Detection and Classification of Anomalous Events in Water Quality Datasets Within a Smart City-Smart Bay Project

Dian Zhang*†, Timothy Sullivan†, Ciprian Briciu-Burghina†, Kevin Murphy†, Kevin McGuinness*,
Noel E. O'Connor*, Alan Smeaton*, and Fiona Regan†
*CLARITY: Centre for Sensor Web Technologies, Dublin City University, Glasnevin, Dublin 9
† MESTECH: Marine and Environmental Sensing Technology Hub, Dublin City University, Glasnevin, Dublin 9
Email: dian.zhang2@mail.dcu.ie, tim.sullivan@dcu.ie
ciprian.briciu@dcu.ie, kevin.murphy@dcu.ie, kevin.mcguinness@dcu.ie
noel.oconnor@dcu.ie,alan.smeaton@dcu.ie,fiona.regan@dcu.ie

***Abstract*—Continual measurement is key to understanding sudden and gradual changes in chemical and biological quality of water, and for taking reactive remedial action in the case of contamination. Monitoring of water bodies will increase in future within Europe to comply with legislative requirements such as the Water Framework Directive and globally owing to pressure from climate change. Establishing high quality long-term monitoring programs is regarded as essential if the implementation of pertinent legislation is to be successful. However, conventional discrete sampling programs and laboratory-based analysis techniques can be costly, and are unlikely to provide timely and reliable estimates of true ranges of deterministic physicochemical variability in a water body with marked temporal or spatial variability. Only continual or near continual measurements have the capacity to detect ephemeral or sporadic events, thus providing the potential for on-line event detection and classification. The aim of this work is to demonstrate the potential role of continuous data acquisition in decision support as part of a smart city project. In this work, a multi-modal smart sensor network system framework for large scale estuarine and marine water quality monitoring is proposed. The application of a number of evolving techniques that allow automated detection and clustering of events from data generated by *in situ* sensors within environmental time series datasets is described. We provide examples of how change in the range of variables such as turbidity and salinity might be detected and clustered to provide the end user with greater ability to detect the onset of environmentally significant events. Finally, we discuss the acquisition of data from *in situ* water quality sensors and its utility within the framework a smart city-smart bay integrated project.**

Keywords*–Continuous water monitoring estuary, marine, decision support, turbidity, salinity, anomaly detection, robust online clustering, pixel-based adaptive segmentation.

I. Introduction

Automated collection and storage of datasets related to environmental water quality is now becoming commonplace, however, challenges remain in automated detection of important events within these datasets and thus determination of the value and ecological significance of collected data for use in decision support systems [1]. This challenge can only increase as the vision of futuristic smart cities containing integrated sensing networks becomes a reality. *In situ* sensors capable of continually sampling chemical and physical parameters offer the potential to reduce costs, provide timely information and improved representation of long-term trends in the fluctuations of pollutant concentrations [2]. *In situ* in the context of environmental sensing means in place or in direct contact with the medium of interest, as opposed to methods such as remote sensing where no contact is made between the sensor and the analyte. Indeed, the ideal aquatic monitoring system of the near future might consist of a network of sensors deployed at key locations, capable of autonomous operation in the field for a year or more [3][4]. Despite the increasing range of techniques available, continuous on-line *in situ*, measurement systems remain largely limited by both poor sensor performance and a lag in availability and application of suitable data analytics. Thus, while measurement and detection methods exist for many environmental pollutants in the laboratory, continuous monitoring on a cost effective basis in the field remains a challenge for these reasons.

A. The Ideal System

The ideal monitoring system of the near future might consist of sensor networks deployed at key locations, capable of near autonomous operation in the field over long time frames (annual to decade time scales). The components necessary to achieve this measurement of multiple parameters, simultaneously and in real-time are available [5]. However, it is clear that as a scientific community, we need to improve the quality and reduce the cost of sensors for many of the desired parameters (for example most nutrients, microbial contaminants), while using simplified devices in robust embedded networks to make this ideal truly achievable. Another consideration is that a common platform for data validation and sensor verification has yet to be universally implemented to improve data quality. Data collected from monitoring stations can be communicated by wireless technology prior to statistical processing and interpretation by expert systems. Indeed wireless data transmission, and the concept of wirelessly networked sensors in particular, has however become one of the most dynamic and important areas of multi-disciplinary research [6][7]. Real time alerts can be raised to relevant personnel, perhaps through an alarm sent to smartphones or e-mail, when trends for any constituent of interest breaches particular thresholds (for example Environmental Quality Standards (EQS)) are detected. Notified personnel can then intercept serious pollution incidents or lead an appropriate response. Detected individual outliers can be combined into event-based information to support the identification of impacts from environmental threats. Events can be further clustered into groups based on some kind of similarity matrix to assist scientists in identifying commonality between groups.

B. The Future: improved signal processing including both on-line and off-line data analytics

Implementation of advanced and user-friendly event-detection software to distinguish between normal conditions and anomalous events is critical if data provided by the ideal sensing system is to be used effectively. Data-driven estimation models with sequential probability updating have been suggested for this purpose [8] and implemented in various forms (see for example CANARY [9][10][11]). Detection of water security threats arising either from intentional or unintentional sudden contamination events implemented in Contamination Warning Systems (CWS) is of particular importance, and it has been suggested that in the region of thirty-three contaminants (pesticides, insecticides, metals, bacteria, etc.) can be utilised as indicators of intentional water contamination [12]. While widespread detection of all thirty-three variables using low-cost autonomous sensor systems is far from achievable at present, the potential advantages of automated early warning systems based on multivariate analysis of datasets collected by such systems are clear. Moreover, considerable expenditure of research effort is required to further develop improved data analytics platforms for the ideal sensing system, including forecasting, modelling and event detection platforms or Early Warning Systems (EWS) [13]. The rapid growth of "big data" provided by social media, concomitant with improved computing capacity has spurred research interest into novel data analytics techniques. Datasets provided by *in situ* sensing in its current form are at present not approaching the scale of those provided by social networking scenarios, however, as the vision of internet-scale sensing heralded by the development of improved sensing becomes reality, such datasets may become widespread. Detection of anomalous events within these datasets would be of widespread interest and a number of Artificial Intelligence (AI) methods, such as artificial neural networks (ANN) and support vector machines (SVM), have already been utilised for this purpose (see for example [14][15][16] and [17] for a critical review of ANN usage in this regard). Generally, these techniques have been used to classify water quality data into normal and anomalous classes after supervised learning training. Other data-mining methods, such as K-means classification and the multivariate nearest-neighbour (MV-NN) algorithms, combining different water-quality parameters and location information, are also used for protecting drinking water systems [18]. Data-fusion methods have been used to combine various types of information, for example, operational data or data from multiple monitoring stations [19] or sensors [20] to improve the detection of water-contamination events while reducing the potential number of false positives. Other approaches have proposed combining residuals for water-quality parameters with autoregressive (AR) models or other methods (see for example [20][21]). An extension of this is to use some form of pattern recognition and matching to detect and and create a multivariate library of known events. A newly detected event can then be matched to the library of historical events to determine if similar events have occurred previously. If so, information gained from historical events can be used to analyse the causes and impacts of current ongoing events. Event clustering would be a typical approach to this problem and has been implemented in several such systems (for example CANARY uses a trajectory clustering-based pattern matching approach). However, existing systems such as CANARY, are focusing on contamination event detection for drinking water systems which are very different from marine or estuarine environments. Drinking water systems normally have a closed supply chain and dare not affected by many factors such as weather, tide, season, dam release, etc., in contrast to open water bodies. A key challenge for on-line automated analysis of environmental datasets lies in dealing with the peculiarities of these datasets themselves, in which missing values are common, disjointed measurement methodologies and techniques are followed or in which large-scale uncertainty can exist due to sensor performance issues. Successful development of useful early warning systems and other on-line data analytics methods for *in situ* sensing platforms must be capable of dealing with these issues. Particular issues of current early warning systems include high proportion of false alarms and false negatives in practical applications, unacceptable computational demands and lack of on-line detection [22].

This work outlines the potential for continuous water quality monitoring in decision support as part of a Smart Bay component [23][24] in the broader context of a connected Smart City project in Dublin. Over the coming years, the SmartBay project will see the expansion of a multi-modal sensor and data network in Dublin Bay for detection of pollution and flood events among others. The latter will consist of a number of sensor deployments, including visual sensing systems, modelling and integration of additional available data sources (for example data provided by citizen monitoring). Datasets collected over the course of the SmartBay project can be utilised for other applications depending on user requirements or emerging applications, with particular emphasis on water in the city, port and coastal areas. In this paper, real data collected from pilot sites in Dublin Bay using continuous autonomous multi-parameter sensing systems are used to demonstrate how machine-learning techniques such as robust online clustering (ROC) and a modified pixel-based adaptive segmentation (MoPBAS) approach can be utilised for such purposes. These techniques will be discussed in terms of anomaly detection, event construction and classification, and the resulting opportunities for development of decision support systems. We show how use of a multi-modal data system demonstrates potential for low-cost sensing in complex aquatic environments such as estuaries. Data from water quality sensors are evaluated and analysed along with data from grab samples, with the latter supporting the observations of trends from water quality monitoring systems. Scenarios presented provide examples of the potential value of such a monitoring system in building a SmartBay infrastructure. The rest of the paper is organised as follows. Section II presents the proposed multi-modal smart sensor network framework for marine environmental monitoring. Section III introduces the pilot site and the deployment of instruments. The implementation of anomaly detection, abnormal event construction and clustering is described in Section IV and the experiment results are shown in Section V. The conclusion of the paper is in Section VI.

II. Multi-Modal Abnormal Event Detection and Classification Framework for Smart-Bay Monitoring System

In order to fully utilise the above ideal system, a novel multi-modal smart sensor network framework for marine environmental monitoring is proposed. Figure 1 illustrates an overview structure of the system. The framework consists of three layers, a wireless sensor network (WSN) layer, backend smart system layer and a decision-making layer. The WSN layer contains observation sites that are equipped with various numbers and type of sensors. The smart system layer has two main components. The data repository collects data generated from all sensors and creates a truly multi modality data pool. The smart system processes these data sets using state-of-the-art machine learning techniques to convert raw sensor measurements into organised knowledge that can be understood by operators. The operators can then make decisions based on this information to avoid or reduce negative impacts. Moreover, the operators can then send feedback to the WSN to indicate whether the current deployed sensor network is sufficient to monitor and subsequently model the observation sites. The overall architecture of the deployments made around Dublin bay is shown schematically in Figure 1.

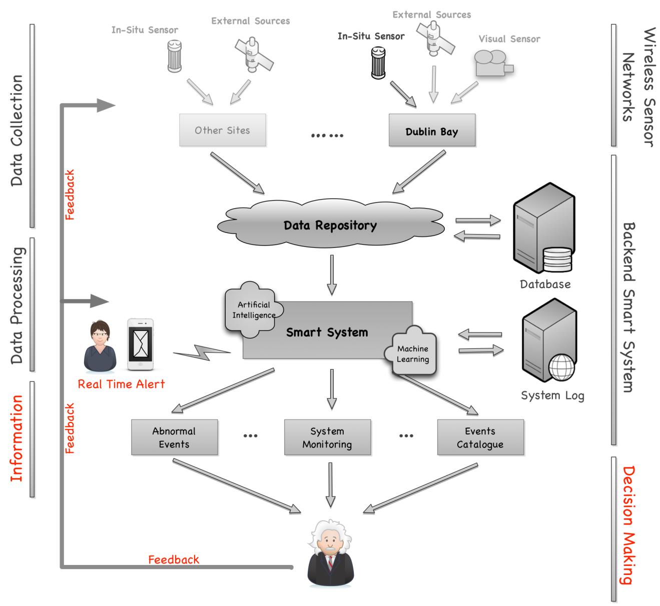

Figure 1. A schematic outlining the architecture of the proposed multi-modal smart monitoring system, including treatment of the data and the feedback mechanisms for decision-making.

This architecture is flexible and extendable allowing other sites and data sources to be added without overly increasing complexity. The data from *in situ* sensors at Dublin Bay can be complemented with data from external sources, for example weather forecasts, whereas other sensing modalities (such as visual sensing) can also provide information on human activities, allowing distinction between anthropogenic and natural re-suspension events. The system can also send real time alerts to operators, so that they can react quickly to avoid or limit negative impacts. In terms of data analysis, the system can be formed as data collection, data process and information stages. The data collection process involves deploying and maintaining the wireless sensor networks, where the data processing level converts raw data into information easily interoperated by operators. The information stage maintains a large indexed content based archive, which allows the user to browse and query events.

III. Monitoring Location and Pilot System Deployed

The following describes Dublin Bay, the site used as the location for this study, along with the equipment used for collection of continuous monitoring data at the site.

A. Test Site

Dublin Bay (latitude: 53°20'39", longitude: -6°12'59") is located on the lower Liffey Estuary Dublin Ireland in a busy port environment (see Figure 2). The estuary is a diverse ecosystem with many micro-environments that include benthic communities, fish and shellfish, seabird populations and marine mammals [25][26]. The area is also a zone of passage for salmon and sea trout migrating to and from feeding and spawning areas [27]. The topography of the estuary has been greatly modified, and is constrained by walls along its whole length and is regularly dredged to remove accumulated sediments. The working site is located in the upper part of the Estuary, where the ship traffic is less intensive. Average water depth in the area is approximately $8m$ and the width of the channel is approximately $260m$. Due to the large amount of activity at the site and its importance from an environmental and ecological perspective, the site was equipped with a multi-parameter *in situ* sensor along with a visual sensing system.

Figure 2. Overview of the Dublin Bay area, indicating the location of the deployed pilot system , which provided the datasets used in this work. Dublin Bay image source: Google Maps. Retrieved: 2014-04-11

B. Instrumentation

A multi-parameter sonde (YSI 6600EDS V2-2), equipped to measure turbidity (Nephelometric Turbidity Units(NTU)), optical dissolved oxygen ($mgL^{-1}/\%$saturation), temperature (°C), conductivity ($mScm^{-1}$), depth (m) and telemetry system (EcoNet) was purchased from YSI Hydrodata UK. The sonde was deployed at a depth of $2.5m$ from the water surface, and data was collected since 1st of Oct 2010 with a sampling interval of $15mins$. Temperature, dissolved oxygen and salinity were checked using a ProPlus handheld multi-parameter instrument (YSI Hydrodata UK) and turbidity was validated using a portable turbidity meter Turb® 430 IR (VWR Ireland). Both hand held instruments were calibrated in the laboratory prior to site visits as per manufacturer's protocols. Site visits were undertaken fortnightly in winter and weekly in spring. Copper tape and mechanical wipers (for the optical oxygen and turbidity sensors) were used to control biofouling of sensor systems.

IV. Methodology

To detect and cluster environmental events, anomalous sensor readings (also referred as outliers) need to be extracted from a continuous data stream. These abnormal sensor measurements are then grouped into events based on proximity in time. A set of features is extracted that is characteristic of different anomalies and is used to assign individual events. Each event might have different temporal characteristics; so to compare their similarities, a bag-of-words approach is adopted to encode these features as constant length descriptors. Each feature set of the detected anomalies is matched against a predefined codebook and the closest matching codeword is used to represent the feature. The event is then represented by the frequency of occurrence of each word. Once the feature vector of the event is constructed, a clustering method is applied to group these events into subclasses based on their similarities. Figure 3 shows the flow diagram of the proposed framework. Each step of the proposed framework is introduced in detail as follows.

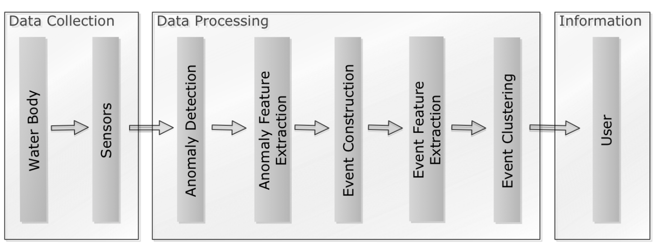

Figure 3. Flow diagram of the proposed system framework

A. Anomaly Detection

To detect abnormal events, we first need to detect unusual sensor measurements in the data stream. An unusual or anomalous sensor measurement, is defined as a sensor reading that differs considerably from recent observations. Thus, an anomaly can be detected by modelling previous sensor measurement trends. To achieve this, we have modified the pixel-based adaptive segmenter (MoPBAS) method originally proposed by Martin Hofmann et al. for image segmentation [28]. A non parametric water quality background trend model is built based on a history of recently observed sensor readings. The classification of an unusual reading depends on a decision threshold, which is adapted based on the variations in the data stream. The model is updated over time according to the dynamics of the measurements. In the following, we describe the process by which the MoPBAS method is used to detect abnormal sensor readings.

1) Background Trend Model and Anomaly Classification: To classify a new incoming value $I(t)$, a sensor reading trend model $B(t)$ is built. $B(t)$ is defined by an array of N recently observed values.

$$B(t) = \{B_1(t), ..., B_k(t), B_N(t)\} \tag{1}$$

In [28], incoming values are classified based on the total number of distances between input value $I(t)$ and all elements in $B(t)$ that are smaller than threshold $T(t)$. We found that just comparing the minimum distance with the threshold is sufficient to differentiate the measurements.

$$I(t) = \begin{cases} 1, & \text{if } \min(\text{ dist}(I(t), B_k(t))) > T(t) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

If the input value is classified as normal ($I(t) = 0$), it can be used for updating the background trend model. The update probability depends on the learning rate $L(t)$.

2) Update of the Decision Threshold: When monitoring water quality of estuarine waters, there can be periods of time where large variations occur in measured variables, such as after heavy rainfall, and time periods with little change or fluctuation. Ideally, for periods of high variability, the threshold $T(t)$ should be increased and for stable conditions, $T(t)$ should be decreased. To quantify this dynamic, the mean $\overline{d}_{min}(t)$ of the previous N minimum distances between input values and trend model are calculated as the measure of the trend variations. For instance, assuming the water quality measurements remain constant, $\overline{d}_{min}(t)$ will be zero. In contrast, $\overline{d}_{min}(t)$ will be higher for more dynamic backgrounds. The decision threshold can then be adapted as follows:

$$T(t) = \begin{cases} T(t) \times (1 - T_{inc/dec}), & \text{if } T(t) > \overline{d}_{min}(t) \times T_{scale} \\ T(t) \times (1 + T_{inc/dec}), & \text{otherwise} \end{cases} \tag{3}$$

where $T_{inc/dec}$ is a static value that controls the threshold update rate and T_{scale} is also a fixed parameter, which stretches $\overline{d}_{min}(t)$ to the same range as $T(t)$. T_{lower} and T_{upper}, which are also fixed values, control the upper and lower bounds of the threshold, thus the threshold will not grow out of range.

3) Update of the Learning Rate: Another important parameter of MoPBAS is the trend model learning rate L_t. Water quality measurements have characteristics that are significantly different from image segmentation data. Values measured by *in situ* sensors are typically very noisy, have lower sampling rates (in terms of minutes compares to fraction of a second in the image processing domain) and vary from a baseline (they change gradually due to "global" effects, such as wind, tide etc.). Unlike background modelling in the image processing domain, in which foreground objects will be slowly merged into the background if it no longer moves, water quality parameters will usually return to a baseline level after an event. Thus, we normalise ($R(t)/R_{upper}$) and invert the learning rate proposed in the original PBAS method. Here, the learning rate is defined as follows:

$$R(t) = \begin{cases} R(t) + \frac{L_{inc}}{\overline{d}_{min}(t))}, & \text{if anomaly = true} \\ R(t) - \frac{L_{dec}}{\overline{d}_{min}(t))}, & \text{if anomaly = false} \end{cases} \tag{4}$$

$$L(t) = 1 - R(t)/R_{upper} \tag{5}$$

Where L_{inc} and L_{dec} are fixed values that control the increasing and decreasing intervals. The variation in $R(t)$ is limited by an upper and lower bound: $R_{lower} < R(t) < R_{upper}$. The learning rate also depends on the background dynamics ($\overline{d}_{min}(t)$). When an event occurs, measured values provided by the sensor will usually deviate greatly from the baseline level. Thus, the trend model should be updated slowly or not updated at all. In contrast, after an event occurs, sensor readings will usually stabilise or return to the baseline, and the trend model should be updated quickly. When an anomaly is first detected ($\overline{d}_{min}(t)$ is small), $R(t)$ increases rapidly, thus the learning rate $L(t)$ decreases sharply. However, $\overline{d}_{min}(t)$ will become large quickly when multiple anomalous readings are detected, which results in $R(t)$ and indeed $L(t)$ remaining constant or only changing slightly. When sensor readings stabilise or return to a normal range, $\overline{d}_{min}(t)$ becomes small and $L(t)$ will increase.

4) Update of the Trend Model: Updating the trend model, B, is essential to capture global effects, such as tide or wind. The learning rate $L(t)$ is used as the update probability and an element in the trend model is randomly chosen and replaced by the incoming value. However, this process is only performed when no anomalous values are detected. This allows the incoming sensor measurement to be "learned" and incorporated into the trend model. In the original PBAS, a randomly chosen neighbouring pixel is also updated, however, as there is no "neighbour" (image data is 2D as opposed to 1D water quality data) and this step is not performed.

5) Distance Calculation: Rather than using common distance metrics, such as Euclidean distance, we use the root of the absolute square difference (RASD) to calculate the distance between incoming value and the ith element in the trend model.

$$D_i(t) = \sqrt{|I(t)^2 - B_i(t)^2|} \tag{6}$$

Figure 4 shows the ratio between our distance metric and the 1-D Euclidean distance (for illustration purposes, the input $I(t)$ range is set from 5 to 104 in steps of 1, background $B_i(t)$ is set to 5). It can be seen from the graph that when the distance is large, the output is approximately equal to the 1-D Euclidean distance. However, the output is enhanced when the different between $I(t)$ and $B_i(t)$ is small. This is a key factor when calculating the background dynamic $\overline{d}_{min}(t)$, as it smooths the effect of an event to $\overline{d}_{min}(t)$. Thus, the value of $\overline{d}_{min}(t)$ will not increase rapidly when an event occurs as shown.

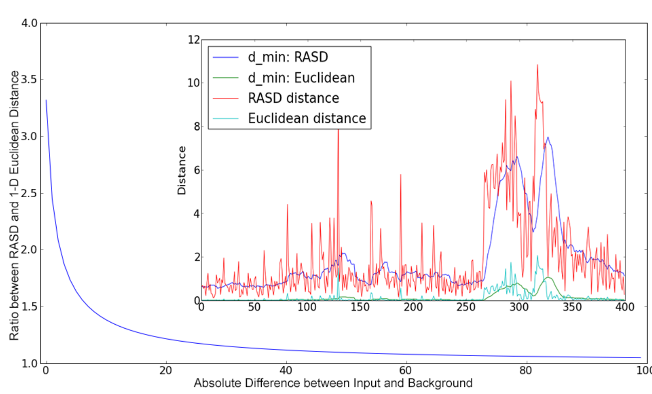

Figure 4. Demonstration of the ratio between RASD distance metrics and 1-D Euclidean distance, the inner graph shows RASD distance method, which enhances small distances, smoothing the variation of the background dynamics $\overline{d}_{min}(t)$.

B. Anomalous Feature Extraction

To capture the similarity in anomalies detected, and for further clustering of anomalous events, we need to extract a set of features that are sufficiently discriminative to allow us to classify unusual readings and subsequent events. The feature set of an anomalous reading has the following components: the difference between the previous sensor measurement $I(t-1)$ and current sensor measurement $I(t)$, current sensor measurement $I(t)$, the difference between current sensor measurement $I(t)$ and the next sensor measurement $I(t+1)$, the minimum distance between sensor measurement and trend model d_{min}, and the distance between the minimum distance d_{min} and the threshold T(t). The feature set $f(anomaly)$ can be represented as:

$$f = [I(t-1) - I(t), I(t), I(t) - I(t+1), d_{min}, d_{min} - T(t)]$$

C. Event Constructing

Anomalies detected by the MoPBAS method are grouped into events according to their temporal information. To achieve this, agglomerative hierarchical clustering is applied. As shown in Figure 5, consecutive anomalies are combined together into a single event. Furthermore, if the gap between a new anomaly and previous outlier is smaller than a threshold, T_{gap}, the new anomalous value will be merged into the same event. In contrast, if this gap is greater than T_{gap}, a new event will be created.

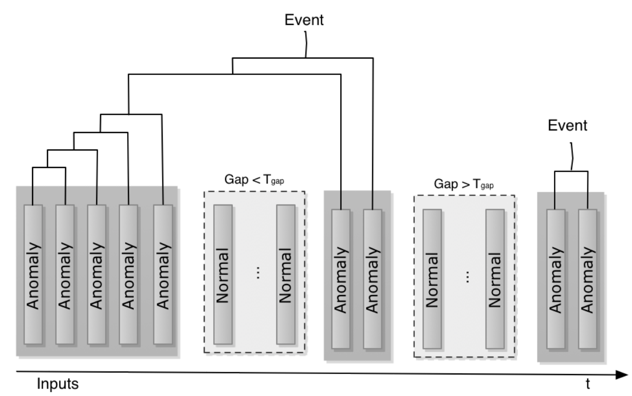

Figure 5. Anomalies are grouped into events using agglomerative hierarchical clustering based on their temporal information

D. Event Clustering

A Bag-of-Words approach is widely used in text document classification [29], content-based image retrieval [30] and image recognition tasks [31], where a document is represented as a bag of its "words" or a bag of small image patches (visual words) in the image processing domain. Most classification or clustering methods require a fixed number of feature dimensions. However, for many tasks, such as text document indexing, the number of features extracted from each file are generally different. The Bag-of-Words method represents these features by counting the frequency of occurrence of each "word" as the descriptor of the object. For text document processing, a "word" generally means an entry in a "codebook", which is the combination of a single word in a dictionary or a phrase. In the image processing domain, a word (some times referred as a "visual word") means a small image patch or fragment. As each environmental event may contain a different number of anomalous values, each outlier feature set is represented by a "sensor word" in order to quantify the similarities between events, and the frequency of their occurrence is reconstructed as the descriptor of the event. To create a codebook, K-means clustering is performed over a set of training data. The centres of the learned clusters are then defined as codewords. Each anomaly feature set in an event is mapped to a certain codeword in the codebook and the event can be represented by the histogram of the occurrence of the codewords.

To divide events into groups, a clustering method known as robust on-line clustering [32] is used. Clustering is the process of dividing instances into groups in such a way that instances in the same group are more similar than elements in other groups. There are many common clustering methods that are widely used such as K-Means or Mean-shift. Current research indicates that there is no known single clustering method that categorically out performs all others in all tasks. The benefit of using robust on-line clustering in this context is that, unlike K-Mean or Mean-Shift, this method is not sensitive to "noisy" data. This is a key requirement for environmental monitoring tasks where highly variable data could indicate a significant event. Moreover, robust on-line clustering is an on-line method that can be used to process a continuous data stream provided by *in situ* sensors.

V. Experiments and Results

This section describes the experiments that carried out to evaluate the proposed system. The initial value of all parameters in the proposed approach are also listed in this section. As a proof of concept, we are focusing on salinity and turbidity measurements, however, the same framework may be applicable to other water quality parameters that have similar time series characteristics.

A. Test Data

The dataset that is used for evaluating the proposed system was collected from deployed remote water quality monitoring systems in Dublin Bay between 1st Oct 2010 and 3rd May 2011 with a total number of 20529 measurements. The data exhibits a wide variety of environmental events that include short-term events such as rainfall as well as long term changes in measurement related to seasonal effects.

B. Parameter Settings

MoPBAS methods consist of a multitude of tuneable parameters, which can be used to control the sensitivity of the anomaly detection process. To obtain an optimized set of parameters for MoPBAS, the standards training, evaluation and testing procedure needs to be carried out. However, a fully annotated dataset, which is required for this process, is not available at the time of this paper is written. Thus, the initial parameter values used in these experiments are set based on the nature of the environment and the knowledge gained from on-site observation and site surveys. In our experiments, the following values were used for salinity anomaly detection:

- $N = 24$: N is the number of elements of the trend model B. Increasing N will reduce the sensitivity of the system as there is high probability that there might be an element that is close to the incoming sensor reading. However, only the normal values will be pushed into the trend model, thus further increases in N only duplicates existing elements (elements in the trend model are similar to each other). N is set to 24 (6 hours with 15 minutes sampling interval) in the following experiments. This is based on the average duration of change from high to low water or visa versa.
- $T_{inc/dec} = 0.02$: The step of the threshold T increases or decreases. Detection performance is not very sensitive to this value and this value is increased if the data exhibit a high degree of variability . This value depends on three main factors, the duration of an event, sampling rate and how fast sensor readings stabilise after an event. The number of $T_{inc/dec}$ should allow an increase of T from minimum to maximum longer than events and roughly the same length as the time required for stabilisation. Setting $T_{inc/dec}$ to 0.02 gives 70 steps from T_{upper} to T_{lower} giving approximately a 17 hour stabilisation period.
- $T_{upper} = 12$: The upper bound of the decision threshold. Increasing this value will reduce the sensitivity of anomaly detection, i.e., only large variations will be classified as anomalies. This value depends on the average of sensor measurements at the site and how the user defines an outlier. At our estuarine pilot site, the mean salinity value is 30.2 ppt over the length of the test dataset described. Readings are generally stable and we define any sudden changes with a magnitude greater than 2.5 ppt as an outlier. Thus, by mapping these values to the distance calculation function, T_{upper} can be calculated:

  $$(T_{upper} \approx \sqrt{|(Sal_{average} \pm 2.5)^2 - Sal_{average}^2|}).$$
- $T_{lower} = 3$: The lower bound of the decision threshold. Reducing this value will increase the sensitivity of anomaly detection, smaller changes will be classified as an anomaly. Similar to T_{upper}, any salinity changes less than ± 0.15 ppt are considered as noise. Thus T_{lower} can be calculated from the same equation above.
- $T_{scale} = 3$: This is the equilibrium factor, which stretches $\bar{d}_{min}(t)$ to the same range as the threshold. Figure 6 shows the distribution of salinity background dynamics $\bar{d}_{min}(t)$. It appears that most of the $\bar{d}_{min}(t)$ values are less than 1 (this variation is generally attributed to sensor measurement error). Thus, to scale the $\bar{d}_{min}(t)$ to the same range of T, T_{scale} is set to 3 $(T_{lower}/1)$
- $L_{inc} = 5$: This is the trend model learning rate control parameter R increasing interval. Figure 6 and Figure 7 show that most of the $\bar{d}_{min}(t)$ values of salinity and turbidity are smaller than 2 and 3, respectively, thus,

we set this value to 5, which is large enough to give a rapid decrease in the learning rate when an significant event occurs.

- $L_{dec} = 0.1$: This is the trend model learning rate control parameter R decreasing interval. The value taken depends on the distribution of the background trend dynamic. Thus, the $\overline{d}_{min}(t)$ varies between 0 and 3 when no events are happening. The chosen value of 0.1 results in a smooth increase in the trend model updating probability.
- $R_{upper} = 3$: The upper bound of learning rate control parameter R. The value taken approximately equals the ratio between L_{inc} and the majority of $\overline{d}_{min}(t)$ values.
- $R_{lower} = 0.1$: The lower bound of the learning rate control parameter R. This takes the form of a small positive number to avoid negative probability. The ratio of R_{upper} and R_{lower} defines how fast the learning rate increases. For example, if $\overline{d}_{min}(t)$ is set to a constant, the learning rate will reach a maximum after 30 samples.

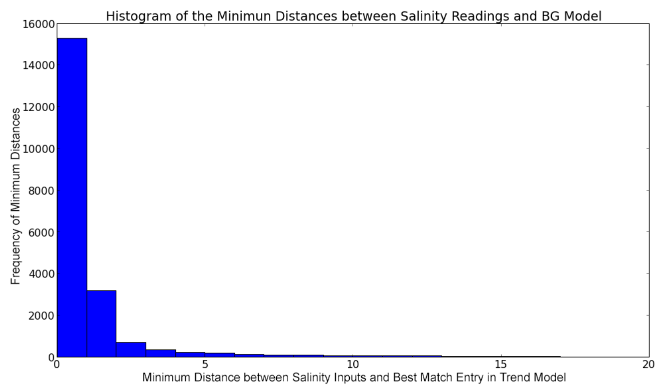

Figure 6. Histogram of salinity background dynamics

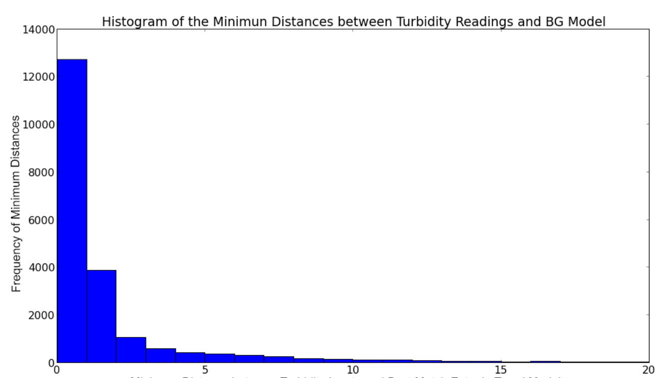

Figure 7. Histogram of turbidity background dynamics

As turbidity readings have different ranges and dynamics when compared with salinity measurements, the definition of a turbidity event are different. Some parameters need to be adjusted. The following parameter values were changed for turbidity anomaly detection: $T_{inc/dec} = 0.05$, $T_{scale} = 4.5$, $T_{upper} = 5$ and $T_{lower} = 1$. $T_{inc/dec}$ is modified because the stabilisation period of turbidity is generally shorter than salinity, which means that the decision threshold for turbidity needs to be updated faster. T_{upper} and T_{lower} are calculated using the same function. For turbidity, if the change is greater than $5NTU$, it will be considered as an anomaly and less than $1NTU$ will be considered as noise. However, the model updating control parameters, L_{inc}, L_{dec}, R_{upper} and R_{lower} remain the same. The number of elements of the trend model B is also set to 24.

For the event construction and clustering purposes, the parameters are the same for both salinity and turbidity. The number of words in the codebook is set to 50. When constructing an event, T_{gap} is set to 1 to avoid noise. This means that two anomalies are merged into the same event if the gap between them is smaller than 2 samples.

There are a number of events that may cause rapid changes in sensor readings, for example rainfall events, flood events, shipping or contamination events. In the present work, the number of clusters is set to 14, which is chosen as it represents approximately twice the number of known events that may occur at the testing site. However, the number of the cluster centres is application dependant. Increasing this number will further spilt the cluster into smaller sub groups.

C. Salinity Experiment Results

Applying the described MoPBAS anomaly detection to our test dataset resulted in 947 out of 20529 measurements being detected as anomalies. Figure 8 shows a 10-day window of the anomaly detection results. The red dots indicate salinity anomalies detected, while the blue line is the sensor measurements and the green solid line is the closest matching entry in the background trend model. As illustrated in Figure 8, most of the abnormal salinity readings are detected accurately.

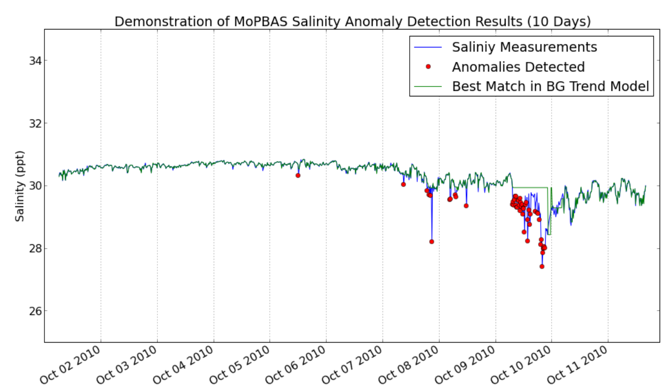

Figure 8. A 10-day window of the MoPBAS salinity anomaly detection results.

Figure 9 demonstrates adaptation of the detection threshold and background learning rate based on variation in the mean minimum distance ($\overline{d}_{min}$) between sensor measurements and background trend model. The red line at the bottom represents the background learning ratio. The decision threshold is shown in blue and the minimum distance between sensor readings and the best match entry in the model is shown in green.

In order to cluster events into groups based on their similarity, detected anomalies are merged into events based on

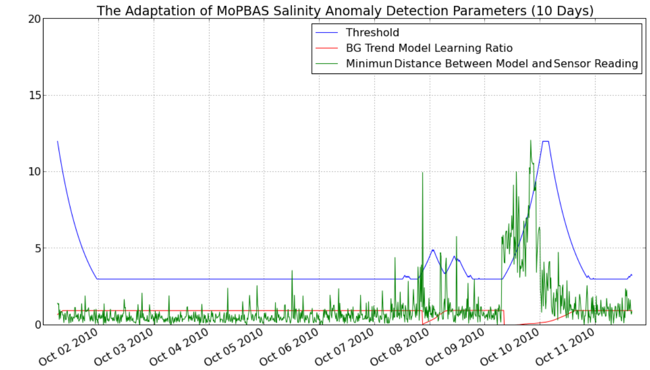

Figure 9. Detection threshold, background trend model learning rate and minimum distance between input value and best match element in model.

timestamps. From the 947 outliers detected, 261 events were constructed in the test datasets using the MoPBAS method. For each salinity anomaly detected, a set of features is extracted as the feature vector of the sample. To normalise these features, a 50-word codebook is created. Due to the limited dataset available, the code book is built using all anomalies. However, when more data is collected the codebook can be reused and does not need to be rebuilt unless the setup of the deployed system is modified or the monitoring site is greatly changed. Thus, each anomalous value is normalised as a "word" and the histogram of the occurrence of each word for each event constructed is used as the feature set of the event.

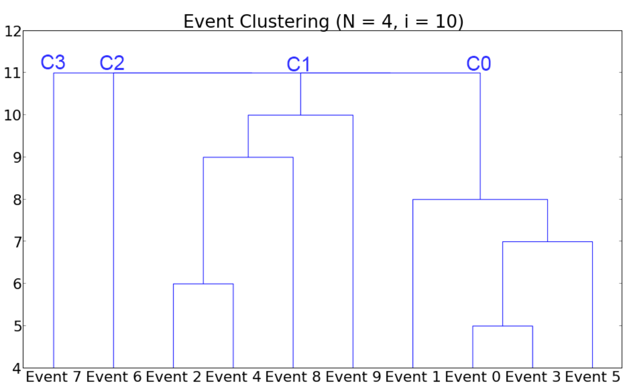

Figure 10. Example of how robust on-line clustering performs, note that for illustration purposes the number of cluster centres is set to 4 (N = 4) and the first 10 events (i = 10) are used.

TABLE I. CLUSTERING RESULTS, SHOWING THE NUMBER OF SIMILAR EVENTS WITHIN EACH CLUSTER GROUP.

Clusters	Number of events	Events
Clusters 0	243	Event 0, 1, 19, 22 etc.
Clusters 1	4	Event 57, 90, 197, 211 etc.
Clusters 2	2	Event 7, 101
Clusters 3	2	Event 204, 256
Clusters 4	2	Event 132, 142
Clusters 5	1	Event 10
Clusters 6	1	Event 34
Clusters 7	1	Event 43
Clusters 8	1	Event 62
Clusters 9	1	Event 80
Clusters 10	1	Event 91
Clusters 11	1	Event 157
Clusters 12	1	Event 173
Clusters 13	1	Event 180

Figure 10 illustrates how similar events are merged into the same cluster. At step 4 ($i = 4$), the cluster buffer is full. At step 5, the two most similar events, event 0 and 3 are merged. Event 2 and event 4 are grouped together when event 5 occurs. As can be seen from the graph, when event 9 occurs it is assigned to cluster 1, other events in cluster 1 (events 2, 4, 8) are the similar events to event 9, which happened in the past.

After applying the described clustering methods to the whole dataset, a total of 14 clusters are created. The number of events in each cluster is shown in Table I. Cluster 0 contains the most number of events. Cluster 1 consists of 4 similar events and there are 2 events in cluster 2, 3, and 4. Cluster 5 to cluster 13 only have 1 event in each.

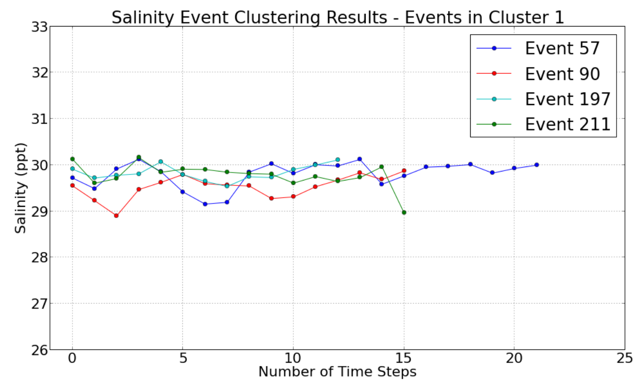

Figure 11. Plot of the salinity measurements of the four events in cluster 1.

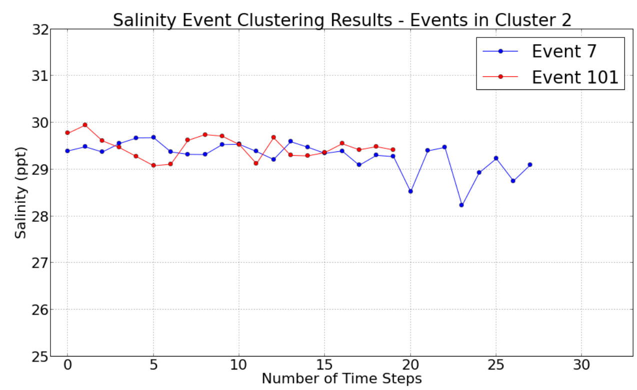

Figure 12. Cluster 2 consists of two events (event 7 and event 101).

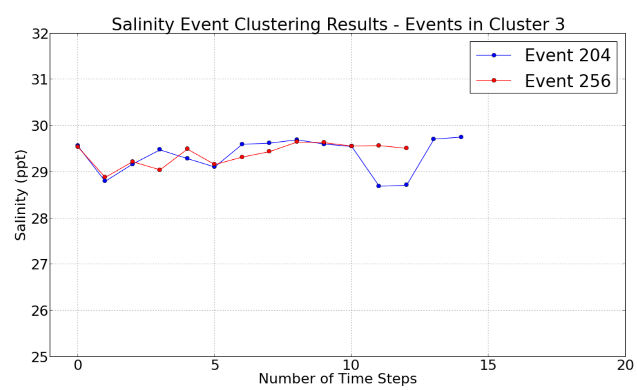

Figure 13. Two events (event 204 and event 256) in cluster 3.

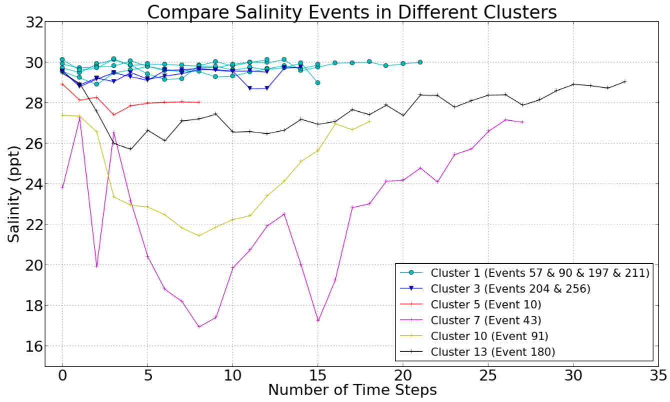

Figure 14. Comparison of salinity events in different clusters.

Figure 11 demonstrates the four events in cluster 1, the results show that events are similar to each other within the cluster. Figure 12 shows all the events in cluster 2 where it can be see that two events do have similar variations. Figure 13 demonstrates events 204 and 256 contained within cluster 3. It can be seen that the two salinity events have a very similar trend until the last two measurements, where event 256 has a small concave but event 204 remain flat. Figure 14 illustrates the difference between events in different clusters. As can be seen from the graph, events within the same cluster have similar trends but events in different clusters have very different profiles.

D. Turbidity Experiments Results

Applying the same procedures to turbidity data, 2096 sensor measurements are classified as anomalies. Figure 15 demonstrates a 10-day subset of turbidity anomaly detection results. The red dots are the turbidity anomalies detected, blue line is the sensor measurements and the green solid line is the closest matching entry in the background trend model. As illustrated in Figure 15, the majority of the abnormal turbidity readings are detected accurately.

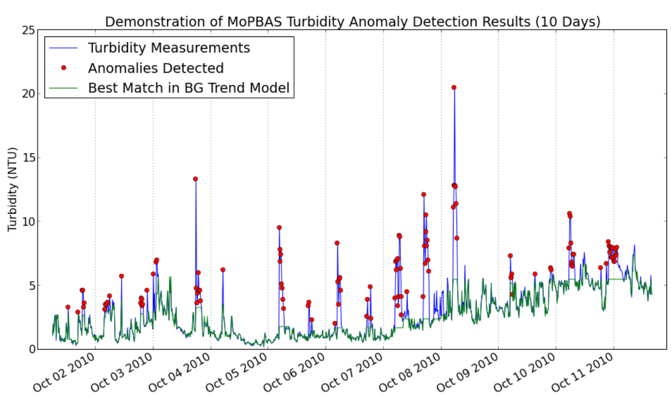

Figure 15. A 10-day window of the MoPBAS turbidity anomaly Detection Results.

Figure 16 demonstrates adaptation of the detection threshold and background learning rate based on variation in the mean minimum distance ($\bar{d}_{min}$) between sensor measurements and background trend model. As with detection of anomalies in the salinity dataset, the classification threshold increases when readings become highly variable and decreases when measurements do not change rapidly. In contrast, the model learning rate decreases sharply when events happen and increases slowly when sensor readings are stable.

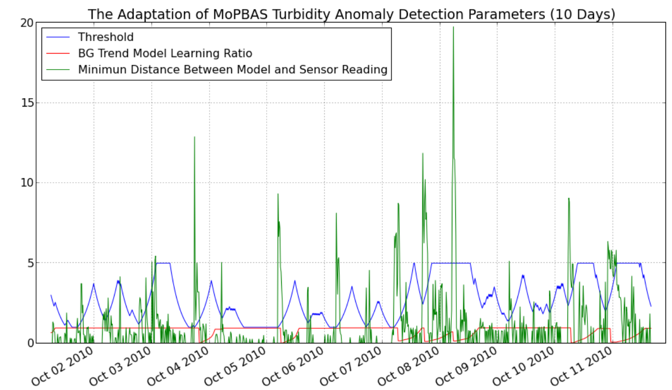

Figure 16. Detection threshold, background trend model learning rate and minimum distance between input value and best matching element in the background model.

Turbidity anomalies are grouped into events according to their timestamps. For the whole dataset, 707 events are constructed from a total number of 2096 classified anomalies. Table II lists the clustering results and the turbidity events in each cluster.

TABLE II. RESULTS OF TURBIDITY EVENT CLUSTERING, SHOWING THE NUMBER OF SIMILAR EVENTS IN EACH CLUSTER GROUP.

Clusters	Number of Turbidity Events	Events
Clusters 0	691	Event 0, 1, 10, 21, 152 etc.
Clusters 1	4	Event 71, 644, 647, 697
Clusters 2	2	Event 124, 253
Clusters 3	1	Event 29
Clusters 4	1	Event 99
Clusters 5	1	Event 102
Clusters 6	1	Event 252
Clusters 7	1	Event 570
Clusters 8	1	Event 606
Clusters 9	1	Event 608
Clusters 10	1	Event 610
Clusters 11	1	Event 705
Clusters 12	1	Event 706
Clusters 13	1	Event 707

Figures 17 and 18 show all events in the corresponding cluster where it can be seen that the events within the same cluster have similar variations. Three out of four events in cluster 1 have a spike shape at the beginning and then settle down. Although, the event 697 does not have a spike shape at the start but its overall trend is very similar to the settle down period of the rest of the events in the cluster. As can be seen in Figure 18, the two events are different in length. This shows the advantage of bag-of-words approach, which encode anomaly features as constant length descriptors. Figures 19 and 20 show that there is only 1 event in each cluster. However, as illustrated in the figures, the two events have unique trends.

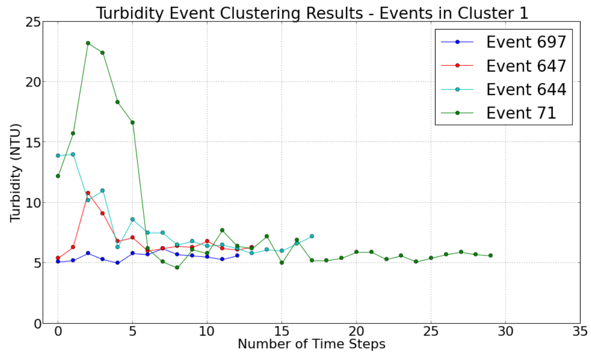

Figure 17. Plot of the turbidity measurements arising from events classified as being in cluster 1.

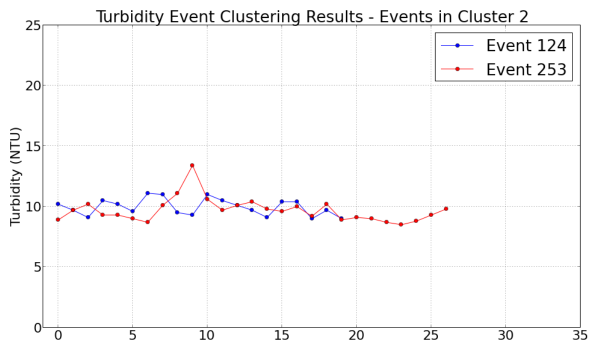

Figure 18. Plot of the turbidity measurements from events in cluster 2.

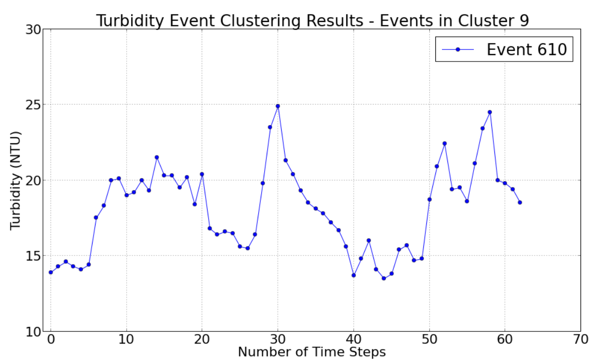

Figure 19. Plot of the turbidity measurements from events in cluster 9.

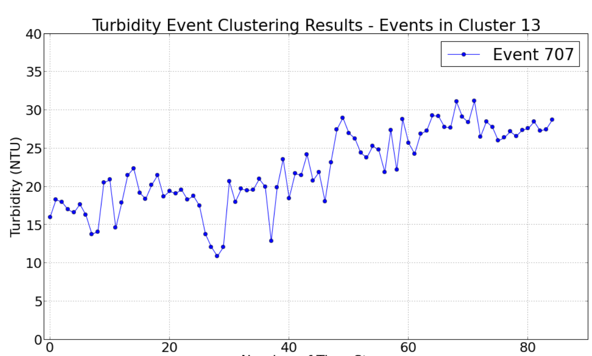

Figure 20. Plot of the turbidity measurements from events in cluster 13.

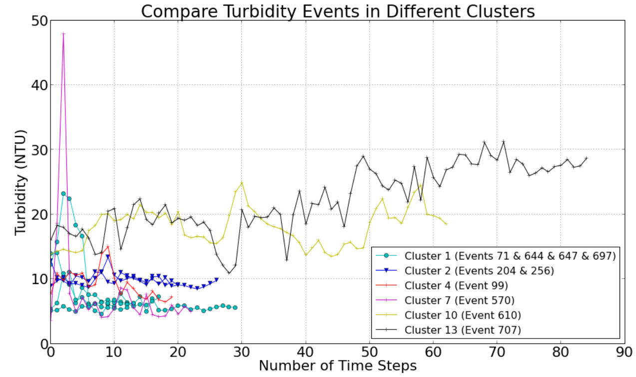

Figure 21. Illustration of the differences in turbidity readings between assigned clusters.

Figure 21 illustrates the difference between events in different clusters. As can be seen from the graph, events within the same cluster have similar trends but events in different clusters have very different shapes.

E. Discussion

We have shown, using time series of both salinity and turbidity, that the MoPBAS method is suitable for detection of anomalous sensor readings. Updating the background trend model provides the capacity to process both highly variable data and gradual changes such as tide or seasonal effects. The dynamic threshold and model update rate are appropriate for detection of environmental events in estuaries. As can be seen from Figures 9 and 16, the classification threshold is increased when an anomaly is detected. This is due to the fact that after an event occurs, there is usually a period of time where the water body settles down. The *in situ* sensor measurements is likely return to a similar trend or slightly offset compare to sensor readings prior the occurrence of the event. During this period, the sensor readings are typically highly variable and alter in step changes rather than monotonic decreases. The rapid increasing of the threshold during a event and slowly decreasing after the event can reduce false positives. Another advantage of use of this adaptive threshold is that the system only detects large variations during periods of high fluctuation while small changes will be captured during periods of relative stability. In contrast, the background learning rate remains high during stable periods and decreases rapidly when an anomaly is detected. This is because the background model should simulate the trend of the water quality parameters. However, as the threshold is raised the input is likely to be classified as normal even it is relatively different to the average trend. So the model learning rate is increased and the trend model will be updated as soon as the sensor readings are returning to normal. One of the benefits of the proposed MoPBAS method is ease of computation, meaning that that the process can be potentially performed on site or event on chip. This provides the opportunity to reduce the data transmission requirement, in which only the information on anomalous events will be sent back to a base station to enable a real-time alert system and normal readings can be logged locally or discarded. This could be a key factor for monitoring sites where the cost of data transmission is very high. Moreover, data transmission over long distance always consumes the majority of power in

WSNs. Thus, applying anomalous detection on site or on chip can extends the deployment time of WSNs which are battery dependant. This novel anomaly detection method inspired from image processing domain provides the fundamental block of creating a dynamic smart wireless sensor network. Above all, it appears that MoPBAS is an suitable anomaly detection approach for wireless sensor networks in the marine environment.

Results in Figures 14 and 21 show that the clustering successfully discriminates between events, assigning them to clusters where events within the same cluster are relativity similar to each other. Unique events are treated as new cluster centres (such as clusters 9 and 13 in the turbidity clustering example). This is a key advantage of using ROC clustering method compare to other commonly used techniques. These unique events are assigned as new cluster centres by ROC, rather than noise in some methods such as K-Means. This feature is very important from a water quality event detection perspective as these events have no analogous events in the past, and thus may be potentially of greater importance to operators. These are the events that would then trigger an alert when detected, thus allowing operators to react accordingly. However, further analysis and determination of the causes and effects of these events require fusion of information from multiple data sources as proposed in the smart sensing system in Figure 1. Another advantage of ROC method is that it is computational inexpensive, which the classification of abnormal events and the update of the model can be performed in real-time. In addition, ROC method is easy to interpret and its tree-like structure can can used to build an event indexing and retrieving system. Defining the number of centres K in ROC is crucial. In this paper, we assigned K equal to 14, which is based on our site surveys and our assumption of the number of possible abnormal event types that may occur at our pilot site. However, finding a suitable K for marine environmental monitoring requires large amount of data collected over long time periods due to seasonal and weather effects.

VI. Conclusion

In this paper, we have demonstrated a novel system of detection and clustering of events in time series datasets of environmental turbidity and salinity measurements. We have modified the pixel-based adaptive segmentation technique from the image processing domain for this purpose and applied robust on-line clustering for classification of events. We have provided this in the context of a component of the proposed framework for an automated sensor network as part of a future smartbay-smartcity project. Such integrated *in situ* sensor networks for environmental applications have the potential to form a significant part of future smart city infrastructure. However, the data generated from such systems must be seamlessly integrated into the overall smart city-smart bay system if they are to be used to full advantage. Use of the generated data for automated data analytics including event detection and classification must be an integral part of this process if such systems are to become ubiquitous and useful in decision support. We altered a state-of-the art background modelling technique from the image processing domain to built a background trend model to detect anomalous events in commonly measured parameters from *in situ* sensor within estuarine systems such as conductivity (derived salinity) and turbidity, using real data generated with in the Dublin Smart City: Smart Bay Network. Furthermore, we have shown the utility of robust online clustering to classify detected events based on their characteristics. The root environmental causes of these events can now be ascertained and a significance level assigned to these events (for example pollution, human activity or storm events). We have demonstrated the utility of this approach in Dublin Bay for detection of abnormal events, and the potential of these techniques for classifying and further enhancing decision support within this framework. By combining the outcome of other parallel research work such as ship traffic detection from visual sensor, the proposed novel multi modality smart sensing system can potentially provide a rich content based indexing and retrieval system to assist the marine scientists better and easier understand and modelling the marine ecosystem. Subsequently, the system can support decision makers in construction of new policies to better protect environmental and coastal resources.

Acknowledgments

The authors acknowledge support from Science Foundation Ireland (grants 07/CE/I147 and SFI/12/RC/2289) and the Beaufort Marine Research Award, carried out under the Sea Change Strategy and the Strategy for Science Technology and Innovation (20062013), with the support of the Marine Institute, funded under the Marine Research Sub-Programme of the National Development Plan 2007-2013.

References

[1] F. Regan, D. Zhang, T. Sullivan, C. Briciu, H. Cooney, K. Murphy, E. O'Connor, N. O'Connor, and A. Smeaton, "A smart city-smart bay project-establishing an integrated water monitoring system for decision support in dublin bay," in *SENSORCOMM 2013, The Seventh International Conference on Sensor Technologies and Applications*, 2013, pp. 75–82.

[2] R. Greenwood, J. Webster, and F. Regan, "Sustainable water: Chemical science priorities. chapter 4 - water monitoring," Royal Society of Chemistry, Tech. Rep., 2007.

[3] J. Grath, R. Ward, A. Scheidleder, and P. Quevauviller, "Report on eu guidance on groundwater monitoring developed under the common implementation strategy of the water framework directive," *J. Environ. Monit.*, vol. 9, pp. 1162–1175, 2007.

[4] A. Lawlor, J. Torres, B. O'Flynn, J. Wallace, and F. Regan, "Deploy: a long term deployment of a water quality sensor monitoring system," *Sensor Review*, vol. 32, no. 1, pp. 29–38, 2012.

[5] B. O'Flynn, F. Regan, A. Lawlor, J. Wallace, J. Torres, and C. O'Mathuna, "Experiences and recommendations in deploying a real-time, water quality monitoring system," *Measurement Science and Technology*, vol. 21, no. 12, p. 124004, 2010.

[6] C. Albaladejo, P. Sánchez, A. Iborra, F. Soto, J. A. López, and R. Torres, "Wireless sensor networks for oceanographic monitoring: A systematic review," *Sensors*, vol. 10, no. 7, pp. 6948–6968, 2010.

[7] P. W. Rundel, E. A. Graham, M. F. Allen, J. C. Fisher, and T. C. Harmon, "Environmental sensor networks in ecological research," *New Phytologist*, vol. 182, no. 3, pp. 589–607, 2009.

[8] J. Ni, C. Zhang, L. Ren, and S. X. Yang, "Abrupt event monitoring for water environment system based on kpca and svm," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 4, pp. 980–989, 2012.

[9] D. Hart, S. A. McKenna, K. Klise, V. Cruz, and M. Wilson, "Canary: a water quality event detection algorithm development tool," in *Proceedings of the World Environmental and Water Resources Congress, ASCE, Reston, VA*, 2007, pp. 1–9.

[10] E. Vugrin, S. A. McKenna, and D. Hart, "Trajectory clustering approach for reducing water quality event false alarms," in *Proc., of ASCE Annual World Environmental and Water Resources Congress, Kansas City, Missouri, May*, 2009, pp. 17–21.

[11] S. A. McKenna, E. D. Vugrin, D. B. Hart, and R. Aumer, “Multivariate trajectory clustering for false positive reduction in online event detection,” *Journal of Water Resources Planning and Management*, vol. 139, no. 1, pp. 3–12, 2012.

[12] L. Perelman, J. Arad, M. Housh, and A. Ostfeld, “Event detection in water distribution systems from multivariate water quality time series,” *Environmental Science & Technology*, vol. 46, no. 15, pp. 8212–8219, 2012.

[13] S. A. McKenna, D. Hart, K. Klise, V. Cruz, and M. Wilson, “Event detection from water quality time series,” in *Proc. ASCE World Envir. & Water Resources Congress, Tampa, Fla*, 2007.

[14] K. P. Singh, N. Basant, and S. Gupta, “Support vector machines in water quality management,” *Analytica Chimica Acta*, vol. 703, no. 2, pp. 152–162, 2011.

[15] K. P. Singh, A. Basant, A. Malik, and G. Jain, “Artificial neural network modeling of the river water qualitya case study,” *Ecological Modelling*, vol. 220, no. 6, pp. 888–895, 2009.

[16] H. R. Maier and G. C. Dandy, “The use of artificial neural networks for the prediction of water quality parameters,” *Water Resources Research*, vol. 32, no. 4, pp. 1013–1022, 1996.

[17] W. Wu, G. C. Dandy, and H. R. Maier, “Protocol for developing ann models and its application to the assessment of the quality of the ann model development process in drinking water quality modelling,” *Environmental Modelling & Software*, vol. 54, pp. 108–127, 2014.

[18] S. A. McKenna, M. Wilson, and K. A. Klise, “Detecting changes in water quality data.” *Journal: American Water Works Association*, vol. 100, no. 1, 2008.

[19] M. W. Koch and S. A. McKenna, “Distributed sensor fusion in water quality event detection,” *Journal of Water Resources Planning and Management*, vol. 137, no. 1, pp. 10–19, 2010.

[20] D. Hou, H. He, P. Huang, G. Zhang, and H. Loaiciga, “Detection of water-quality contamination events based on multi-sensor fusion using an extented dempster–shafer method,” *Measurement Science and Technology*, vol. 24, no. 5, p. 055801, 2013.

[21] D. Ömer Faruk, “A hybrid neural network and arima model for water quality time series prediction,” *Engineering Applications of Artificial Intelligence*, vol. 23, no. 4, pp. 586–594, 2010.

[22] M. V. Storey, B. van der Gaag, and B. P. Burns, “Advances in on-line drinking water quality monitoring and early warning systems,” *Water Research*, vol. 45, no. 2, pp. 741–747, 2011.

[23] F. Regan, A. Lawler, and A. McCarthy, *SmartCoast Project- Smart Water Quality Monitoring System*. Environmental Protection Agency, Ireland, 2009.

[24] F. Regan, A. Lawlor, B. OFlynn, and J. Wallace, “Deploy: Smart demonstration of online water quality monitoring on the river lee, cork, ireland,” Environmental Protection Agency, Tech. Rep., 2011.

[25] S. Roth and J. G. Wilson, “Functional analysis by trophic guilds of macrobenthic community structure in dublin bay, ireland,” *Journal of Experimental Marine Biology and Ecology*, vol. 222, no. 1, pp. 195 – 217, 1998.

[26] J. Wilson, “Productivity, fisheries and aquaculture in temperate estuaries,” *Estuarine, Coastal and Shelf Science*, vol. 55, no. 6, pp. 953 – 967, 2002.

[27] DCC, *Dublin to Waste Project. Environmental Impact Statement*, D. C. Council, Ed. Dublin City Council, 2006, no. 246851, chapter 12.

[28] M. Hofmann, P. Tiefenbacher, and G. Rigoll, “Background segmentation with feedback: The pixel-based adaptive segmenter,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 38–43.

[29] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[30] G. Qiu, “Indexing chromatic and achromatic patterns for content-based colour image retrieval,” *Pattern Recognition*, vol. 35, no. 8, pp. 1675–1686, 2002.

[31] L. Fei-Fei and P. Perona, “A bayesian hierarchical model for learning natural scene categories,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, June 2005, pp. 524–531 vol. 2.

[32] D. Zhang, S. Chen, and K. Tan, “Improving the robustness of online agglomerative clustering methodbased on kernel-induce distance measures,” *Neural Processing Letters*, vol. 21, no. 1, pp. 45–51, 2005.