators
- The problem is to identify significant events requiring operator actions
- Review existing test methods
- Explore alternative approaches to enhance on-line event detection system performance

Standard Algorithms

Linear Prediction Filter

Filter and then perform linear trend analysis using previous window of data to predict next measurement. Difference between prediction and measurement is compared to a threshold value.

Multivariate Nearest Neighbor

Map previous window of data into m-dimensional multivariate space. Plot newest measurement and compute distance to nearest historical cluster. Distance to nearest cluster is compared to a threshold value.

Set-Point Proximity

Provide a ramped warning that indicates the probability of an event occurring as the water quality signal approaches different set point limits.

Water Quality Pattern Matching

A library of low-order polynomials are constructed from events the user deems representative of common water quality patterns for each parameter. Recent data is compared to the pattern library and checked for a match.

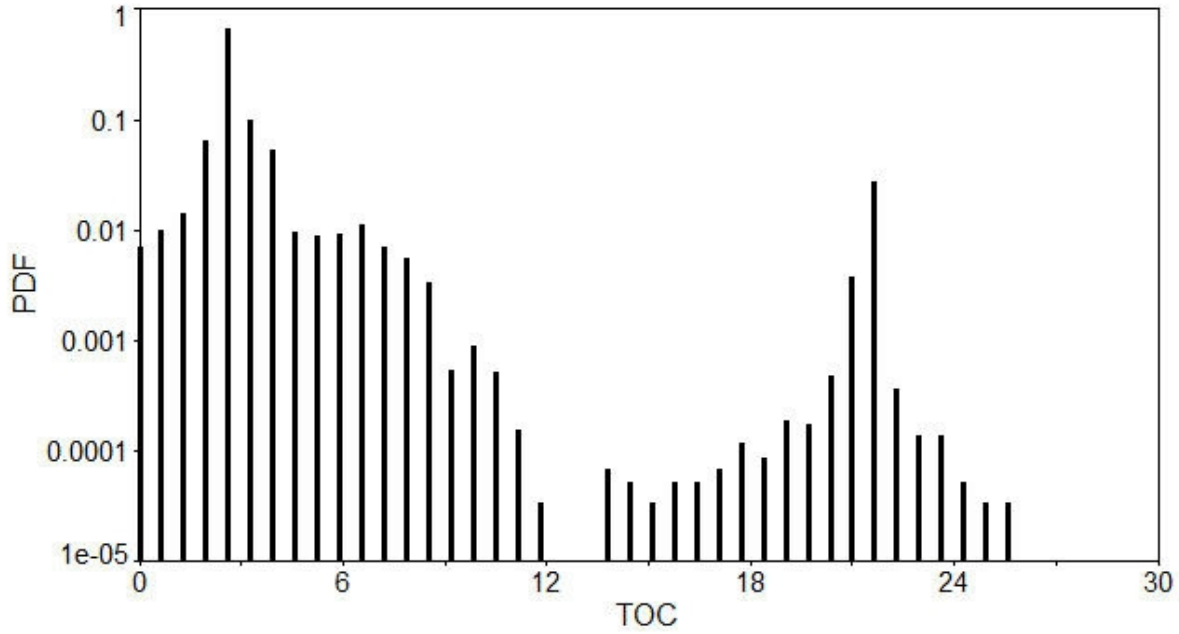
Nature Of The Recorded Data And Limitations Of Standard Algorithms

Observed Data	Standard Assumption
Multiple magnitudes of scale	Bell shaped distribution of values (Gaussian)
Multiple frequencies superimposed	Linear variation over time at a fixed frequency
Random initiation of scale & frequency changes	Patterns assume a fixed sequence of values over time

Standard Assumption Is A Gaussian Distribution

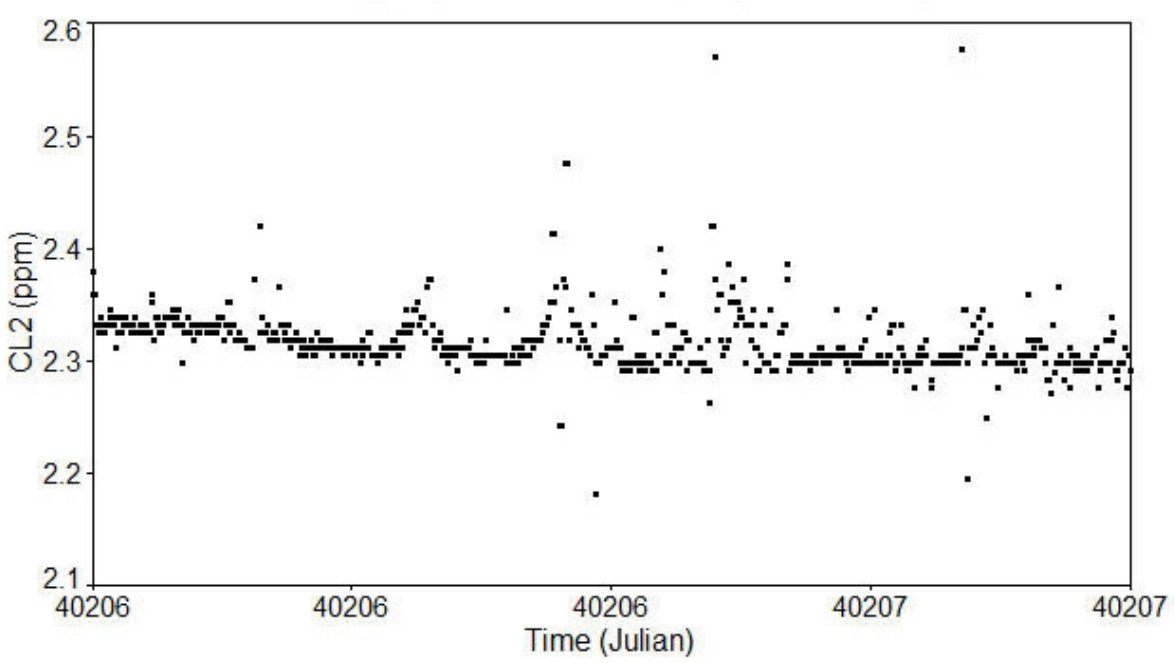
Actual Data Contains Multiple Magnitudes Of Scale

For Example- A Bivariate Distribution Of Total Organic Carbon



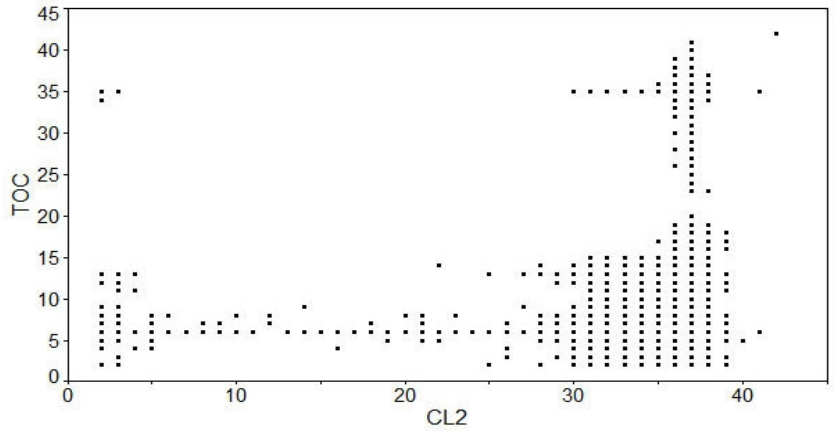
Standard Assumption Is A Linear Variation Over Time At A Fixed Frequency

Actual Data Contains Multiple Scales & Harmonics In A Data Set



Additional Attributes Observed

- **Bivariate and trivariate frequencies**
- **Many combinations of values never occur**
- **Correlation matrix is sparsely populated, as seen in this NxM matrix**



Summary Of Data Characteristics And Prior Assumptions

- Plot of parameter value as function of time shows multiple harmonics in the data
- Plot of the probability density function (PDF) of data shows multiple modes in data
- Basic assumption of central tendency is invalid
- Many combinations of data values for two or more parameters are not observed

Desired Features of New Algorithm

- **Ability to detect covariate relationships between two or more parameters, as shown in the EPA contaminate-spike formulas**
- **Simultaneously control the number of both false positives and false negatives**
- **Need to custom fit algorithm to each site using an automatic fitting routine**

Algorithm To Create Correlation Matrix For Two Parameters

- Pick historical record of approximately 10,000 values
- Find minimum and maximum value for each parameter
- Subdivide ranges of two parameters into N and M intervals, creating a NxM matrix
- Place each data pair into the matrix, counting the number of values in each cell

False Positive Count

- Take an additional 10,000 data values
- Find the cell of the p-dimensional correlation matrix that corresponds to the new data values
- Count as a false positive value if there are J or less historical values in that cell: increment as FalsePos +1
- Repeat for all 10,000 data pairs

False Negative Count

- Take the same 10,000 data values
- Add a contaminate spike to data values using Eddies Contaminate Library with a specified level of dose
- Compare spiked values to correlation matrix. Count as a false negative if there are K or more historical values in that cell: increment as FalseNeg +1
- Repeat for all 10,000 data pairs

Testing Different Combinations

- Repeat the above process for a different number of cells for each parameter range, varying cell count between 4 and 15
- Repeat all of the above processes for a different combination of two or more water quality parameters
- The Rare Event Detection Algorithm uses that combination with the smallest sum of FalsePos + FalseNeg

Using Algorithm

- **The NxM matrix and paired parameters with the smallest FalsePos+FalseNeg count is periodically re-calculated off-line (tuned every 3-5 months)**
- **Each new data value is compared to the NxM matrix. An event flag is triggered if any new data pair falls into a cell with less than J historical values**

2-D Analysis Of Rare Event Detection Routine With First Data Set

Spike with Dose = 2	Best Parameter Pair	Number of False Positives	Number of False Negatives
A	TOC & CL2	16	16
B	Temp & Cond	712	7187
C	Temp & TOC	626	5415
D	Temp & pH	715	6879
E	TOC & CL2	15	4
F	pH & CL2	9	74
G	Temp & TOC	626	5414
H	Cond & CL2	0	0
I	TOC & CL2	12	1

2-D Analysis Of Rare Event Detection Routine With Second Data Set

Spike with Dose = 2	Best Parameter Pair	Number of False Positives	Number of False Negatives
A	TOC & CL2	5	1
B	Temp & Cond	4583	4585
C	TOC & ORP	0	0
D	Cond & pH	3381	3287
E	TOC & ORP	0	0
F	TOC & ORP	0	0
G	TOC & ORP	7	25
H	TOC & ORP	0	0
I	TOC & ORP	0	0

Summary

Rare Event Detection Algortihm works extremely well when:

- **All water quality parameters are actually measured**
- **The magnitude of the contaminate spike is sufficiently different from background**
- **Routine has been tuned to a site-specific spike**