MONASH
BUSINESS
SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://business.monash.edu/econometrics-and-business-statistics/research/publications

A feature-based approach to outlier detection in water-quality data from in situ sensors

Priyanga Dilini Talagala, Rob J Hyndman,
Catherine Leigh, Kerrie Mengersen,
Kate Smith-Miles

October 2018

Working Paper no/yr

A feature-based approach to outlier detection in water-quality data from in situ sensors

Priyanga Dilini Talagala
Department of Econometrics and Business Statistics, Monash University, Australia, and
ARC Centre of Excellence for Mathematics and Statistical Frontiers, Australia
Email: dilini.talagala@monash.edu
Corresponding author

Rob J Hyndman
Department of Econometrics and Business Statistics, Monash University, Australia, and
ARC Centre of Excellence for Mathematics and Statistical Frontiers, Australia

Catherine Leigh
Institute for Future Environments, Science and Engineering Faculty, Queensland University of Technology, Australia, and
ARC Centre of Excellence for Mathematics and Statistical Frontiers, Australia

Kerrie Mengersen
Science and Engineering Faculty, Queensland University of Technology, Australia, and
ARC Centre of Excellence for Mathematics and Statistical Frontiers, Australia

Kate Smith-Miles
School of Mathematics and Statistics, University of Melbourne, Australia, and
ARC Centre of Excellence for Mathematics and Statistical Frontiers, Australia

5 October 2018

JEL classification: C10 C14 C22

A feature-based approach to outlier detection in water-quality data from in situ sensors

Abstract

We propose a framework that provides early detection of outliers in water-quality data from *in situ* sensors. The proposed framework first identifies the data characteristics that differentiate outlying instances from typical behaviours. Then different transformation methods are applied to make the outlying instances stand out in the transformed data space, and unsupervised outlying detection algorithms are applied to the transformed data space. Using two data sets obtained from *in situ* sensors at study sites, we demonstrate the wide applicability and usefulness of our proposed framework. We show that the proposed framework can perform well with outlying types such as sudden isolated spikes, sudden isolated drops and level shifts, while maintaining very low false detection rates. This framework is implemented in the open source R package *oddwater*.

Keywords: anomaly detection, time series features, water quality sensors, extreme value theory

1 Introduction

Water-quality monitoring largely relies on water samples collected manually. The samples are then analysed within laboratories according to the information required. This type of rigorous laboratory analysis of field-collected samples is crucial in making natural resources management decisions that affect human welfare and environmental conditions. However, with the rapid advances in hardware technology, the use of *in situ* water-quality sensors positioned at different geographic sites has become common practice to acquire real-time measurements of environmental and water-quality variables. Though only a subset of the required water-quality variables can be measured by these sensors, they have several advantages. Their ability to collect large quantities of data and to archive historic records allows for deeper analysis of water-quality variables and thereby improves understanding about the field conditions and water-quality processes (Glasgow et al. 2004). Near-real-time monitoring also allows the operators to identify

and respond to potential issues quickly and thereby manage the operations efficiently. Further, the use of *in situ* sensors can greatly reduce the labor involved in field sampling and laboratory analysis.

Water-quality sensors are exposed to changing environments and extreme weather conditions, and thus are prone to technical error, including failure. Automatic detection of outliers due to technical errors in the water-quality data from *in situ* sensors has therefore captured the attention of many researchers both in the ecology and data science communities (Hill & Minsker 2006; Hill, Minsker & Amir 2009; Archer, Baptista & Leen 2003; Raciti, Cucurull & Nadjm-Tehrani 2012) as manual anomaly detection is not feasible given the amount of data the sensors produce. Therefore, developing powerful automated methods for near-real-time detection of anomalies in water-quality sensor data will have far reaching benefits for both ecologists and natural resource managers and policy makers (Hill, Minsker & Amir 2009).

1.1 Possible types of water-quality anomalies

Anomalies in water-quality data from sensors can be due to many reasons and ultimately decrease the accuracy of the measurements. Here, *anomalies* are specifically defined as due to technical issues with the sensor equipment, i.e., not real water-quality events. As such, anomalies can be associated with human activities. Maintenance issues and calibration failures, for example, can give rise to abnormal values such as large sudden spikes, sudden isolated drops, sudden shifts, drift and impossible values (Horsburgh et al. 2015). Faults or damage to sensor equipment, including battery failures, can also give rise to anomalies, such as missing values, sudden shifts and very low or high variability in the data. Biofouling, which occurs when biological material accumulates on and around the sensors, may cause drift, missing values or sudden spikes in the data (Archer, Baptista & Leen 2003). Air bubbles interfering with the sensors, freezing temperatures, debris and sediment accumulation may also produce anomalies.

These anomalies give rise to several analytical challenges such as outlier detection (Hill, Minsker & Amir 2009; Raciti, Cucurull & Nadjm-Tehrani 2012; Hill & Minsker 2006), drift detection (Archer, Baptista & Leen 2003) and variability detection (or change-point detection) (Ba & McKenna 2015). The anomaly types such as large sudden spikes and sudden isolated drops, which are much higher or lower than surrounding data, sudden shifts and clusters of spikes can be the focus of outlier detection methods (the difference between the two terms 'anomaly' and 'outlier' is further discussed in Section 2.1). Variability detection methods (or change-point

detection methods) can be used to detect very low or high variability sections in the data that are unusual when compared to typical variability during natural daily cycles. Drift detection is a unique problem with many challenges and therefore has emerged as a separate field of study (Archer, Baptista & Leen 2003). With the help of water-quality experts, rule-based methods (Thottan & Ji 2003) can also be incorporated into anomaly detection frameworks to detect anomalies such as impossible values (e.g. negative conductivity values), out-of-range, and missing values.

Our aim in this work is limited to the problem of outlier detection which covers large sudden spikes, sudden isolated drops, sudden shifts and clusters of spikes. Rule-based methods are also incorporated into the proposed framework to flag occurrences of impossible, out-of-range, and missing values.

1.2 Importance of automating the process of detecting outliers

The problem of outlier detection in water-quality data from *in situ* sensors can be divided into two sub-topics according to their focus: (1) identifying errors and corruptions in the data that make the data unreliable and untrustworthy; and (2) identifying real events (e.g. rare but sudden spikes in turbidity associated with rare but sudden high-flow events). Both problems are equally important when making natural resource management decisions that affect human welfare and environmental conditions. Problem 1 can also be considered as a data preprocessing phase before addressing Problem 2. Hereafter, the outliers under Problem 1 will be referred to as 'technical outliers'.

In this work our focus is only on detecting technical outliers, that is detecting unusual measurements in the data that make the data unreliable and untrustworthy, and which have a direct impact on the performance of the subsequent data analysis under Problem 2. According to Keith et al. (1983), the level and degree of confidence of the sensor measurements is one of the main requirements for a properly defined environmental analysis procedure. Any ignorance of Phase 1 can lead to many problems. Researchers and policy makers are unable to use water-quality data containing technical outliers with confidence for decision making and reporting purposes, because erroneous conclusions regarding the quality of the water being monitored could ensue, leading, for example, to inappropriate or unnecessary water treatment, land management or warning alerts to the public (Kotamäki et al. 2009; Rangeti et al. 2015). Missing values and corrupted data can also have an adverse impact on water-quality model building and calibration processes (Archer, Baptista & Leen 2003). Early detection of these technical outliers will limit the

use of corrupted data for subsequent analysis. However, because data arrive near continuously at high speed in large quantities, manual monitoring is highly unlikely to be able to capture all the corruptions. These issues have therefore increased the importance of developing automated methods for early detection of outliers in water-quality data from *in situ* sensors. Furthermore, automating the detection of technical outliers will limit the use of corrupted data in real-time forecasting and online applications such as on-line drinking water-quality monitoring and early warning systems (Storey, Van der Gaag & Burns 2011), predicting algal bloom outbreaks leading to fish kill events and potential human health impacts, forecasting water level and currents etc. (Glasgow et al. 2004; Archer, Baptista & Leen 2003; Hill & Minsker 2006).

1.3 Main contributions

This paper makes three fundamental contributions. First, we propose an unsupervised framework that provides early detection of technical outliers in water-quality data from *in situ* sensors. Second, we provide a comparative analysis to show how effective and reliable density and nearest neighbour distance-based outlier detection techniques for high dimensional data are for the detection of technical outliers in such data. Third, we introduce an R package, `oddwater` (Talagala & Hyndman 2018), that implements the proposed framework and related functions.

Our proposed framework has many advantages. (1) It can take the correlation structure of the water-quality variables into account when detecting anomalies. (2) It can be applied to both univariate and multivariate problems. (3) The anomaly detection methods that we are considering in the proposed framework are unsupervised, data-driven approaches and therefore do not require training data sets for the model building process, and can be extended easily to other time series from other sites. (4) The anomalous thresholds have a probabilistic interpretation. (5) The framework can easily be extended to streaming data such that it can provide near-real-time support. (6) The proposed framework also has the ability to deal with irregular time series.

In Section 2 we present the background work on outlier detection in water-quality sensor data. In Section 3 we introduce the proposed framework for detection of technical outliers in the data. The methods are trialed on data in Section 4, and the R package `oddwater` is introduced in Section 5. Section 6 concludes the paper.

2 Background

This section reviews the background work on outlier detection and lays the foundation for the framework proposed in Section 3.

2.1 Definitions

Solutions to the problem of detecting unusual behaviours in systems of interest can be influenced heavily by the way in which anomalies are defined. Three terms are used commonly and interchangeably in the literature to describe work related to the topic: *anomaly* (Hyndman, Wang & Laptev 2015; Kumar et al. 2016), *outlier* (Wilkinson 2018; Schwarz 2008) and *novelty* (Clifton, Hugueny & Tarassenko 2011; Hugueny 2013). However, Faria et al. (2016) differentiate between these three terms, using the terms anomaly and outlier to refer to the idea of an undesired pattern, but novelty to refer to the emergence of a new concept that needs to be incorporated into the typical behaviour of the system. The two terms *anomaly* and *outlier* are commonly used interchangeably; however, in this article, we will limit the use of the term *outlier* to define a point which is significantly different (with respect to distance or density) from the majority of a given data space. We use the term *anomaly* to describe any behavior which is unusual or unexpected with respect to water-quality considerations (for example, a drift in the sensor measurement due to biofouling). Thus, *outliers* can be considered as a subset of the class of *anomalies*. In particular, sudden spikes, sudden isolated drops, and sudden shifts are categorized as *outliers*, and any unexpected behaviours (such as impossible values, out of range values, missing values high variability, low variability etc.) including *outliers* are labeled as *anomalies*. In this work, our focus is on *outlier* detection. Rule-based methods are also incorporated in the proposed framework to detect certain types of anomalies such as impossible values, out of range values and missing values (Thottan & Ji 2003).

In line with our *outlier* definition, Grubbs (1969) also defined an outlier as an observation that deviates markedly from other members of the dataset. This deviation can be defined in terms of either distance or density. Burridge & Taylor (2006), Wilkinson (2018) and Schwarz (2008) have proposed methods for outlier detection by defining an outlier in terms of distance. In contrast, Hyndman (1996), Clifton, Hugueny & Tarassenko (2011), Hugueny, Tarassenko & Clifton (2008) and Talagala et al. (2018) proposed methods that define an outlier with respect to either the density or the probability of the occurrence of observations. Madsen (2018) also provides a series of distance- and density-based outlier detection algorithms.

2.2 Categorizations of outliers

The problem of outlier detection for temporal data has different categorizations. According to Gupta et al. (2014), the problem is threefold: (a) detection of contextual outliers within a given series; (b) detection of outlying sub-sequences within a given series; and (c) detection

of outlying series within a space of a collection of series. In this categorization, contextual outliers are referred to as single observations that are surprisingly large or small, independent of the neighbouring observations. On the other hand, Goldstein & Uchida (2016) provide a different categorization for temporal data, namely (a) point outliers, (b) collective outliers and (c) contextual outliers. Although there is a clear overlap between the two categorizations, they assign slightly different definitions in each case. Goldstein & Uchida (2016) use the term point outlier instead of contextual outliers to define single outlying instances. Their definition of contextual anomalies differs slightly from that of Gupta et al. (2014). In Goldstein & Uchida (2016), a contextual outlier is a point that can be seen as typical in general, but when context is taken into account the point is seen to be an outlier. Time can be considered as the most common occurring context. For example, 28°C may be normal in summer, but such a high temperature during winter could be considered as an outlier in some regions. On the other hand, a collective outlier is represented as a set of many points. That is, each of the individual points is not necessarily a point outlier, but taken together they are unusual and constitute a collective outlier. According to Goldstein & Uchida (2016), to address collective outliers, correlation, aggregation and grouping can be used to generate a new data set with a different representation of features. ???? Sometimes these categorizations are more relevant and meaningful for certain applications than others. In our water-quality data, for example, most outliers are recognized as contextual outliers. For instance, a sudden shift in turbidity is not necessarily an outlier if it is followed by a gradually decaying tail (which would be more likely to be a real event) but if the sudden shift is followed by a sudden drop, then it would be an outlier, possibly due to a technical fault in the sensor. Further the correlation between different variables also needs to be taken into account when deciding outlying behavior.

The problem of outlier detection in the non-temporal context can also be divided into three categorizations: (a) detection of global outliers; (b) detection of local outliers; and (c) detection of micro clusters or clusters of outliers (Goldstein & Uchida 2016). Most of the existing outlier detection methods for high dimensional data can easily recognize global outliers as they are very different from the dense area with respect to their attributes. In contrast, a local outlier is only an outlier when it is distinct from and compared with its local neighbourhood. Madsen (2018) introduces a set algorithms based on a density or distance definition of outliers, which mainly focuses on local outliers in multidimensional domains. Micro clusters or clusters of outliers may cause masking problems. The recently proposed HDoutlier algorithm (Wilkinson 2018) addresses this problem by incorporating a clustering step into the outlier detection algorithm.

Following Goldstein & Uchida (2016), to address collective outliers in the water-quality data we applied different transformations to the original time series and thereby converted the original problem of outlier detection in a temporal context to a non-temporal context. Our focus was then on identifying the global, local and micro clusters in the transformed space. The corresponding points in the original time series were then declared as outliers in the final step.

2.3 Outlier detection in water-quality data from *in situ* sensors

Hill & Minsker (2006) addressed the problem of outlier detection in environmental sensors using regression-based time series models. In this work they addressed the scenario as a univariate problem. Their prediction models are based on four data driven methods: naïve, clustering, perceptron, and Artificial Neural Networks (ANN). Measurements that fell outside the bounds of an established prediction interval were then declared as outliers. They also considered two strategies: anomaly detection (AD) and anomaly detection and mitigation (ADAM) for the detection process. ADAM replaces detected outliers with the predicted value prior to the next predictions whereas AD simply uses the previous measurements without making any alteration to the detected outliers. These types of data-driven methods develop models using sets of training examples containing a feature set and a target output. Later, Hill, Minsker & Amir (2009) addressed the problem by developing three automated anomaly detection methods using dynamic Bayesian networks (DBN) and showed that DBN-based detectors using either robust Kalman filtering or Rao-Blackwellized particle filtering, outperform that of Kalman filtering.

The second most common approach for detecting outliers in environmental sensor data is based on the residuals (the differences between predicted and actual values). Due to the ability of ANNs to model a wide range of complex non-linear phenomena, Moatar, Fessant & Poirel (1999) also used ANN techniques to detect anomalies such as abnormal values, discontinuities, and drifts in pH readings. After developing the pH model, the Student t-test and the cumulative Page–Hinkley test were applied to detect changes in the mean of the residuals to detect measurement error occurring over a short period of time. Later, Moatar, Miquel & Poirel (2001) expanded their work to a multivariate scenario with some additional water-quality variables including dissolved oxygen, electrical conductivity, pH and temperature. The proposed algorithm used both deterministic and stochastic approaches for the model building process. The measures of the variables were then compared with the model forecasts using a set of classical statistical tests to detect outliers in the measurements. This work also demonstrates the effectiveness and advantages of the multimodel approach.

Detection of failures in the water-quality sensors due to biofouling is the focus of the work done by Archer, Baptista & Leen (2003). When biological materials accumulates on and around the sensors, it could cause drifts in the sensor measurements and thereby can give rise to the analytical challenges related to drift detection. The approaches by Archer, Baptista & Leen (2003) for early detection of biofouling are mainly based on the sequential likelihood ratio test. These detectors also have the ability to provide estimates of biofouling onset time, which is useful for the subsequent step of correction of anomalies in the water-quality data.

A common feature of all of the above methods is that they are supervised or semi-supervised and require training data. In certain applications, all the possible outliers are not known in advance and can arise spontaneously as new outlying behaviors during the test phase. In such situations, supervised methods may fail to detect those outliers. Semi-supervised methods are also unsuitable for certain applications due to the unavailability of training data containing only typical instances that are free from outliers (Goldstein & Uchida 2016). The data sets considered under Section 4 suffer from both of these limitations and therefore highlight the need for a more general approach.

3 Methodology

This section introduces the proposed framework to detect technical outliers in water-quality data from *in situ* sensors. The proposed framework involves: (a) identifying the data characteristics or patterns that differentiate outlying instances from typical behaviours; (b) identifying suitable transformation methods to make the outlying instances stand out in the data space; and (c) applying outlier detection algorithms for high dimensional data to the new data space constructed by the transformed series. Our proposed algorithm is unsupervised and applicable to both univariate and multivariate data.

The structure of this section is organised according to the main steps involved in the proposed framework: (1) variable selection; (2) rule-based approaches; (3) data preprocessing; (4) outlier score calculation; (5) outlier threshold calculation; and (6) performance evaluation.

3.1 Variable Description

The current work is limited to three water-quality variables: turbidity, conductivity and river level. However the proposed framework can be easily generalized to other water-quality variables All *in situ* sensors are housed within a gauging station on the bank of the stream. This setup also means that the sensors are less susceptible to biofouling. Water is pumped up from

the stream to the station approximately every 60 or 90 minutes for readings to be logged by the sensors.

Turbidity

Turbidity, measured in Nephelometric Turbidity Units (NTU), is a visual property of water indicative of its clarity due to suspended particles that absorb and scatter light. According to Panguluri et al. (2009), turbidy is considered one of the principal physical characteristics of water. Turbidity tends to increase during high inflow events in rivers which carry high suspended sediment loads (e.g. from runoff-derived erosion). It can also become high during times of low flow due to concentration of suspended particles and when there are high concentrations of algae in the water column (which often occurs when temperatures and water residence times are high). The greater the scattering of light, the higher the turbidity. Low turbidity values therefore indicate high water clarity whereas high values indicate low water clarity. As can be seen in Figure 1(a), the rate at which turbidity rises is typically faster than the rate at which it falls.

Conductivity

Conductivity, measured in Siemens per centimeter (μS/cm), is another key water-quality variable which determines several fundamental physical properties of water (Tutmez, Hatipoglu & Kaymak 2006). It reflects the ability of water to pass an electrical current and is mainly affected by the presence of inorganic dissolved solids such as chloride, nitrate, sulfate, sodium, magnesium, calcium, iron, etc.; and temperature. The warmer the water, the higher the conductivity. For this reason, conductivity is usually reported at a given temperature (usually at 25 °C).

The higher the salt concentration, the more easily ions pass through water. Most freshwater ranges from ~10 to 1000 μS/cm. It is generally more stable than dissolved oxygen and temperature but can change rapidly in response to inputs of new water. Due to the concentration effect, conductivity typically decreases with freshwater inflows (e.g. during wetter months) and increases when river water levels and discharge volumes decline (e.g. during drier months). In other words, conductivity will generally increase through time in the absence of freshwater inflows. The greatest drop in conductivity will typically occur on the first inflow event after a period of low/base flow. Its fall rate is generally faster than its rise rate.

River level

River level, measured in meters (m), reflects the depth of water at a location in space and time. Level increases (usually rapidly) when inflow events occur and decreases (usually more slowly) in the absence of such events. Its rates of rise are typically fast compared with fall rates (Figure 1(c)). However, exceptions can occur in managed waterways where releases of water can be controlled. In highly seasonal regions, water levels will typically be much lower on average in the drier months than the wetter months.

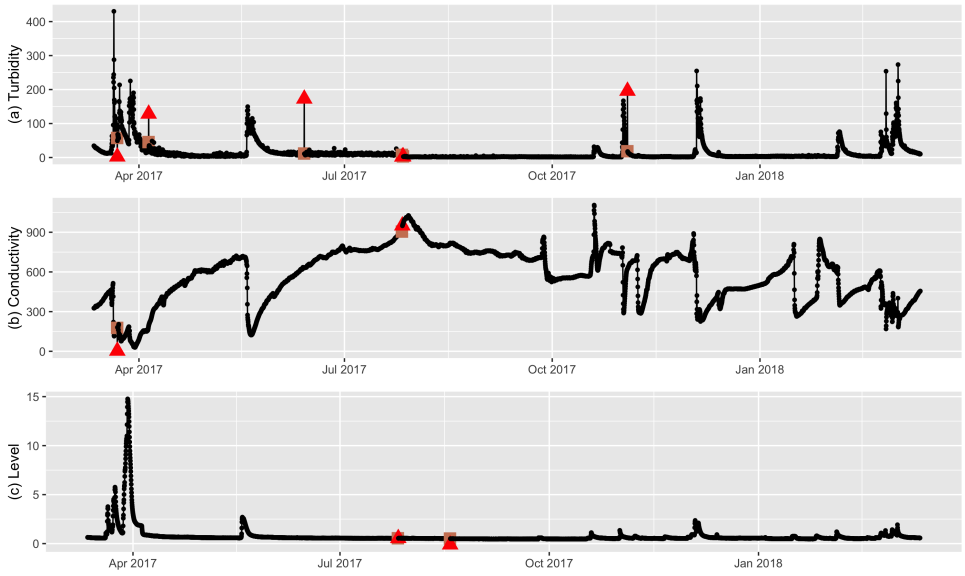

Figure 1: *Time series for turbidity (NTU), conductivity (µS/cm) and river level (m) measured by in situ sensors at Sandy Creek. In each plot, outliers determined by water-quality experts are shown in red. Typical points are shown in black.*

In general, we expect turbidity and conductivity to have a negative relationship; that is conductivity decreases as turbidity increases. Turbidity and river level are expected to show a positive relationship (turbidity increases with level), but there may be a time offset (e.g. peak turbidity may occur just prior to or after the peak river level (clockwise or anti-clockwise hysteresis) and the rate of increase in turbidity may be highest at the lower end of changes in level (i.e. a change from 0 to 5m vs 5 to 15m). Conductivity and river level may display different relationships for different levels. Conductivity may gradually increase to very high values as levels decline, particularly if the river level has remained low for some time. Much lower conductivity is expected at medium to high river levels. These bivariate relationships are apparent in Figure 2 which is based on the water-quality data measured by *in situ* sensors at Sandy Creek introduced in Section 4.

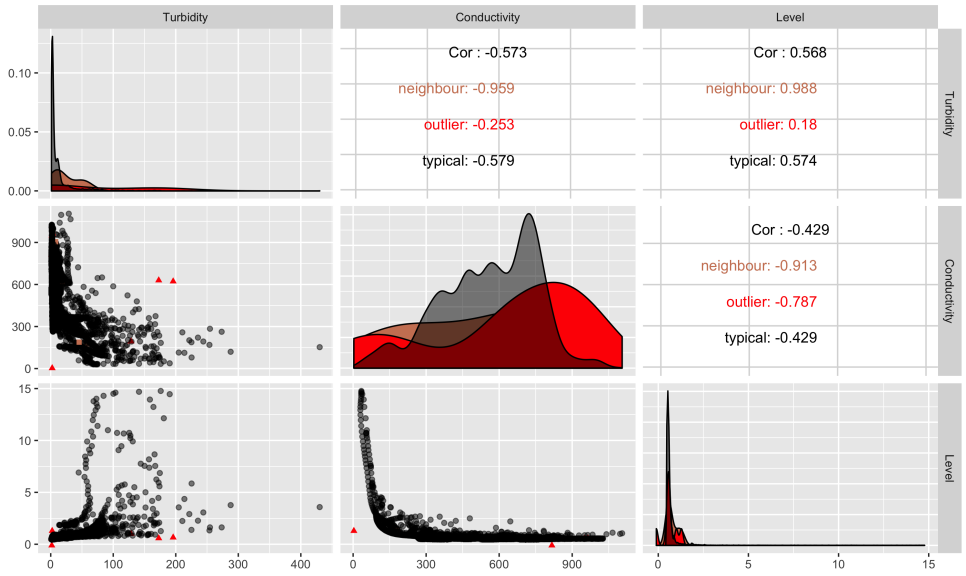

Figure 2: *Bivariate relationships between water-quality variables (turbidity (NTU), conductivity (µS/cm) and river level (m)) measured by in situ sensors at Sandy Creek. In each scatter plot outliers determined by water-quality experts are shown in red, while typical points are shown in black. Diagonal plots show the distribution of individual variables.*

3.2 Rule-based approaches

Early work in the area of fault or anomaly detection used rule-based approaches which rely on expert knowledge regarding typical behaviour and the possible types of anomalies of a given system. Following Thottan & Ji (2003), we incorporated simple rules into our algorithm to detect anomalies such as out-of-range values, impossible values (e.g. negative values) and missing values.

Water-quality sensors have their own specification. If a sensor reading is outside the corresponding sensor detection range it is then marked as an outlier. Negative readings are also inaccurate and impossible for river turbidity, conductivity and level. A simple constraint is imposed on the algorithm to filter these values and mark them as outliers. Missing values are also frequently encountered in water-quality sensor data (Rangeti et al. 2015). Flagging missing values is important as they could be due to damage to sensor equipment, battery failures or biofouling that requires immediate action. Missing values are detected by calculating the time gaps between readings. If a gap exceeds the maximum allowable time difference between any two consecutive readings as determined by water-quality experts, the corresponding time stamp is then marked as an anomaly due to missingness.

3.3 Data preprocessing

Our next task involves identifying candidate input *data features* for the classifier by understanding the common characteristics or patterns of the possible types of outliers in water-quality data that differentiate them from typical instances or events. For turbidity for example, "extreme" deviations upward are more likely than deviations downwards. The opposite is true for conductivity. Further, in a turbidity time series a sudden isolated upward shift (spike) is a point anomaly, but if the sudden upward shift is followed by a gradually decaying tail then it becomes part of the typical behaviour. In general, isolated data points that are outside the general trend are outliers. Experts' knowledge was heavily involved in this step to identify the common characteristics of the possible types of outliers in water-quality data from *in situ* sensors. Hereafter, these characteristics will be referred as 'data features'.

After identifying the data features, different transformations were then applied to the time series with the aim of highlighting different types of outliers, with a focus on sudden isolated spikes, sudden isolated drops, sudden shifts, and clusters of spikes. Table 1 summaries the transformation methods used to highlight different types of outliers in water-quality data from *in situ* sensors.

Figure 3 shows the evolution of the data space according to the different transformation explained in Table 1. Figure 3(a) corresponds to the original series where only few target outliers (labeled by water-quality experts) are somewhat visible from its majority. Figure 3(e) which is based on one sided derivative transformation (as explained in Table 1) gives a clear separation between most of the target outlying points and the typical points. Figure 4 gives the one sided derivative transformed series of the original series given in Figure 1 and it demonstrates how one sided derivatives assist in separating sudden isolated spikes (outliers) from its typical behavior with sudden shift with a gradually decaying tail. The corresponding bivariate relationship between the transformed series are given in Figure 1.

3.4 Outlier score calculation

For our current framework we considered eight unsupervised outlier detection algorithms for high dimensional data.

HDoutliers algorithm

The HDoutliers algorithm (Wilkinson 2018) is an unsupervised outlier detection algorithm that searches for outliers in high dimensional data assuming there is a large distance between

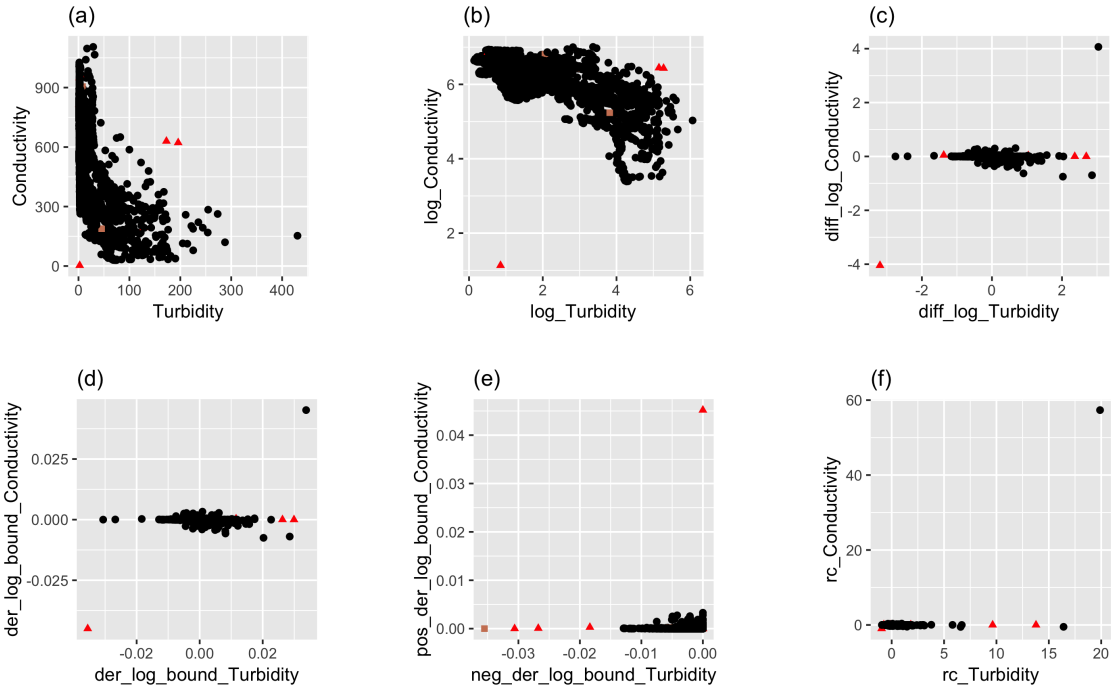

Figure 3: *Bivariate relationships between transformed series of turbidity and conductivity measured by in situ sensors at Sandy Creek. In each scatter plot outliers determined by water-quality experts are shown in red, while typical points are shown in black. (a) Original series, (b) Log transformation, (c) First difference, (d) First derivative, (e) One sided derivative, and (f) Rate of change.*

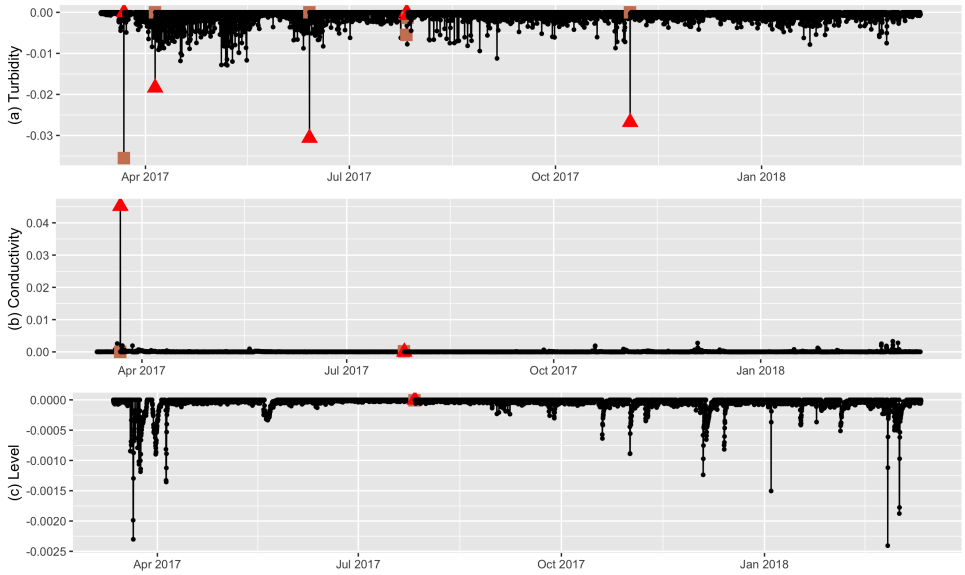

Figure 4: *Transformed series (one sided derivatives) of turbidity (NTU), conductivity (µS/cm) and river level (m) measured by in situ sensors at Sandy Creek. In each plot outliers determined by water-quality experts are shown in red, while typical points are shown in black.*

Table 1: *Transformation methods to highlight different types of outliers in water-quality sensor data. Let Y_t represent an original series from one of the three variables : turbidity, conductivity and level at time t.*

Transformation	Formula	Data Feature	Focus
Log transformation	$\log(y_t)$	High variability of the data.	To stabilize the variance across time series and make the patterns more visible (e.g. level shifts)
First difference	$\log(y_t/y_{t-1})$	Isolated spikes (to both positive and negative directions) that are outside the general trend are considered as outliers. Under typical behaviour sudden upward (downward) shift are possible for turbidity (conductivity). But their rate of fall (rise) is generally slower than the rate of rise (fall) under typical behaviour.	To separate isolated spikes from the general upward/downward trend patterns.
Time gap	Δt		To identify missing values
First derivative	$x_t = \log(y_t/y_{t-1})/\Delta t$	Data are unevenly spaced time series.	To handle irregular time series. Data points with large gaps will get small value. Large gaps indicate the lack of information to make a claim regarding the points.
One sided derivative			
Turbidity or level	$\min\{x_t, 0\}$	Extreme upward trend in Turbidity and level under typical behaviour.	To separate spikes from typical upward trends.
Conductivity	$\max\{x_t, 0\}$	Extreme downward trend in Conductivity under typical behaviour.	To separate isolated drops from typical downward trends.
Rate of change	$(y_t - y_{t-1})/y_t$	High or low variability in the data.	To detect change points in variance.

outliers and the typical data. Nearest neighbour distances between points are used to detect outliers. However, variables with large variance can bring disproportional influence on Euclidean distance calculation. Therefore, the columns of the data sets are first normalized such that the data are bounded by the unit hyper-cube. To deal with data sets with a large number of observations, the Leader algorithm (Hartigan 1975) is used to form several clusters of points in one pass through the data set. This is done with the aim of handling micro clusters (Goldstein & Uchida 2016). After forming small mutually exclusive clusters, a representative member is selected from each cluster. The nearest neighbour distances are then calculated for the selected representative members. Using extreme value theory, an outlying threshold is calculated to differentiate outliers from typical points. Section 3.5 further elaborates on the outlying threshold calculation.

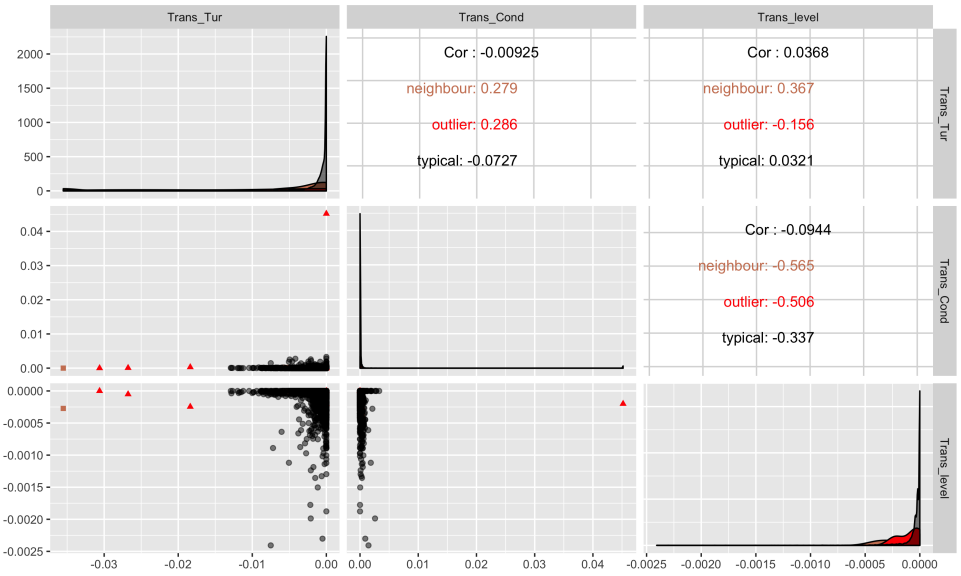

Figure 5: *Bivariate relationships between transformed series (one sided derivative) of turbidity (NTU), conductivity (µS/cm) and river level (m) measured by in situ sensors at Sandy Creek. In each scatter plot outliers determined by water-quality experts are are shown in red, while typical points are shown in black. Diagonal plots show the distribution of individual transformed series.*

KNN-AGG and KNN-SUM algorithms

The HDoutliers algorithm uses only nearest neighbour distances to detect outliers under the assumption that any outlying point (or outlying clusters of points) present in the data set is isolated. For example, if there are two outlying points (or two outlying clusters of points) that are close to one another, but are far away from the rest of the valid data points, then the two outlying points (or two outlying clusters of points) become nearest neighbors to one another and give a small nearest neighbour distance for each outlying point (or each outlying cluster of points). Since the HDoutlier algorithm is totally dependent on the nearest neighbour distances, and since the two outlying points (or two outlying clusters of points) do not show any significant deviation from other typical points with respect to nearest neighbour distance, the HDoutliers algorithm now fails to detect these points as outliers.

Following Angiulli & Pizzuti (2002), Madsen (2018) proposed two algorithms: aggregated k-nearest neighbour distance (KNN-AGG); and sum of distance of k-nearest neighbours (KNN-SUM) to overcome this limitation by incorporating k nearest neighbour distance for the outlier score calculation. The algorithms start by calculating the k nearest neighbour distances for each point. The k-dimensional tree (kd-tree) algorithm (Bentley 1975) is used to identify the k nearest neighbours of each point in a fast and efficient manner. A weight is then calculated using the k nearest neighbour distances and the observations are ranked such that outliers are those points having the largest weights. For KNN-SUM, the weight is calculated by taking the summation of

the distances to the k nearest neighbours. For KNN-AGG, the weight is calculated by taking a weighted sum of distances to k nearest neighbours, assigning nearest neighbours higher weight relative to the neighbours father apart.

LOF algorithm

The Local Outlier Factor (LOF) algorithm (Breunig et al. 2000) calculates an outlier score based on how isolated a point is with respect to its surrounding neighbours. Here the data points with a lower density than their surrounding points are identified as outliers. The local reachable density of a point is calculated by taking the inverse of the average readability distance based on the k (user defined) nearest neighbours. This density is then compared to the density of the corresponding nearest neighbours by taking the average of the ratio of the local reachability density of a given point and that of its nearest neighbours.

COF algorithm

One limitation of LOF is that it assumes that the outlying points are isolated and therefore fails to detect outlying clusters of points that share few outlying neighbours if k is not properly selected (Tang et al. 2002). This is generally known as a masking problem (Hadi 1992). That is, LOF assumes both low density and isolation to detect outliers. However isolation can imply low density, but the other direction is not always true. In general, low density outliers result from deviating from a high density region and an isolated outlier results from deviating from a connected pattern. Tang et al. (2002) addressed this problem by introducing a Connectivity-based Outlier Factor (COF) that compares the average chaining distances between points subject to outlier scoring and the average of that of its neighbouring to their own k-distance neighbours.

INFLO algorithm

Detection of outliers is challenging when data sets contain adjacent multiple clusters with different density distributions (Jin et al. 2006). For example, if a point from a sparse cluster is close to a dense cluster, this could be misclassified as an outlier with respect to the local neighbourhood as the density of the point could be derived from the dense cluster instead of the sparse cluster itself. This is another limitation of LOF (Breunig et al. 2000). The Influenced Outlierness (INFLO) algorithm (Jin et al. 2006) overcomes this problem by considering both the k nearest neighbours (KNNs) and reverse nearest neighbours (RNNs) which allows it to obtain a better estimation of the neighbourhood's density distribution. The RNNs of an object, p for example, are essentially the objects that have p as one of their k nearest neighbours. It helps to

distinguish typical points from outlying points as they have no RNNs. To reduce the expensive cost incurred by searching a large number of KNNs and RNNs, the kd-tree algorithm was used during the search process.

LDOF algorithm

The Local Distance-based Outlier Factor (LDOF) algorithm (Zhang, Hutter & Jin 2009) also uses the relative location of a point to its nearest neighbours to determine the degree to which the point deviates from its neighbourhood. It computes the distance for an observation to its k-nearest neighbours and compares the distance with the average distances of the point's nearest neighbours. In contrast to LOF (Breunig et al. 2000) which uses local density, LDOF now uses relative distances to quantify the deviation of a point from its neighbourhood system. One of the main differences between the two approaches LDOF and LOF is that, in LDOF, the typical pattern of the data set is represented by scattered points rather than crowded main clusters as in LOF (Zhang, Hutter & Jin 2009).

RKOF algorithm with Gaussian kernel

According to Gao et al. (2011), LOF is not accurate enough to detect outliers in complex and large data sets. Furthermore the performance of LOF depends on the parameter k that determines the scale of the local neighbourhood. The Robust Kernel-based Outlier Factor (RKOF) algorithm (Gao et al. 2011) tries to overcome these problems by incorporating variable kernel density estimates to address the first problem and weighted neighbourhood density estimates to address the second problem.

3.5 Outlier threshold calculation

As explained in Section 3.4, outlier scores assign each point a degree of being an outlier. However, for certain applications it is also important to categorize typical and outlying points. For example, if the correction process is automated upon the detection of outliers then filtering the outlying observations is important prior to the subsequent data correction process.

Ideally we prefer a universal threshold to unambiguously distinguish outlying points from typical points. Following Schwarz (2008), the HDoutliers algorithm (Wilkinson 2018) defines an anomalous threshold based on extreme value theory, a branch of probability theory that relates to the behaviour of extreme order statistics in a given sample (Galambos, Lechner & Simiu 2013).

The anomalous threshold calculation in Schwarz (2008); Burridge & Taylor (2006); Wilkinson (2018) is an application of Weissman's spacing theorem (Weissman 1978) that is applicable to the distribution of data covered by the maximum domain of attraction of Gumbel distribution. Fortunately, this requirement is satisfied by a wide range of distributions, ranging from those with light tails to moderately heavy tails that decrease to zero faster than any power function (Embrechts, Klüppelberg & Mikosch 2013).

Following Schwarz (2008); Burridge & Taylor (2006); Wilkinson (2018), we start our anomalous threshold calculation from a subset containing the 50% of observations with the smallest outlier scores, under the assumption that this subset contains the outlier scores corresponding to typical data points and the remaining subset contains the scores corresponding to the possible candidates for outliers. Following Weissman's spacing theorem (Weissman 1978), the algorithm then fits an exponential distribution to the upper tail of the outlier scores of the first subset, and computes the upper $1 - \alpha$ points of the fitted cumulative distribution function, thereby defining an outlying threshold for the next outlier score. Then from the remaining subset it selects the point with the smallest outlier score. If this outlier score exceeds the cutoff point, it flags all the points in the remaining subset as outliers and stops searching for outliers. Otherwise, it declares the point as a non-outlier and adds it to the subset of the typical points. It then updates the cutoff point including the latest addition. This searching algorithm continues until it finds an outlier score that exceeds the latest cutoff point. This algorithm is known as a "bottom up searching" algorithm (Schwarz 2008). This threshold calculation was done under the assumption that the distribution of outlier scores produced by each algorithm (explained in Section 3.4) is in the maximum domain of attraction of the Gumbel distribution which covers a wide range of distributions.

3.6 Performance evaluation

In this step, we perform an experimental evaluation on the accuracy and computational efficiency of our proposed framework with respect to the eight outlier detection algorithms for high dimensional data explained in Section 3.4, using the different transformations and different combinations of variables. Our goal was to detect suitable transformations, combinations of variables, and the algorithms for outlier score calculation for the two sites introduced in Section 4. This interplay may depend on the characteristics of the time series (site and time dependent for example), and what is best for one site may not be the best for another site.

All the experiments were evaluated with respect to some common measures for binary classification based on the values of the confusion matrix which summarizes the false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN). The measures include accuracy $((\text{TP}+\text{TN})/(\text{TP}+\text{FN}+\text{FN}+\text{TN}))$ which explains the overall effectiveness of a classifier; error rate $((\text{FN}+\text{FN})/(\text{TP}+\text{FN}+\text{FN}+\text{TN}))$ which explains the misclassification of the classifier and geometric-mean $(\sqrt{\text{TP}*\text{TN}})$ which explains the relative balance of TP and TN of the classifier (Sokolova & Lapalme 2009). According to Hossin & Sulaiman (2015), these measures are not enough to capture the poor performance of the classifiers in the presence of imbalanced data sets. Since our data sets obtained from *in situ* sensors are highly imbalanced and are negatively dependent (i.e. containing many more negatives than positives), we recorded three additional measures which are recommended for imbalanced two class problems by Ranawana & Palade (2006): the negative predictive values $(\text{NPV}=\text{TN}/(\text{FN}+\text{TN}))$ which measures the probability of a negatively predicted pattern actually being negative; positive predictive value $(\text{PPV}=\text{TP}/(\text{TP}+\text{FN})$ which measures the probability of a positively predicted pattern actually being positive, and optimized precision (OP) which is a combination of accuracy, sensitivity and specificity metrics (Ranawana & Palade 2006).

The proposed framework is implemented using the R programming language (R Core Team 2018). The computation time of the framework depends on the number and the length of the time series. We timed our code with the `microbenchmark` package (Mersmann 2018) for different combinations of algorithms, transformations and variable combinations on a MacBook Air with a 2.2 GHz Intel Core i7 CPU and 8 GB memory.

4 Application

We now evaluate the effectiveness of our proposed framework using two data sets obtained from *in situ* sensors positioned at two study sites that are in freshwater reaches, one on Pioneer River and one on Sandy Creek in Australia's northeast with catchment areas of 1466 km^2 and 326 km^2, respectively.

4.1 Analysis of water-quality data from *in situ* sensors at Sandy Creek

The water-quality data obtained from *in situ* sensors located at Sandy Creek were available from 12 March 2017 to 12 March 2018. The data set included 5402 recorded points. These time series are irregular time series (i.e. the frequency of observations is not constant) as shown in Figure 6 with minimum time gap of 10 minutes and maximum time gap around 4 hours.

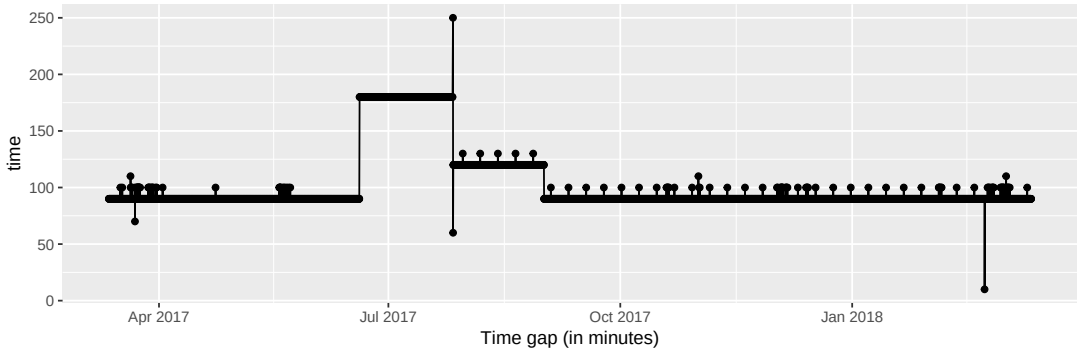

Figure 6: *Time gap between recorded time points of the sensors positioned at Sandy Creek.*

According to Figure 2 which gives the bivariate relationship between original water-quality variables (turbidity (NTU), conductivity (μS/cm) and river level (m)) measured at the same time by *in situ* sensors at Sandy Creek, there is no clear separation between the target outliers (labeled by water-quality experts) and the typical points in the original data space. However a clear separation is apparent between the two sets of points once the one sided derivative transformation (an appropriate transformation for unevenly spaced data (Figure 5 is applied to the original series. Having this type of a separation between outlier and typical points is important before applying unsupervised outlier detection for high dimensional data as the methods are usually based on the definition of outliers in terms of distance or density.

Table 2 summarizes the performance metrics of outlier detection algorithms performed on the transformed series (based on derivatives and one sided derivatives) of the data (turbidity, conductivity and level) obtained from *in situ* sensors at Sandy Creek and it is organised in descending order of NPV values. According to Ranawana & Palade (2006), NPV is the most recommended measurement for negatively dependent data (that is, more negatives (typical points) than positives (outliers) within the data set) and the focus is more on sensitivity (the proportion of positive patterns being correctly recognized as being positive) than specificity.

According to Table 2, the one sided derivative transformation outperforms the first derivative transformation (rows 1–5). Not surprisingly, the false positive rate for the first derivative is somewhat higher than that of the one sided derivative. This is because, in an occurrence of a sudden spike (or isolated drop) the first derivative assigns high values to the actual outlying point as well as the neighbouring point. The goal of taking the one sided derivative is to differentiate an actual outlying point from its immediate neighbouring point which is also claimed to be an outlier. Therefore deriving one sided derivatives from the corresponding first derivatives sometimes leads to the generation of false negatives. For example, in an occurrence of a sudden level (mean) shift in the series one sided derivative could identify the neighbouring

Table 2: *Performance metrics of outlier detection algorithms performed on multivariate water-quality time series data from in situ sensors at Sandy Creek. (Rows are arranged in descending order of NPV values)*

i	TimeSeries	Tranformation	Method	TN	FN	FP	TP	Accuracy	Error_Rate	Geometric_mean	Optimised_Precision	PPV	NPV
1	**T-C**	**One sided Derivative**	**HDoutliers**	**5389**	**7**	**1**	**5**	**0.9985**	**0.0015**	**164.1493**	**0.5868**	**0.8333**	**0.9987**
2	**T-C**	**One sided Derivative**	**KNN-AGG**	**5389**	**7**	**1**	**5**	**0.9985**	**0.0015**	**164.1493**	**0.5868**	**0.8333**	**0.9987**
3	**T-C**	**One sided Derivative**	**KNN-SUM**	**5389**	**7**	**1**	**5**	**0.9985**	**0.0015**	**164.1493**	**0.5868**	**0.8333**	**0.9987**
4	**T-C-L**	**One sided Derivative**	**KNN-AGG**	**5387**	**8**	**1**	**6**	**0.9983**	**0.0017**	**179.7832**	**0.5984**	**0.8571**	**0.9985**
5	**T-C-L**	**One sided Derivative**	**KNN-SUM**	**5387**	**8**	**1**	**6**	**0.9983**	**0.0017**	**179.7832**	**0.5984**	**0.8571**	**0.9985**
6	T-C	First Derivative	HDoutliers	5382	9	8	3	0.9969	0.0031	127.0669	0.3973	0.2727	0.9983
7	T-C	First Derivative	KNN-AGG	5382	9	8	3	0.9969	0.0031	127.0669	0.3973	0.2727	0.9983
8	T-C	One sided Derivative	LDOF	5389	10	1	2	0.9980	0.0020	103.8171	0.2837	0.6667	0.9981
9	T-C	One sided Derivative	RKOF	5374	10	16	2	0.9952	0.0048	103.6726	0.2816	0.1111	0.9981
10	T-C	First Derivative	KNN-SUM	5388	11	2	1	0.9976	0.0024	73.4030	0.1515	0.3333	0.9980
11	T-C	First Derivative	INFLO	5385	11	5	1	0.9970	0.0030	73.3826	0.1510	0.1667	0.9980
12	T-C	First Derivative	COF	5389	11	1	1	0.9978	0.0022	73.4098	0.1517	0.5000	0.9980
13	T-C	First Derivative	LDOF	5387	11	3	1	0.9974	0.0026	73.3962	0.1513	0.2500	0.9980
14	T-C	First Derivative	LOF	5387	11	3	1	0.9974	0.0026	73.3962	0.1513	0.2500	0.9980
15	T-C	First Derivative	RKOF	5385	11	5	1	0.9970	0.0030	73.3826	0.1510	0.1667	0.9980
16	T-C	One sided Derivative	INFLO	5388	11	2	1	0.9976	0.0024	73.4030	0.1515	0.3333	0.9980
17	T-C	One sided Derivative	COF	5388	11	2	1	0.9976	0.0024	73.4030	0.1515	0.3333	0.9980
18	T-C	One sided Derivative	LOF	5378	11	12	1	0.9957	0.0043	73.3348	0.1499	0.0769	0.9980
19	T-C-L	First Derivative	HDoutliers	5387	12	1	2	0.9976	0.0024	103.7979	0.2476	0.6667	0.9978
20	T-C-L	First Derivative	KNN-AGG	5386	12	2	2	0.9974	0.0026	103.7882	0.2475	0.5000	0.9978
21	T-C-L	First Derivative	KNN-SUM	5386	12	2	2	0.9974	0.0026	103.7882	0.2475	0.5000	0.9978
22	T-C-L	First Derivative	INFLO	5373	12	15	2	0.9950	0.0050	103.6629	0.2456	0.1176	0.9978
23	T-C-L	First Derivative	COF	5386	12	2	2	0.9974	0.0026	103.7882	0.2475	0.5000	0.9978
24	T-C-L	First Derivative	RKOF	5372	12	16	2	0.9948	0.0052	103.6533	0.2455	0.1111	0.9978
25	T-C-L	One sided Derivative	HDoutliers	5388	12	0	2	0.9978	0.0022	103.8075	0.2478	1.0000	0.9978
26	T-C-L	One sided Derivative	INFLO	5385	12	3	2	0.9972	0.0028	103.7786	0.2473	0.4000	0.9978
27	T-C-L	One sided Derivative	COF	5386	12	2	2	0.9974	0.0026	103.7882	0.2475	0.5000	0.9978
28	T-C-L	One sided Derivative	LDOF	5385	12	3	2	0.9972	0.0028	103.7786	0.2473	0.4000	0.9978
29	T-C-L	One sided Derivative	LOF	5385	12	3	2	0.9972	0.0028	103.7786	0.2473	0.4000	0.9978
30	T-C-L	One sided Derivative	RKOF	5380	12	8	2	0.9963	0.0037	103.7304	0.2466	0.2000	0.9978
31	T-C-L	First Derivative	LDOF	5388	13	0	1	0.9976	0.0024	73.4030	0.1309	1.0000	0.9976
32	T-C-L	First Derivative	LOF	5388	13	0	1	0.9976	0.0024	73.4030	0.1309	1.0000	0.9976

point as the outlier instead of the actual outlier. However, since this data set contains more spikes than level shift one sided derivatives are recommended to reduce the false positive rate.

Further, as can be seen in Table 2, the distance-based outlier detection algorithms: HDoutliers, KNN-AGG and KNN-SUM, outperform the others. Among the three methods the performance of k nearest neighbour distance-based algorithms are slightly higher in comparison with the HDoutliers algorithm, which is based only on the nearest neighbour distance (Figure 7(d–f)). Although outlying points are clearly separated from their majority, which corresponds to the typical behaviors, the individual outliers are not isolated and are surrounded by the other outlying points (Figure 2 column 1 row 2: the scatter plot between turbidity (*Trans_Tur*) and conductivity (*Trans_Cond*)). Because HDoutliers has the additional requirement of isolation in addition to clear separation between outlying points and typical points, it now shows poor performance in comparison to the two KNN distance-based algorithms (KNN-AGG and KNN-SUM) as they are not restricted to the single most nearest neighbour. For the current work k is set to 10, the maximum default values of k in Madsen (2018), as too large a value of k could skew the focus towards global outliers alone (Zhang, Hutter & Jin 2009) and make the algorithms computationally inefficient. On the other hand, too small a value of k could incorporate an additional assumption of isolation into the algorithm, as in HDoutliers algorithm where $k = 1$. Furthermore, considering level for the detection of outliers in the water-quality sensors does

not greatly affect on the performance. Both T-C-L combinations (combination of turbidity, conductivity and river level) (row 1 to 3 in Table 2) and T-C combinations (combination of turbidity a conductivity) (row 4 to 5 in Table 2) perform similarly well. Not surprisingly, LOF performs the least well (row 32) due to its additional assumption of isolation of outliers which is not satisfied by this data sets as the outlying points are somewhat surrounded by other outliers.

The leading combinations of Table 2 (row 1 to 5) also have the highest PPV. Therefore our proposed framework also maintains a better balance between false positive and false negative rates as shown in Figure 7, 13, 14 and 15.

4.2 Analysis of water-quality data from *in situ* sensors at Pioneer River

The data obtained from Pioneer River were available from 12 March 2017 to 12 March 2018, including 6280 recorded points. Many missing values were observed during the initial part of all the three series: turbidity, conductivity and level (Figure 8) which are highly likely to occur due to technical errors of the sensors. With the help of water-quality experts points were labeled (outlier or not) with the aim of evaluating the performance of the proposed framework. However, the outliers which were labeled as sudden spikes (by water-quality experts) only deviate slightly from the general trend (Figure 9). As expected a negative relationship is clearly visible between the two variables: turbidity and conductivity, even though conductivity shows different relationships for different values of levels.

Since most of the target outliers are masked by the typical points in the original space (as seen in Figure 9) different transformations (as explained in Table 1) were applied with the aim of getting a better separation between the outliers and the typical points. Similar to Sandy Creek, data obtained from the sensors at Pioneer River also give a better separation under the one sided derivative transformation (Figure 10 and 11). However, the sudden spikes labeled as outliers (by water-quality experts) could not be separated from the majority with a large distance gap and were only visible as a small group (micro cluster) in the boundary defined by the typical points. The two large spikes observed in the transformed series of conductivity (Figure 11(b)) correspond to the sudden drops in the original conductivity series (Figure 8(b)).

From the performance analysis (Table 3) it is observed that turbidity and conductivity together produce better results than when combined with level, which tends to reduce the performance level while increasing the false negative rate. In the proposed framework, the correlation structure between the variables is also taken into account when detecting outliers and no clear relationship is observed between level and the remaining two variables. This could be one

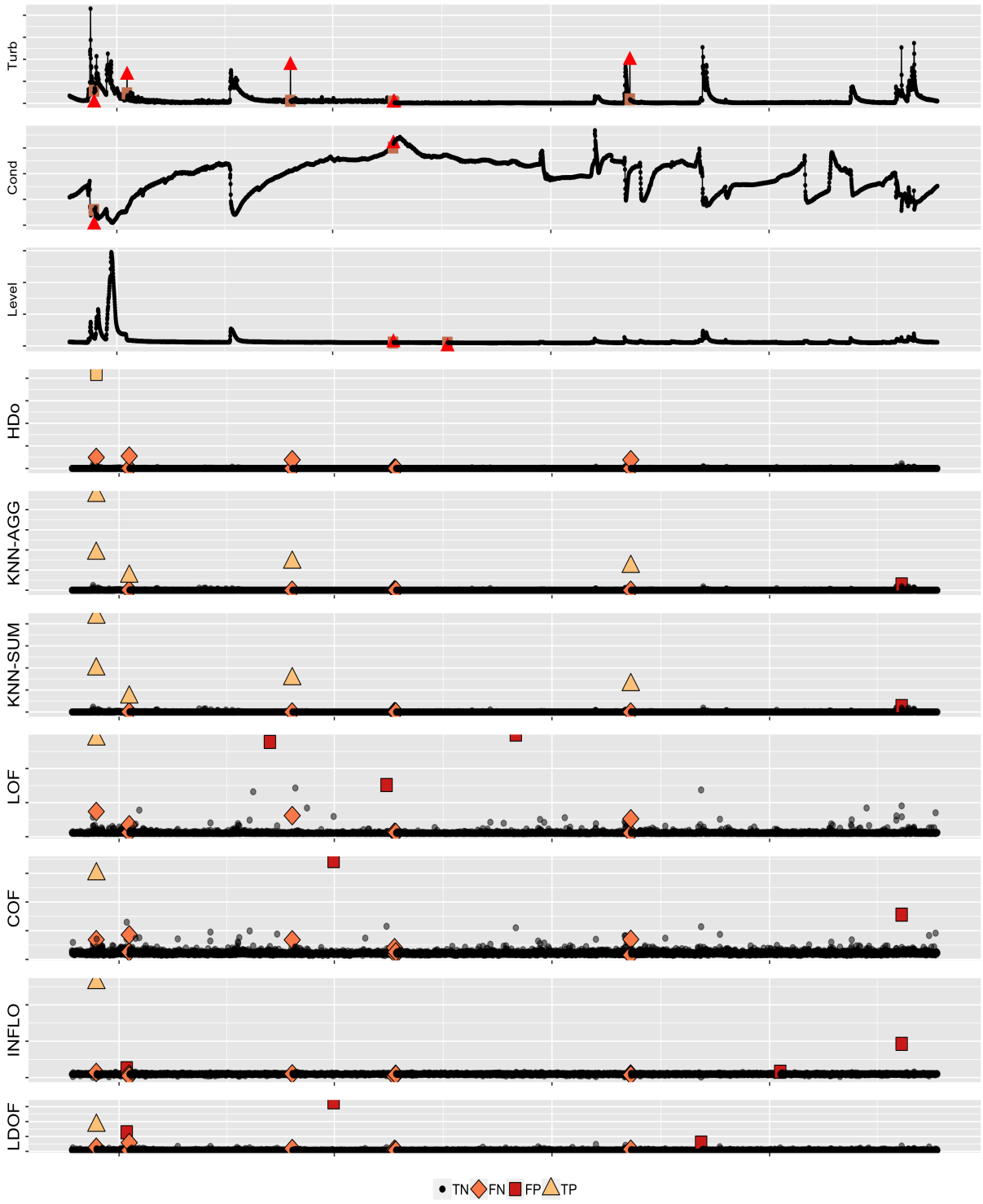

Figure 7: *Classification of outlier scores produced from different algorithms as true negatives (TN), true positives (TP), false negatives (FN), false positives (FP). The top three panels (a, b, c) correspond to the original series (turbidity, conductivity and river level) measured by in situ sensors at Sandy Creek. The target outliers (detected by water-quality experts) are shown in red, while typical points are shown in black. The remaining panels (d–j) give outlier scores produced by different outlier detection algorithms for high dimensional data when applied to the transformed series (one sided derivative) of the three variables: turbidity, conductivity and level.*

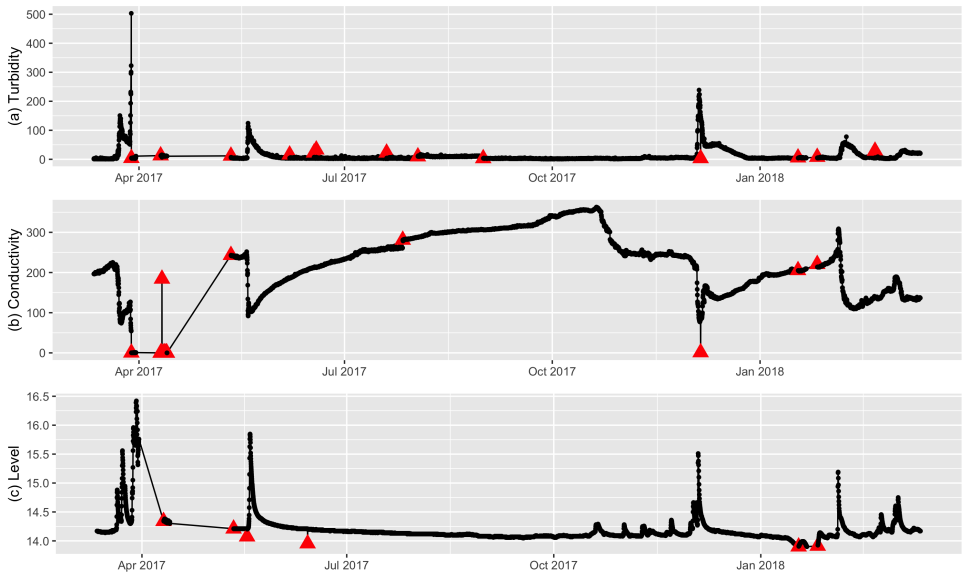

Figure 8: *Time series for turbidity (NTU), conductivity (µS/cm) and river level (m) measured by in situ sensors at Pioneer River. In each plot, outliers determined by water-quality experts are shown in red, while typical points are shown in black.*

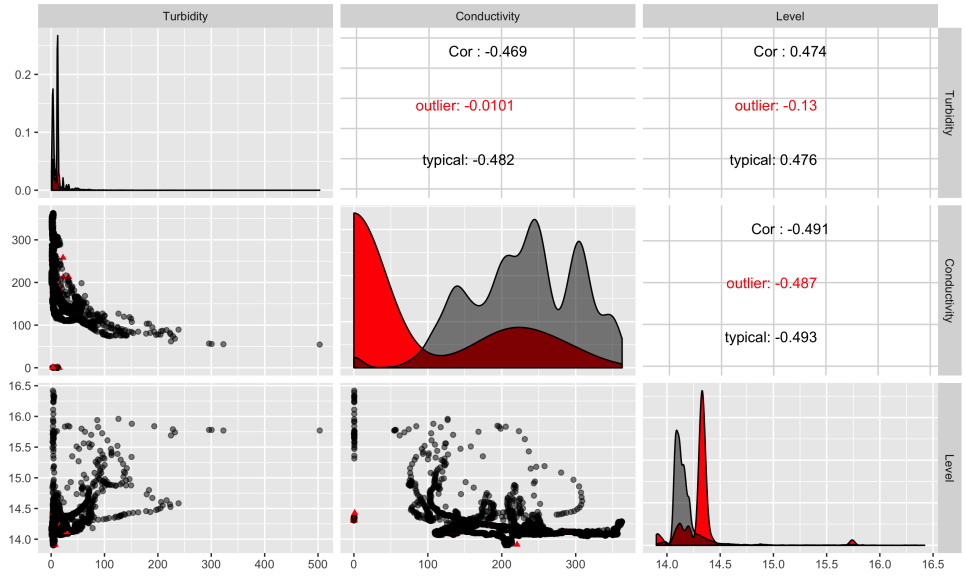

Figure 9: *Bivariate relationships between water-quality variables (turbidity (NTU), conductivity (µS/cm) and river level (m)) measured by in situ sensors at Pioneer River. In each scatter plot, outliers determined by water-quality experts are shown in red, while typical points are shown in black. Diagonal plots show the distribution of individual variables.*

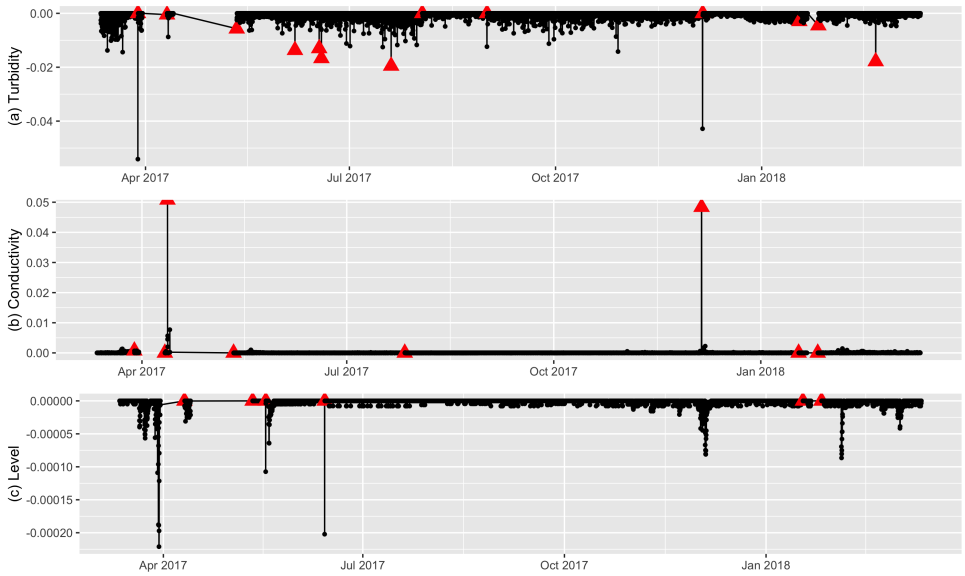

Figure 10: *Transformed series (one sided derivatives) of turbidity (NTU), conductivity (µS/cm) and river level (m) measured by in situ sensors at Pioneer River. In each plot, outliers determined by water-quality experts are shown in red, while typical points are shown in black.*

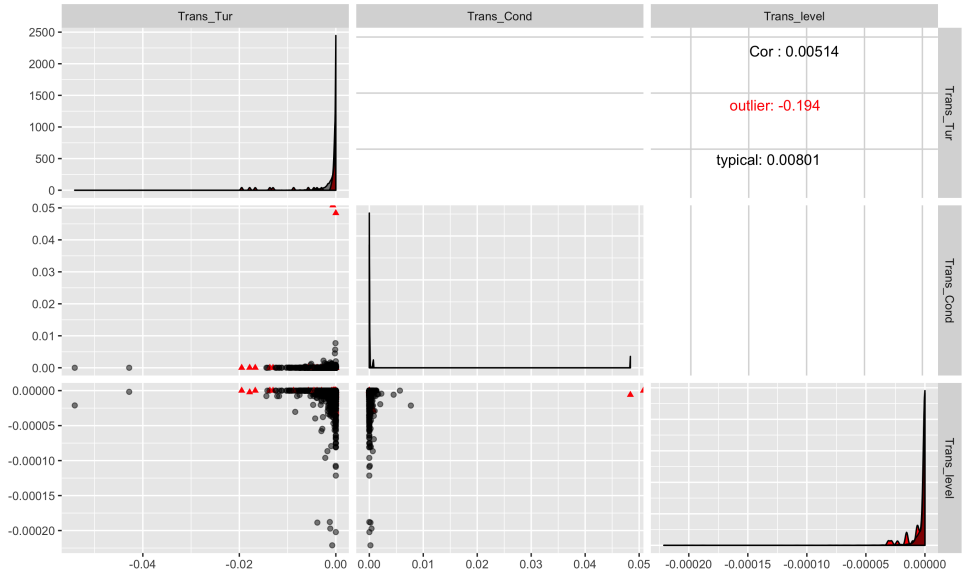

Figure 11: *Bivariate relationships between transformed series (one sided derivatives) of turbidity (NTU), conductivity (µS/cm) and river level (m) measured by in situ sensors at Pioneer River. In each scatter plot, outliers determined by water-quality experts are shown in red, while typical points are shown in black. Diagonal plots show the distribution of individual transformed series.*

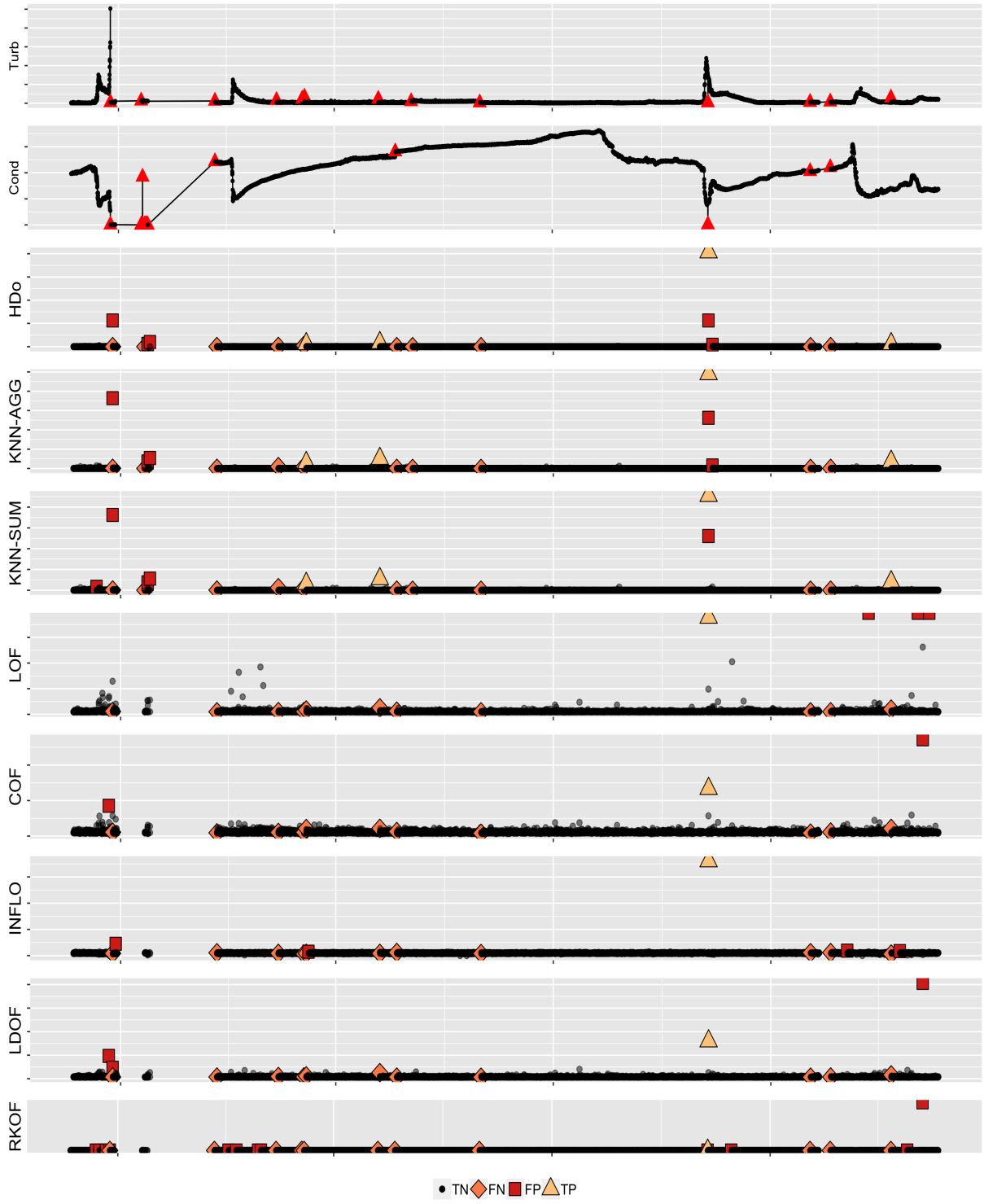

Figure 12: *Classification of outlier scores produced from different algorithms as true negatives (TN), true positives (TP), false negatives (FN), false positives (FP). The top two panels (a and b) correspond to the original series (turbidity and conductivity) measured by in situ sensors at Pioneer River. The target outliers (detected by water-quality experts) are shown in red, while typical points are shown in black. The remaining panels (c–j) give outlier scores produced by different outlier detection algorithms for high dimensional data when applied to the transformed series (one sided derivative) of the two variables: turbidity and conductivity.*

Table 3: *Performance metrics of outlier detection algorithms performed on multivariate water-quality time series data from in situ sensors at Pioneer River. (Rows are arranged in descending order of NPV values)*

i	TimeSeries	Tranformation	Method	TN	FN	FP	TP	Accuracy	Error_Rate	Geometric_mean	Optimised_Precision	PPV	NPV
1	**T-C**	**One sided Derivative**	**HDoutliers**	**6224**	**10**	**7**	**39**	**0.9973**	**0.0027**	**492.6825**	**0.8842**	**0.8478**	**0.9984**
2	**T-C**	**One sided Derivative**	**KNN-AGG**	**6225**	**10**	**6**	**39**	**0.9975**	**0.0025**	**492.7220**	**0.8843**	**0.8667**	**0.9984**
3	**T-C**	**One sided Derivative**	**KNN-SUM**	**6225**	**10**	**6**	**39**	**0.9975**	**0.0025**	**492.7220**	**0.8843**	**0.8667**	**0.9984**
4	**T-C**	**First Derivative**	**HDoutliers**	**6229**	**11**	**2**	**38**	**0.9979**	**0.0021**	**486.5203**	**0.8717**	**0.9500**	**0.9982**
5	**T-C**	**First Derivative**	**KNN-AGG**	**6229**	**11**	**2**	**38**	**0.9979**	**0.0021**	**486.5203**	**0.8717**	**0.9500**	**0.9982**
6	T-C	First Derivative	KNN-SUM	6229	11	2	38	0.9979	0.0021	486.5203	0.8717	0.9500	0.9982
7	T-C	First Derivative	LOF	6228	12	3	37	0.9976	0.0024	480.0375	0.8583	0.9250	0.9981
8	T-C	First Derivative	RKOF	6223	12	8	37	0.9968	0.0032	479.8448	0.8579	0.8222	0.9981
9	T-C-L	One sided Derivative	KNN-AGG	6223	12	6	39	0.9971	0.0029	492.6429	0.8643	0.8667	0.9981
10	T-C-L	One sided Derivative	KNN-SUM	6223	12	6	39	0.9971	0.0029	492.6429	0.8643	0.8667	0.9981
11	T-C-L	First Derivative	KNN-AGG	6227	13	2	38	0.9976	0.0024	486.4422	0.8517	0.9500	0.9979
12	T-C-L	First Derivative	KNN-SUM	6227	13	2	38	0.9976	0.0024	486.4422	0.8517	0.9500	0.9979
13	T-C-L	First Derivative	RKOF	6210	13	19	38	0.9949	0.0051	485.7777	0.8503	0.6667	0.9979
14	T-C	One sided Derivative	INFLO	6227	13	4	36	0.9973	0.0027	473.4681	0.8447	0.9000	0.9979
15	T-C	One sided Derivative	COF	6229	13	2	36	0.9976	0.0024	473.5441	0.8448	0.9474	0.9979
16	T-C	One sided Derivative	LDOF	6228	13	3	36	0.9975	0.0025	473.5061	0.8447	0.9231	0.9979
17	T-C	One sided Derivative	LOF	6228	13	3	36	0.9975	0.0025	473.5061	0.8447	0.9231	0.9979
18	T-C	One sided Derivative	RKOF	6218	13	13	36	0.9959	0.0041	473.1258	0.8439	0.7347	0.9979
19	T-C-L	First Derivative	HDoutliers	6227	14	2	37	0.9975	0.0025	479.9990	0.8385	0.9487	0.9978
20	T-C	First Derivative	INFLO	6223	14	8	35	0.9965	0.0035	466.6958	0.8305	0.8140	0.9978
21	T-C	First Derivative	COF	6229	14	2	35	0.9975	0.0025	466.9208	0.8309	0.9459	0.9978
22	T-C	First Derivative	LDOF	6229	14	2	35	0.9975	0.0025	466.9208	0.8309	0.9459	0.9978
23	T-C-L	One sided Derivative	HDoutliers	6227	15	2	36	0.9973	0.0027	473.4681	0.8250	0.9474	0.9976
24	T-C-L	One sided Derivative	INFLO	6226	15	3	36	0.9971	0.0029	473.4300	0.8250	0.9231	0.9976
25	T-C-L	One sided Derivative	COF	6227	15	2	36	0.9973	0.0027	473.4681	0.8250	0.9474	0.9976
26	T-C-L	One sided Derivative	LDOF	6227	15	2	36	0.9973	0.0027	473.4681	0.8250	0.9474	0.9976
27	T-C-L	One sided Derivative	RKOF	6214	15	15	36	0.9952	0.0048	472.9736	0.8240	0.7059	0.9976
28	T-C-L	First Derivative	INFLO	6229	16	0	35	0.9975	0.0025	466.9208	0.8114	1.0000	0.9974
29	T-C-L	First Derivative	COF	6227	16	2	35	0.9971	0.0029	466.8458	0.8112	0.9459	0.9974
30	T-C-L	First Derivative	LDOF	6227	16	2	35	0.9971	0.0029	466.8458	0.8112	0.9459	0.9974
31	T-C-L	First Derivative	LOF	6229	16	0	35	0.9975	0.0025	466.9208	0.8114	1.0000	0.9974
32	T-C-L	One sided Derivative	LOF	6223	16	6	35	0.9965	0.0035	466.6958	0.8109	0.8537	0.9974

reason for this combination of variables to give poor results. Otherwise the results are somewhat similar to those of Sandy Creek. The three methods KNN-AGG, KNN-SUM and HDoutliers stand out from the other methods (Table 3 row 1–5) with highest accuracy (0.9993), lowest error rates (0.0007), highest geometric means and highest optimized precision. Despite the challenge given by the small spikes which could not clearly separated from the typical points, the three methods: KNN-AGG, KNN-SUM and HDoutliers with one sided derivatives of turbidity and conductivity still detected some of those points as outliers while maintaining a low false negative and false positive rate as shown in Figure 12, 16, 17 and 18.

5 R package `oddwater`

The proposed framework explained in Section 3 is implemented in the open source R package `oddwater`. It provides a growing list of transformation methods and outlier detection methods for high dimensional data together with visualization and performance evaluation techniques. Version 0.1.0 of the package was used for the results presented herein and is available from Github (github.com/pridiltal/oddwater). The datasets used for this article are also available in the package `oddwater`.

6 Conclusion

This paper proposes a methodology for the detection of outliers in water-quality data from *in situ* sensors. Here outliers are specifically defined as due to technical errors that make the data unreliable and untrustworthy. The fundamental aim of the proposed framework is to transform the data space produced by the original series into a space such that outlying points and typical points are clearly separated. Outlier detection algorithms for high dimensional data can then be applied to the new space generated by the transformed series. For the current work we selected transformation methods mainly focused on identifying sudden isolated spikes, sudden isolated drops, and level shifts. Further work is recommended to explore the ability of other transformations to capture other anomaly types such as high and low variability and drift. Our preliminary analysis using data obtained from *in situ* sensors positioned at two study sites, Sandy Creek and Pioneer River, reveals that the proposed framework can perform well with outlier types such as sudden isolated spikes, sudden isolated drops, and level shifts while maintaining low false detection rates. Since our proposed framework is unsupervised, it is easily extended to other water-quality variables, other sites and also to other outlier detection tasks in other application domains. The only requirement is to select suitable transformation methods according to the *data features* that differentiate the outlying instances from the typical behaviours of a given system.

7 Acknowledgements

Funding for this project was provided by the Queensland Department of Environment and Science (DES) and the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS).

Appendix A

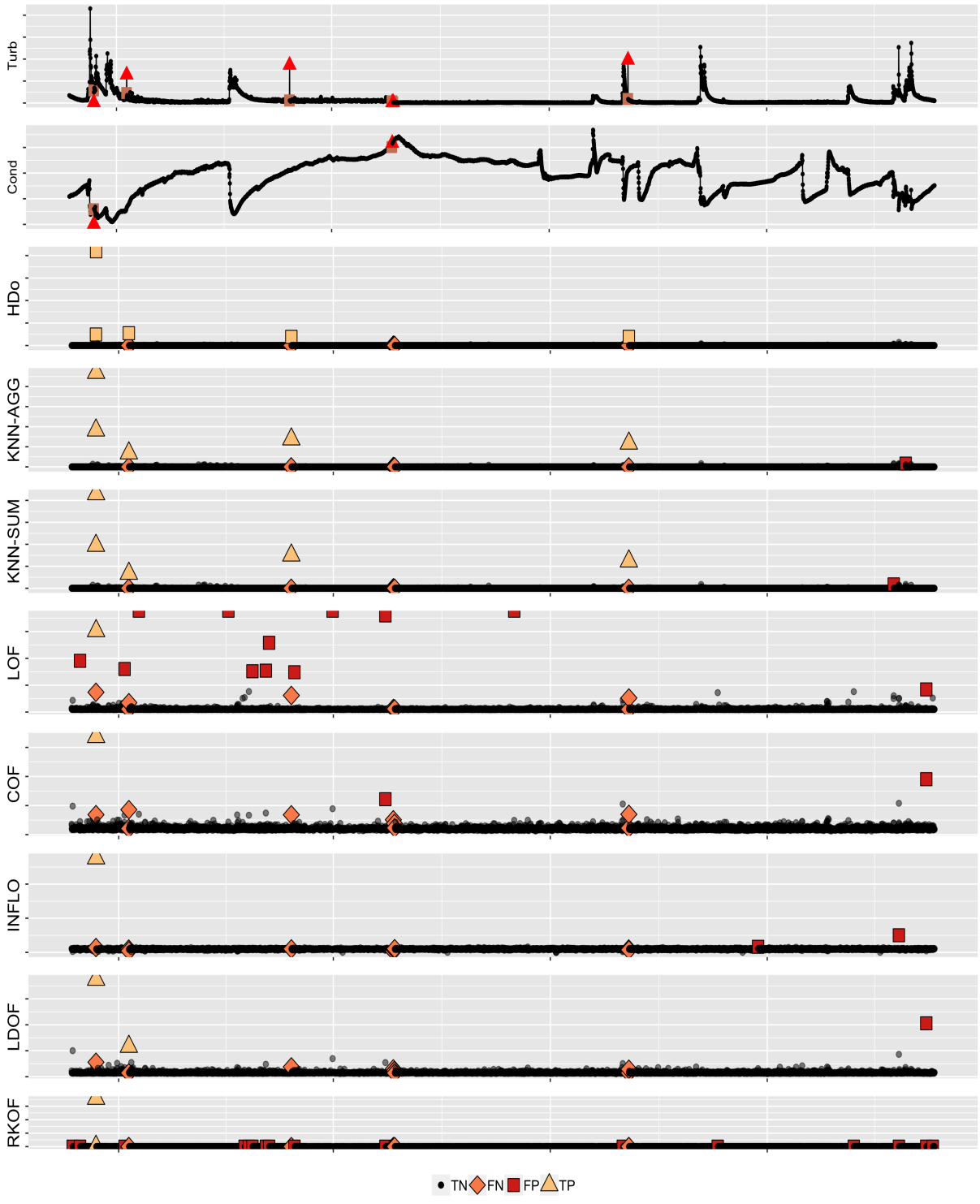

Figure 13: *Classification of outlier scores produced from different algorithms as true negatives (TN), true positives (TP), false negatives (FN), false positives (FP). The top three panels (a and b) correspond to the original series (turbidity and conductivity) measured by in situ sensors at Sandy Creek. The target outliers (detected by water-quality experts) are shown in red, while typical points are shown in black. The remaining panels (c–j) give outlier scores produced by different outlier detection algorithms for high dimensional data when applied to the transformed series (one sided derivative) of the two variables: turbidity and conductivity.*

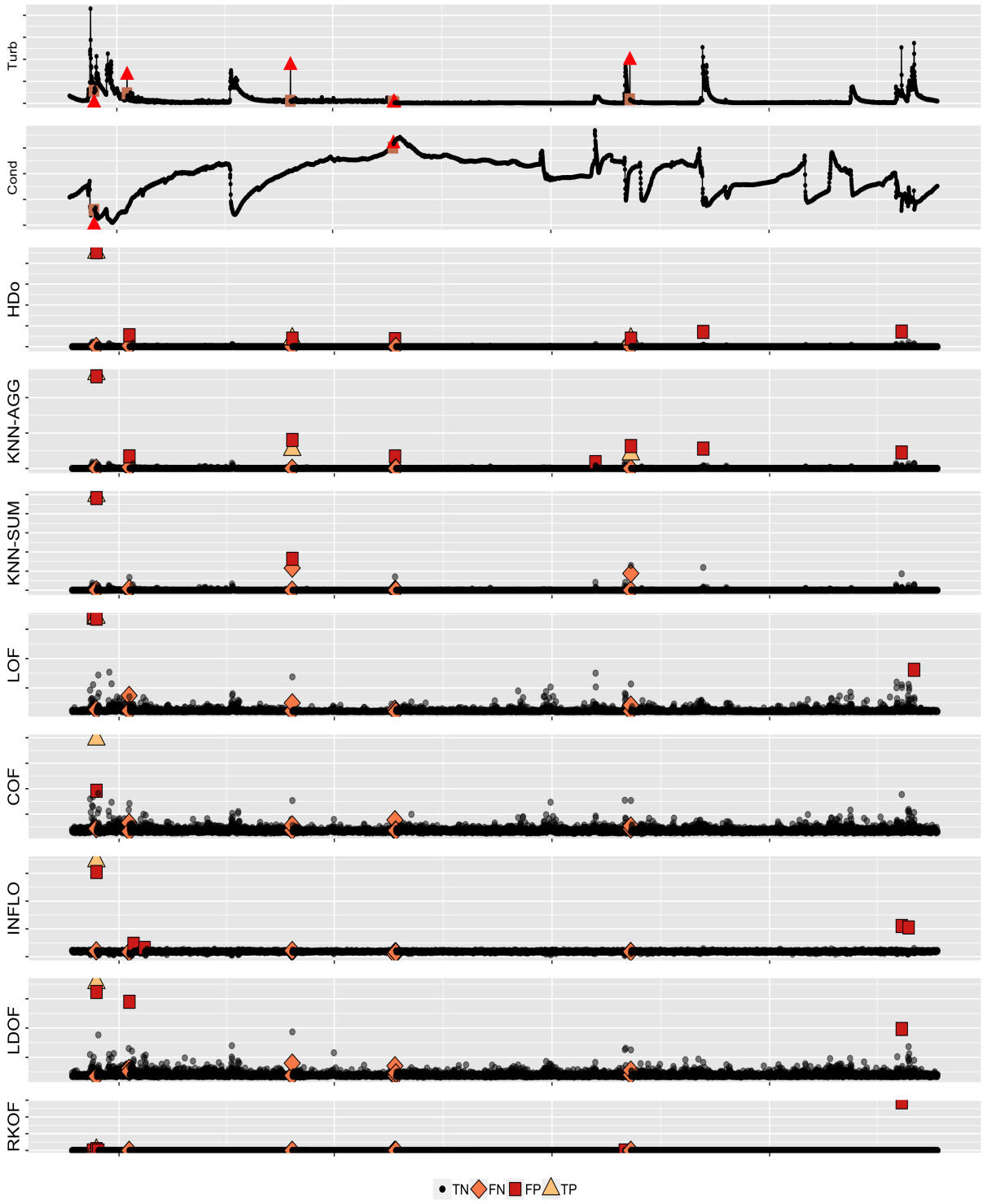

Figure 14: *Classification of outlier scores produced from different algorithms as true negatives (TN), true positives (TP), false negatives (FN), false positives (FP). The top three panels (a and b) correspond to the original series (turbidity and conductivity) measured by in situ sensors at Sandy Creek. The target outliers (detected by water-quality experts) are shown in red, while typical points are shown in black. The remaining panels (c–j) give outlier scores produced by different outlier detection algorithms for high dimensional data when applied to the transformed series (first derivative) of the two variables: turbidity and conductivity.*

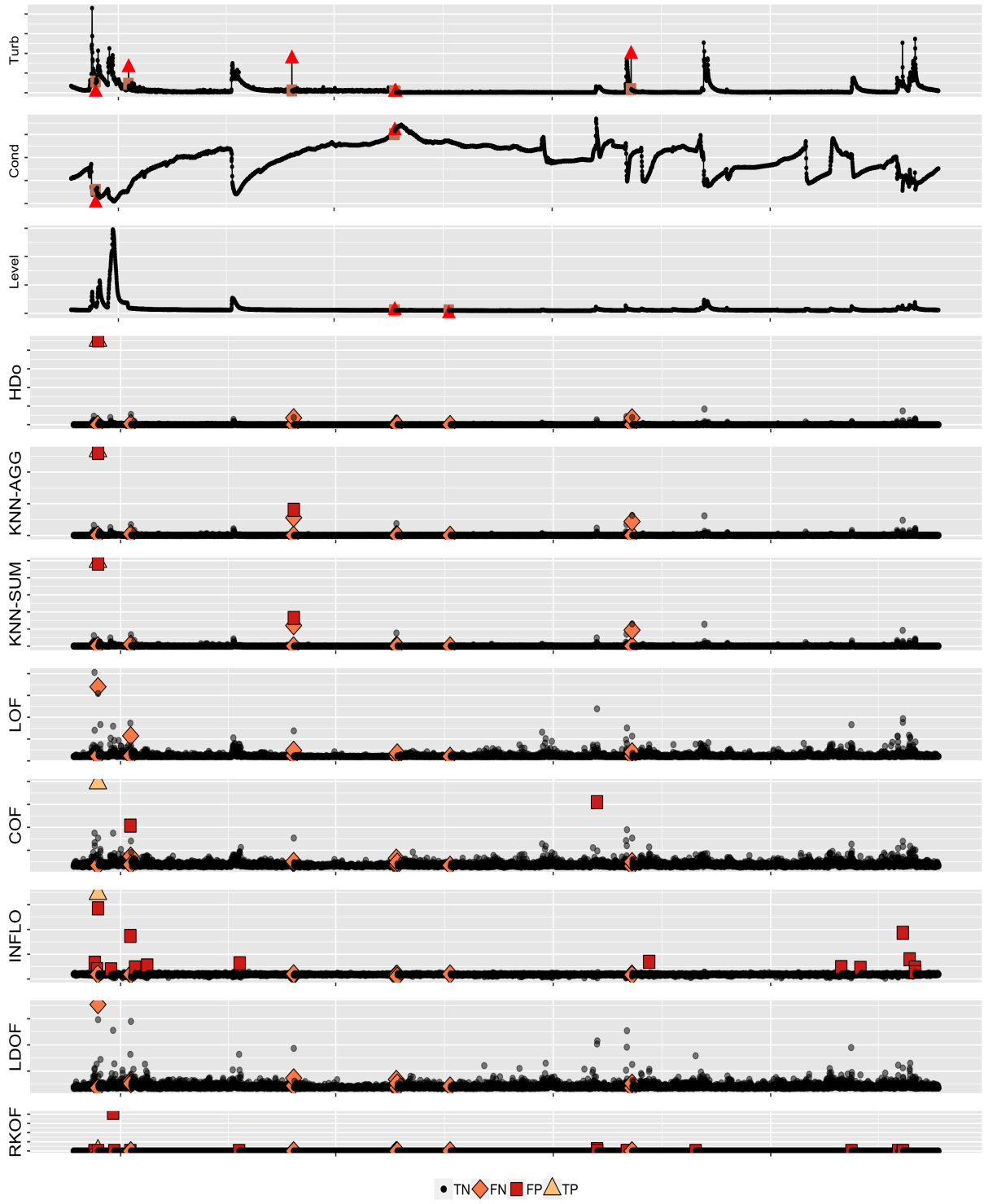

Figure 15: *Classification of outlier scores produced from different algorithms as true negatives (TN), true positives (TP), false negatives (FN), false positives (FP). The top three panels (a, b, c) correspond to the original series (turbidity, conductivity and level) measured by in situ sensors at Sandy Creek. The target outliers (detected by water-quality experts) are shown in red, while typical points are shown in black. The remaining panels (d–k) give outlier scores produced by different outlier detection algorithms for high dimensional data when applied to the transformed series (first derivative) of the three variables: turbidity, conductivity and level.*

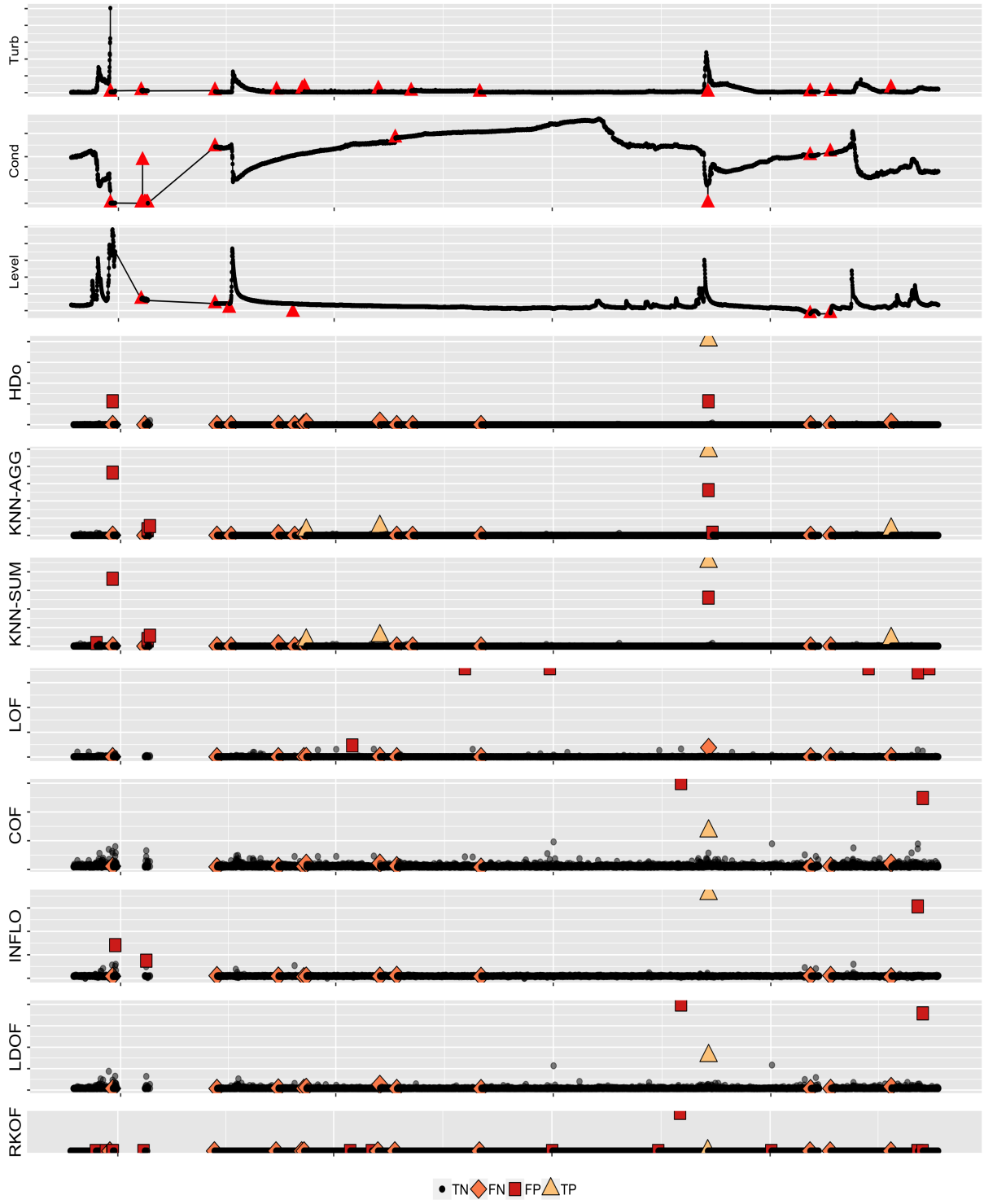

Figure 16: *Classification of outlier scores produced from different algorithms as true negatives (TN), true positives (TP), false negatives (FN), false positives (FP). The top three panels (a, b, c) correspond to the original series (turbidity, conductivity and level) measured by in situ sensors at Pioneer River. The target outliers (detected by water-quality experts) are shown in red, while typical points are shown in black. The remaining panels (d–k) give outlier scores produced by different outlier detection algorithms for high dimensional data when applied to the transformed series (one sided derivative) of the three variables: turbidity, conductivity and level.*

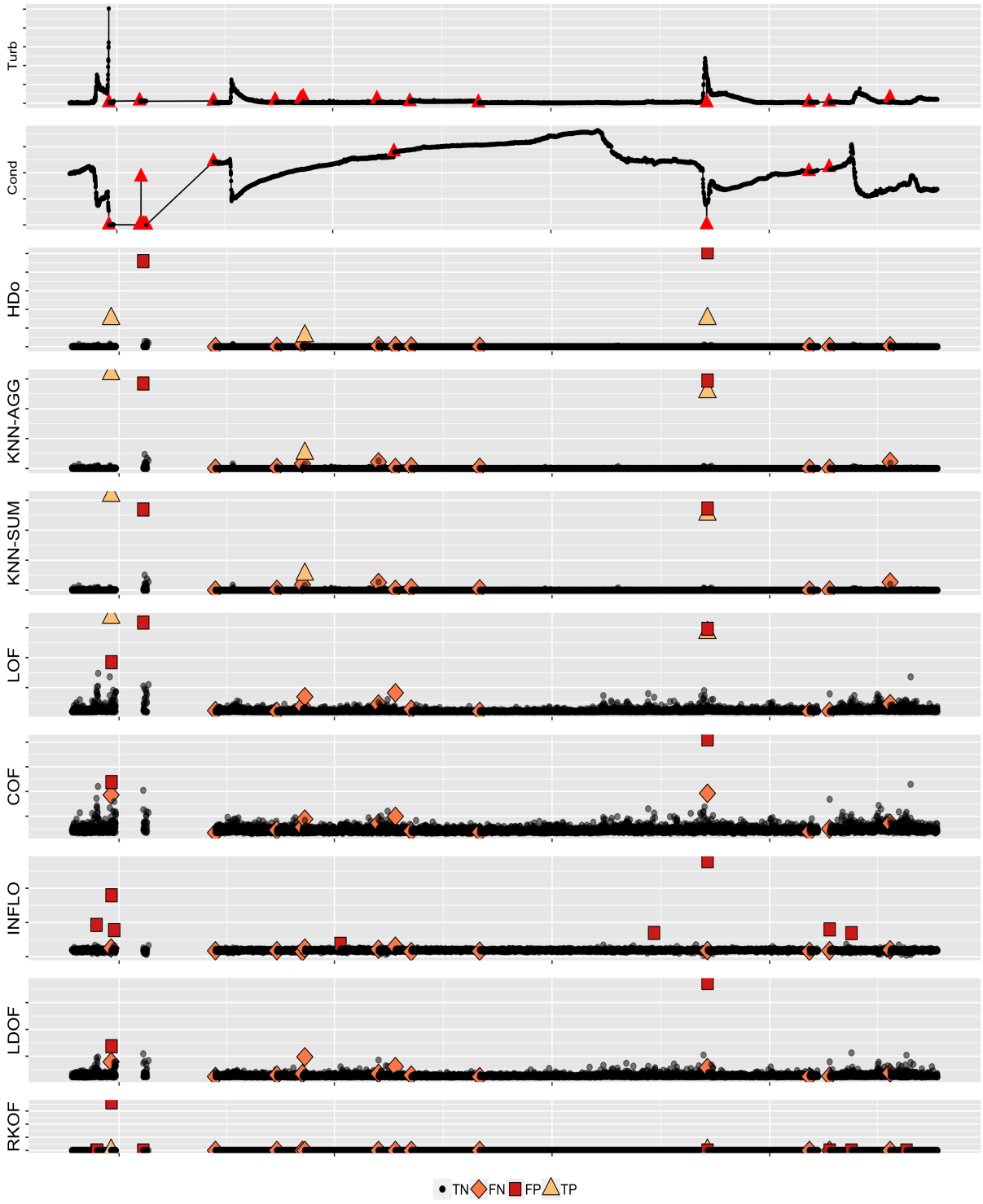

Figure 17: *Classification of outlier scores produced from different algorithms as true negatives (TN), true positives (TP), false negatives (FN), false positives (FP). The top two panels (a and b) correspond to the original series (turbidity and conductivity) measured by in situ sensors at Pioneer River. The target outliers (detected by water-quality experts) are shown in red, while typical points are shown in black. The remaining panels (c–j) give outlier scores produced by different outlier detection algorithms for high dimensional data when applied to the transformed series (first derivative) of the two variables: turbidity and conductivity.*

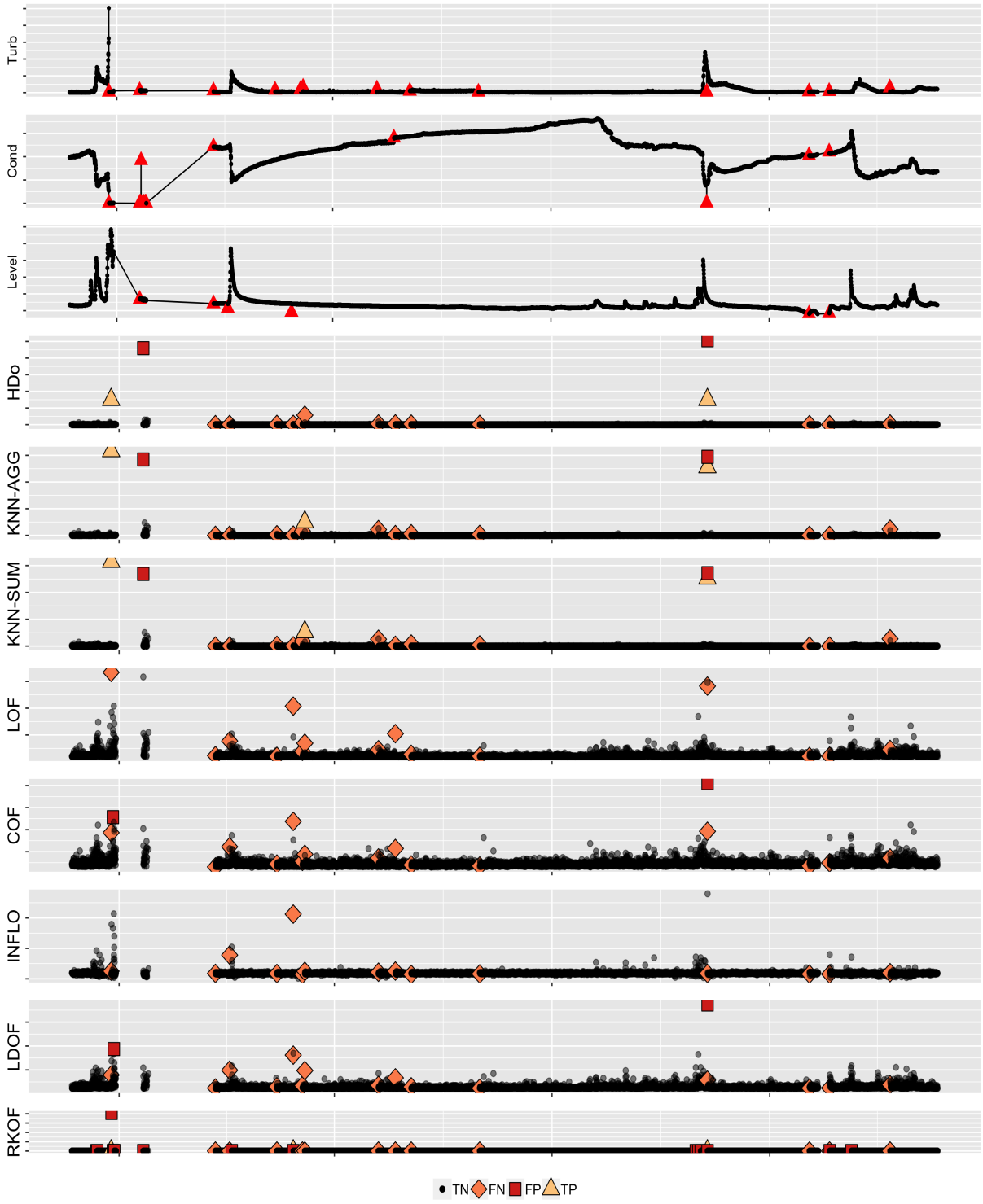

Figure 18: *Classification of outlier scores produced from different algorithms as true negatives (TN), true positives (TP), false negatives (FN), false positives (FP). The top three panels (a, b, c) correspond to the original series (turbidity, conductivity and level) measured by in situ sensors at Pioneer River. The target outliers (detected by water-quality experts) are shown in red, while typical points are shown in black. The remaining panels (d–k) give outlier scores produced by different outlier detection algorithms for high dimensional data when applied to the transformed series (first derivative) of the three variables: turbidity, conductivity and level.*

References

Angiulli, F & C Pizzuti (2002). Fast outlier detection in high dimensional spaces. In: *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, pp.15–27.

Archer, C, A Baptista & TK Leen (2003). Fault detection for salinity sensors in the Columbia estuary. *Water Resources Research* **39**(3).

Ba, A & SA McKenna (2015). Water quality monitoring with online change-point detection methods. *Journal of Hydroinformatics* **17**(1), 7–19.

Bentley, JL (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM* **18**(9), 509–517.

Breunig, MM, HP Kriegel, RT Ng & J Sander (2000). LOF: identifying density-based local outliers. In: *ACM sigmod record*. Vol. 29. 2. ACM, pp.93–104.

Burridge, P & AMR Taylor (2006). Additive Outlier Detection Via Extreme-Value Theory. *Journal of Time Series Analysis* **27**(5), 685–701.

Clifton, DA, S Hugueny & L Tarassenko (2011). Novelty detection with multivariate extreme value statistics. *Journal of signal processing systems* **65**(3), 371–389.

Embrechts, P, C Klüppelberg & T Mikosch (2013). *Modelling Extremal Events: for Insurance and Finance*. Vol. 33. Springer.

Faria, ER, IJ Gonçalves, AC de Carvalho & J Gama (2016). Novelty detection in data streams. *Artificial Intelligence Review* **45**(2), 235–269.

Galambos, J, J Lechner & E Simiu (2013). *Extreme Value Theory and Applications: Proceedings of the Conference on Extreme Value Theory and Applications, Volume 1 Gaithersburg Maryland 1993*. Extreme Value Theory and Applications: Proceedings of the Conference on Extreme Value Theory and Applications, Gaithersburg, Maryland, 1993. Springer US. https://books.google.com.au/books?id=XMPkBwAAQBAJ.

Gao, J, W Hu, ZM Zhang, X Zhang & O Wu (2011). RKOF: robust kernel-based local outlier detection. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp.270–283.

Glasgow, HB, JM Burkholder, RE Reed, AJ Lewitus & JE Kleinman (2004). Real-time remote monitoring of water quality: a review of current applications, and advancements in sensor, telemetry, and computing technologies. *Journal of Experimental Marine Biology and Ecology* **300**(1-2), 409–448.

Goldstein, M & S Uchida (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one* **11**(4), e0152173.

Grubbs, FE (1969). Procedures for detecting outlying observations in samples. *Technometrics* **11**(1), 1–21.

Gupta, M, J Gao, C Aggarwal & J Han (2014). Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering* **26**(9), 2250–2267.

Hadi, AS (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 761–771.

Hartigan, JA (1975). Clustering algorithms.

Hill, DJ & BS Minsker (2006). Automated fault detection for in-situ environmental sensors. In: *Proceedings of the 7th International Conference on Hydroinformatics*.

Hill, DJ, BS Minsker & E Amir (2009). Real-time Bayesian anomaly detection in streaming environmental data. *Water Resources Research* **45**(4).

Horsburgh, JS, SL Reeder, AS Jones & J Meline (2015). Open source software for visualization and quality control of continuous hydrologic and water quality sensor data. *Environmental Modelling & Software* **70**, 32–44.

Hossin, M & M Sulaiman (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process* **5**(2), 1.

Hugueny, S (2013). "Novelty detection with extreme value theory in vital-sign monitoring". PhD thesis. University of Oxford.

Hugueny, S, L Tarassenko & DA Clifton (2008). Extreme Value Theory for Novelty Detection in Vital Sign Monitoring.

Hyndman, RJ (1996). Computing and graphing highest density regions. *The American Statistician* **50**(2), 120–126.

Hyndman, RJ, E Wang & N Laptev (2015). Large-scale unusual time series detection. In: *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. IEEE, pp.1616–1619.

Jin, W, AK Tung, J Han & W Wang (2006). Ranking outliers using symmetric neighborhood relationship. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp.577–593.

Keith, LH, W Crummett, J Deegan, RA Libby, JK Taylor & G Wentler (1983). Principles of environmental analysis. *Analytical chemistry* **55**(14), 2210–2218.

Kotamäki, N, S Thessler, J Koskiaho, AO Hannukkala, H Huitu, T Huttula, J Havento & M Järvenpää (2009). Wireless in-situ sensor network for agriculture and water monitoring on a river basin scale in southern Finland: Evaluation from a data user's perspective. *Sensors* **9**(4), 2862–2883.

Kumar, D, JC Bezdek, S Rajasegarar, M Palaniswami, C Leckie, J Chan & J Gubbi (2016). Adaptive cluster tendency visualization and anomaly detection for streaming data. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **11**(2), 24.

Madsen, JH (2018). *DDoutlier: Distance and Density-Based Outlier Detection*. R package version 0.1.0. https://CRAN.R-project.org/package=DDoutlier.

Mersmann, O (2018). *microbenchmark: Accurate Timing Functions*. R package version 1.4-4. https://CRAN.R-project.org/package=microbenchmark.

Moatar, F, J Miquel & A Poirel (2001). A quality-control method for physical and chemical monitoring data. Application to dissolved oxygen levels in the river Loire (France). *Journal of Hydrology* **252**(1-4), 25–36.

Moatar, F, F Fessant & A Poirel (1999). pH modelling by neural networks. Application of control and validation data series in the Middle Loire river. *Ecological Modelling* **120**(2-3), 141–156.

Panguluri, S, G Meiners, J Hall & J Szabo (2009). Distribution system water quality monitoring: Sensor technology evaluation methodology and results. *US Environ. Protection Agency, Washington, DC, USA, Tech. Rep. EPA/600/R-09/076* **2772**.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Raciti, M, J Cucurull & S Nadjm-Tehrani (2012). "Anomaly detection in water management systems". In: *Critical infrastructure protection*. Springer, pp.98–119.

Ranawana, R & V Palade (2006). Optimized Precision-A new measure for classifier performance evaluation. In: *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on*. IEEE, pp.2254–2261.

Rangeti, I, B Dzwairo, GJ Barratt & FA Otieno (2015). "Validity and Errors in Water Quality Data—A Review". In: *Research and Practices in Water Quality*. InTech.

Schwarz, KT (2008). *Wind dispersion of carbon dioxide leaking from underground sequestration, and outlier detection in eddy covariance data using extreme value theory*. ProQuest.

Sokolova, M & G Lapalme (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management* **45**(4), 427–437.

Storey, MV, B Van der Gaag & BP Burns (2011). Advances in on-line drinking water quality monitoring and early warning systems. *Water research* **45**(2), 741–747.

Talagala, PD & RJ Hyndman (2018). *oddwater: A package for Outlier Detection in water quality sensor data*. https://github.com/pridiltal/oddwater.

Talagala, P, R Hyndman, K Smith-Miles, S Kandanaarachchi, M Munoz, et al. (2018). *Anomaly detection in streaming nonstationary temporal data*. Tech. rep. Monash University, Department of Econometrics and Business Statistics.

Tang, J, Z Chen, AWC Fu & DW Cheung (2002). Enhancing effectiveness of outlier detections for low density patterns. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp.535–548.

Thottan, M & C Ji (2003). Anomaly detection in IP networks. *IEEE Transactions on signal processing* **51**(8), 2191–2204.

Tutmez, B, Z Hatipoglu & U Kaymak (2006). Modelling electrical conductivity of groundwater using an adaptive neuro-fuzzy inference system. *Computers & geosciences* **32**(4), 421–433.

Weissman, I (1978). Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association* **73**(364), 812–815.

Wilkinson, L (2018). Visualizing Big Data Outliers through Distributed Aggregation. *IEEE transactions on visualization and computer graphics* **24**(1), 256–266.

Zhang, K, M Hutter & H Jin (2009). A new local distance-based outlier detection approach for scattered real-world data. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp.813–822.