Outlier Detection Test Best Practices

Document Control Number: DWR-1-BST-005

Document Type: BST

Version: 2.0

Status: Effective

Effective Date: 1/19/2026

California Department of Water Resources

Division of Integrated Science and Engineering

3500 Industrial Boulevard

West Sacramento, California 95691

Contents

Authorship

Authors (alphabetical): David Bosworth[1], Michael Koohafkan[2], Amanda Maguire[3], Cat Pien[4], John Franco Saraceno[1], Patrick Scott[3], Ted Swift[1], Daniel Wisheropp[5]

[1] Division of Integrated Science and Engineering.

[2] U.S. Army Corps of Engineers.

[3] California Department of Fish and Wildlife.

[4] U.S. Bureau of Reclamation.

[5] Division of Operations and Maintenance, Headquarters.

Acronyms & Abbreviations

CDEC — California Data Exchange Center

DISE — Division of Integrated Science and Engineering

DWR — California Department of Water Resources

IOOS — U.S. Integrated Ocean Observing System

NaN or **NAN** — Not a Number

ODWG — Outlier Detection Working Group

QA — Quality Assurance

QC — Quality Control

SME(s) — Subject Matter Expert(s)

SOP(s) — Standard Operating Procedure(s)

WQA — Water Quality Assessment (Unit in DISE)

Scope and application

This document applies to Department of Water Resources (DWR) staff involved in the collection and review of water quality data. DWR's Quality Assurance Program outlines general considerations and has recommended procedures applicable to data review and quality control of discrete and continuous water quality data. The goal of this document is to establish a robust and consistent department-wide set of recommendations for outlier detection practices for water quality associated data. Consistent and robust procedures in data collection and review are of fundamental importance because they support development of data-driven management decisions. Following consistent and well-documented procedures supports DWR's fundamental principles of providing high quality data collected with integrity.

The elements outlined in this document are intended to be used in developing program-specific data review and Quality Control (QC) standard procedures for the projects and programs within DWR that collect applicable data. The scope of this document is limited to guidance on outlier detection practices for water quality data. This document will not address field methods guidance, nor most laboratory quality control data, and is limited to methods from data recording to reporting. This document should be used in combination with other data handling guidance, such as "Discrete Water Quality Data Review Best Practices," DWR-1-BST-001.

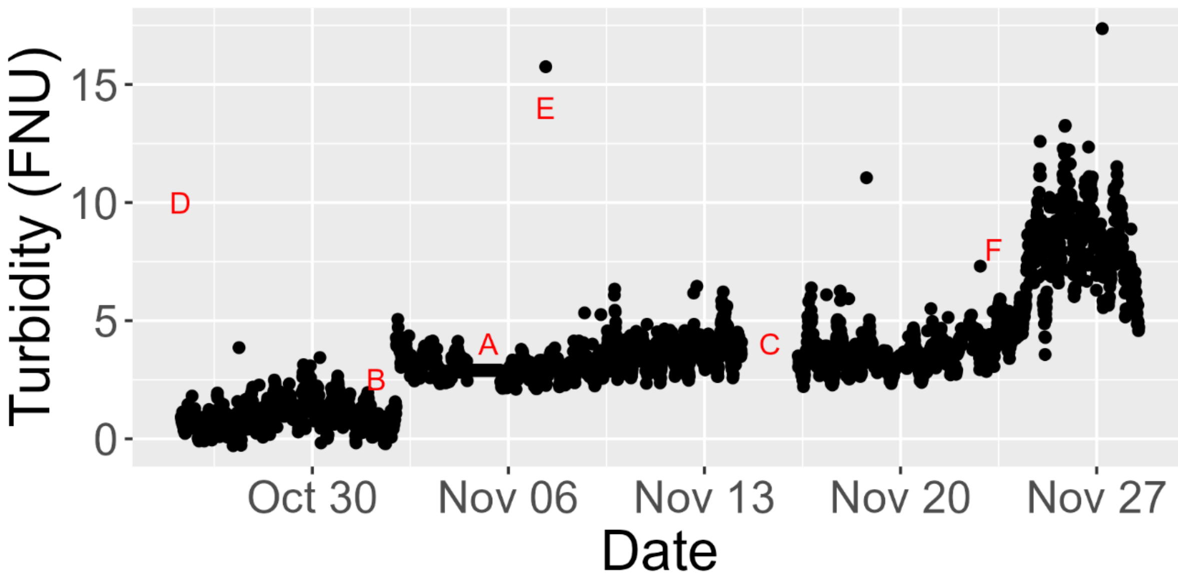
Introduction and Background

If one looks up the definition of "outlier," one may be surprised to find that it doesn't necessarily mean "incorrect" or "erroneous," but merely something unusually far from the distribution of other data. In the field of environmental monitoring and ecological research, as in most scientific fields, it is the unusual (but valid) observations that are often the most valuable.

> The most exciting phrase to hear in science, the one that heralds new discoveries, is not "Eureka!" (I found it!) but "That's funny…"
>
> — Isaac Asimov

The central challenge of outlier detection is to accurately identify and subsequently flag data within a dataset that are "wrong," i.e., not representative of the variable being measured, while not removing "correct" data that are merely unusual (Figure 1). We seek to minimize the number of "wrong" data that we fail to identify, called *false negatives* (Type-II error), while also minimizing the number of "correct" data flagged as "wrong," called *false positives* (Type-I error).

Figure 1. Examples of many of the kinds of anomalies typically seen in a high frequency time series. Specific tests are discussed further in the text.



Timeseries features identified in an initial data review may include tests for a (A) flatline, (B) data shift, (C) gap, (D) gross range (off scale), (E) spike, (F) rate of change.

Outlier detection is a broad field because outliers are caused by a wide variety of mechanisms. Thus, the search for a single “best” algorithm for outlier detection leads to the conclusion that no single approach will catch all outliers, and a layered approach is more appropriate. This document presents a set of phased outlier detection guidelines or steps that can be used as a framework for Standard Operating Procedures (SOPs) and software tools development.

Recommendations

The Outlier Detection Working Group (ODWG) recommends a layered approach to outlier detection and removal, ranked in three groups below, following the framework of the U.S. Integrated Ocean Observing System (IOOS) recommendations discussed in the Literature Review (Appendix 1). The following tests are grouped according to the degree of human supervision required, ease of implementation, and discernment, from “Highly Recommended” (for which there is little debate whether the tests are worthwhile) to “Suggested or Experimental” tests (which may require additional expert involvement, process supervision, and subjective decision-making).

Outlier detection and removal may occur at several stages in the data collection and analysis process. For example, some monitoring groups have implemented unsupervised outlier removal within data-logger software at field stations. These remove data based on a set of rules or criteria set up by a human operator, without human supervision. Data collection programs such as Campbell Scientific's *LoggerNet* can display and flag anomalous data. California Data Exchange Center (CDEC) graphs often represent "raw" data from which outliers can often be identified visually, though the CDEC system doesn't lend itself to data flagging.

Group 1: Highly Recommended Tests

These are tests that check for the most basic data accuracy requirements, such as correct time stamps, data structure, location, and documented sensor and seasonal measurement ranges.

1. Gap/Time Test: Applies to continuous data. Checks that the most recent data point has been measured and received within the expected time window and has the correct time stamp. Operators must be mindful of the time zone, time change (e.g., daylight savings time), and time stamp format. The test checks to see if there is a gap in the data based on an expected time frequency (e.g., 15 minutes, 1 hour, etc.).

2. Syntax Test: Applies to continuous data. Checks that the received data message contains the proper structure without any indicators of flawed transmission such as parity errors, incomplete timestamp or data values, or line format anomalies. For example, the expected number of characters for fixed-length messages equals the number of characters received, or a data logger and multi-sensor sonde are incorrectly set up, recording incorrect parameter order or units. This test may not be necessary given our telemetry systems.

3. Location Test: Checks that the reported present physical location (latitude/longitude) is within operator-determined limits. Coordinates are within an expected geographical area or bounding box. This test is most appropriate to data produced by mobile ships or buoys than fixed monitoring stations, but may be applicable to some data streams, e.g., transects or stations that change in location. It may be appropriate for some kinds of discrete sample data.

4. Gross Range Test: Checks if a data point exceeds sensor minimum or maximum values. All sensors have a limited output range, and this can form the most rudimentary gross range check. Choices of limits in this test are guided by the operational range of the sensor and the physically realistic range of values. For instance, water temperature is not likely to be greater than 35 ^{o}C (with very unusual exceptions worthy of further human investigation), pH will always be in the range 0 to 14, and most physical measurements are greater than or equal to zero. Additionally, the operator

can select a smaller span based upon local knowledge or a desire to draw attention to extreme values (see Test #5, Tuned Gross Range Test).

5. Tuned Gross Range Test (Hydrological/constituent/geographical): A variation on the Gross Range Test (Test #4, where the thresholds are determined and adjusted by the operator on a monthly or seasonal basis, or some other predefined time period. Subject matter expertise is required to determine reasonable seasonal or spatial ranges. Understanding of the historical data through documented analysis will help determine operator-chosen range thresholds. This test can be implemented using a range of approaches, such as a geographic area bounded by polygons on a map of stations within which data ranges are expected to vary.

   Determining Regional/Seasonal Range Values: To establish and document range values, each monitoring group should determine a set of criteria appropriate for their program. Criteria for developing range values may be determined by grouping data seasonally or regionally, evaluating several years of monitoring data, and identifying realistic ranges seen across multiple years and/or decades. The program's SOP document should include a discussion of methods and justification for the choice of range values, as well as guidance on revisiting and documenting changes in constituent ranges. Changes in range values should be documented in program SOPs and/or historical station metadata. Best practice is to develop a stepwise process to avoid subjective "personal views;" e.g., use input from historical reports and/or station history documents with summary statistics by constituent and water year type. Additionally, it's best to include extreme years when establishing range values, such as very wet or dry years, if possible. For new field monitoring sites without a historical data context, the analyst will have to rely on statistical tests of the data themselves.

   Number of Data Points Used to Develop Regional/Seasonal Range Values: For new stations, use the instrument-based Gross Range Test (Test #4. To implement a localized range test, use at least a one-year span of the clean period of record (do not use raw data). For very sparse data, statistical tests will have lower statistical power. The usual rule of thumb is that approximately thirty values are needed to have confidence in summary statistics such as the mean and standard deviation. This pertains to discrete data collected at relatively low frequency (e.g., monthly), but will not be meaningful for continuous data, where a longer period of record is needed to capture realistic variation. For continuous data, thirty days throughout a year or more will be better at capturing variation. If nearby stations are grouped together to establish range values, the rationale for grouping should be documented, as grouping can introduce unintended changes in the data characteristics.

Spatial Tests: Within a watershed or catchment, historical data ranges may provide the only guidance for a new station. Start with the Gross Range Test (Test #4 until there is sufficient data to implement the Tuned Gross Range Test (Test #5.

6. Flat Line Test: Applies to continuous data. This test compares the present observation to several previous observations to identify a continuously repeated observation of the same value, which is a common result when sensors and/or data collection platforms fail. A naturally occurring flat line is an unusual pattern in continuous data; however, some environments may have subtle changes that are below the resolution of the sensor, producing what appear to be flat lines. This may be a reporting resolution issue, and the data reviewer needs to keep that in mind. This test is relatively easy to implement as an unsupervised test.

Group 2: Recommended Tests

Tests for outliers that may pass the tests in Group 1 (Test #1 through Test #6 but are still physically extremely unlikely to be “real”.

Challenges: Environmental data are rarely normally distributed. Often, they are approximately log-normal, with many values in the lower range and relatively few large values. Outlier detection approaches that assume normality therefore need to be used with caution, and near-normality confirmed (e.g., using a quantile-quantile plot (QQ plot) before proceeding further. Data can be transformed (e.g., by taking the log, square root, etc.) to make the data approximately normal. But one must then keep in mind that any summary statistics are in transformed units: For example, for a true log-normal data set, the mean of the log-transformed data is no longer the mean, but the median of the original data set. The standard deviation, and confidence intervals, will be similarly transformed.

7. Z-score and Modified Z-score Tests: The Z-score is calculated as the number of standard deviations between a data point and the mean of neighboring data points. The Z-score test is affected by the presence of the outliers that it is intended to detect. Often outliers are removed in several rounds when implementing the Z-score test, since removing the outlier lowers the standard deviation, changing the Z-score itself. IOOS guidance suggests limiting successive eliminations to three rounds. Additionally, Z-score tests are susceptible to being blinded by large real variation (false negative) or, in a time series, by sudden changes after periods of small variation (false positive). Both of these scenarios should be considered when using this test.

An alternative to the Z-score test is the Modified Z-score test, which is more robust to outliers (Leys et al. 2013). The Modified Z-score is similar to the Z-score but uses the median and MAD (median absolute deviation) in place of the mean and standard deviation, respectively. The MAD is defined as: MAD = median(|Yi – median(Y)|), or the "median of the absolute values of individual value departures from the data set median." A modified Z-score greater than 3.5 is considered an outlier candidate (Gorrie 2016 and Pien 2020). The Modified Z-score test may be more applicable to the type of data seen in environmental monitoring and research, as it is less sensitive to the often log-normal distribution of environmental data.

8. Tukey's Interquartile Range (IQR) Test: Tukey's IQR test is an example of a test that relies on the variation of the data. It defines a range outside of which a data point may be expected to be an outlier: First, the interquartile range (Q3 – Q1) of the dataset is calculated. Then, values outside of 1.5*IQR the first (Q1) or third (Q3) quartile are identified as potential outliers. Extreme outliers can be identified as 3*IQR. This test requires a relatively long data set and may be better for discrete data.

9. Spike Test: Applies to continuous data or frequently sampled discrete data. Checks for single value spikes but requires tuning. The mean or median of adjacent data points is calculated to form a reference. If the data point exceeds the threshold value difference from the adjacent values, it is flagged for further evaluation (Basu and Meckesheimer 2007). The difference may be positive or negative and may be flagged separately. The absolute value of the spike is tested to capture either positive or negative spikes. Multiple thresholds can be used to differentiate between large and small spikes, which may require tuning by the subject matter expert (SME). Since the spike test uses data on either side of the data point being tested, it does not lend itself to incoming real-time data until one or more subsequent data are collected.

10. Rate of Change Test: Applies to continuous data or frequently sampled discrete data. Inspects the time series for a rate of change between the value being tested and the prior value and then compares the difference to a threshold value identified by the operator. A typical rate of change threshold for this test is a specified number of standard deviations of a specified number of prior observations. For example, the threshold could be 3 standard deviations of the data collected 25 hours prior to the value being tested, which could work well for continuous data collected in a tidal environment. Another example is to use a window of prior observations that capture the sudden water quality changes that occur at monitoring stations near pumping plants, for which some kinds of outliers aren't flagged by automated tests.

Tests that rely on neighboring data values require care when there are multiple bad data points, or a single bad data point on one side but not the other (E. Ateljevich pers. comm. 2022). In these cases, most of these tests may still require staff to manually perform visual checks to identify extended periods of erroneous data.

Group 3: Suggested / Experimental Tests

11. Multivariate Test: A variation of the Rate of Change Test (Test #**Error! Reference source not found.**. This test helps reduce false positives. The Rate of Change Test is conducted for one environmental variable with a more restrictive threshold; if this test fails, a second Rate of Change Test operating on a second, related, variable is conducted. For example, a temperature, salinity, or depth sensor can provide a check on another water quality parameter to determine whether a sensor went out of the water during a field deployment. This approach involves the data quality for one sensor being used to evaluate the data quality of other sensors: If several sensors are co-located, such as on a multiparameter sonde, anomalies in one sensor may show up as anomalies in other sensors. For example, submerged vegetation, periphyton, or phytoplankton can cause significant changes in pH during a day-night cycle. In discrete samples, anion and cation ratios often co-vary; an outlier candidate in one analyte can be compared to other analytes. Relationships between constituents should be documented in SOPs and/or station metadata.

12. Attenuated Signal Test: Inspects for a standard deviation or range variation that fails to exceed threshold values over a selected time period. Designed to identify a sensor failure mode where a data series is nearly but not exactly a flat line, e.g., if the sensor were to become wrapped in debris. This test may be challenging to use consistently in practice, as it may be difficult to avoid subjective reasoning when independent evidence is not available.

13. Neighbor Test: This test is the same as the Multivariate Test (Test #11, but a second sensor is used rather than a second environmental variable. Ideally, redundant sensors using different technology would be co-located and alternately serviced at different intervals, but in some cases a nearby (but not co-located) sensor can provide a useful QC check as well. The selected thresholds depend entirely upon the relationship between the two sensors as determined by the local knowledge of the operator. This class of test may apply to data collected in a transect, where data are collected across space, as well as time. The definition of “nearby” is at the discretion of the SME but must be documented in metadata. Care and subject matter expertise is required. Some stations may be geographically close but have very different constituent behavior due to hydrology or other factors. Examples include the Banks and Jones Delta export pumping plants, some channels in the Cache

Slough — Yolo Bypass region, and stream monitoring stations in nearby but separate upstream watersheds. Constituent behavior may vary in time, also, requiring lagging one constituent with respect to the other to develop the best correlation.

14. Curve/Space Test: Originally a test that relied on the relationship between temperature and salinity in the open ocean. The authors note the temperature-salinity test itself will probably not be useful in a tidally mixed estuary. However, if two variables are known to behave together in a predictable way, they can be used as a cross-check to detect outliers. For example, cations and anion ratios vary in predictable ways. The vertical profile of water temperature in a lake should be monotonically increasing or decreasing, due to water density varying with temperature. Another example would be to compare constituent values across space on a map such as applying moving averages to transects, which is similar to the Spike Test (Test #9, but substituting location for time.

    Some environmental variables tend to co-vary for hydrologic or biochemical reasons. This provides the potential of using one measured variable to check another as they vary in time or space. One could use all or a large set of one's data to develop a predictive model (simple two variable regression, multivariable regression, etc.). When one plots the predicted vs. actual data, "true" outliers should fall far from the predictive model line (or surface, for multidimensional models). One would define a maximum error threshold (e.g., a percentage) for outlier detection, based on expected variation, and calculate an error distance between model output and data. Measures such as Cook's Distance can be used to identify individual points that unduly influence the regression model.

    Example approaches from the Sacramento San Joaquin River Delta include, but are not limited to:

    - First-flush events tend to produce pulse increases in turbidity and other analyte concentrations. Knowledge of the flow conditions (discharge, time derivative of discharge) could be used to predict a rapid change in water quality constituents. Predictive models or expert systems could be constructed in such a way as to use an expected rapid change to, for example, relax a Z-score threshold after an extended period of small variance (e.g., organic carbon at river tributaries).

    - Wind-driven sediment resuspension causes turbidity increases. Wind data from a meteorological station near the sensor in question can be used to produce a predictive model that turbidity will increase when windspeeds exceed a threshold (e.g., turbidity at Banks Pumping Plant due to resuspension in Clifton Court Forebay, resuspension in Grizzly Bay).

- Daily cycles of sunlight-driven photosynthesis increase dissolved oxygen concentrations and drive up pH levels as photosynthesis removes CO_2 (as carbonic acid, H_2CO_3) from the water. Nighttime biological respiration tends to decrease pH as CO_2 concentrations increase. A periodic model based on time of day could predict pH levels, which could then be subtracted and reveal outliers (e.g., parameters in the San Joaquin River at Vernalis and Mossdale, at monitoring stations along the South Bay Aqueduct).

- A water channel may discharge past a sensor at a specific tidal stage, producing a distinct signal correlated with that stage level, repeating daily but aligned with the tide time, not time of day. Plotting the signal (e.g., conductivity) versus water stage reveals the pattern that might be hard to diagnose from plotting the signal versus time.

15. Sophisticated tests: In principle, sophisticated tests based on training software to recognize outliers might be implemented in, for instance, neural networks or other forms of Machine Learning (ML) or artificial intelligence methods. As with other tests in this section, these would very likely require careful supervision by subject matter experts during the training phase and afterwards.

Graphical Methods

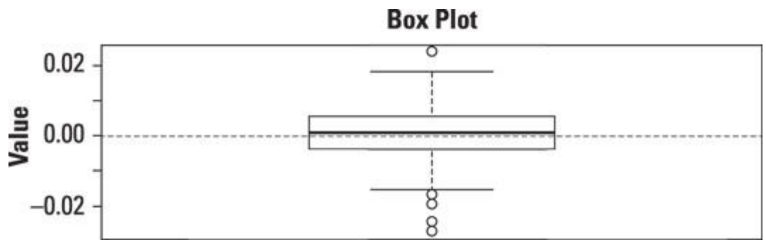
For many of these tests, plotting the data graphically can be a powerful tool. As always, the outlier candidates identified by these graphical methods may be truly invalid data, or they may be unusual but very valuable “real” data. Determining which is which will still require human evaluation.

One can often detect outliers by plotting two co-varying variables against each other. Outliers will appear as points far from the regression line or curve (see also Li 2019, for clustering example). However, this method does not immediately tell you which variable is the outlier. An outlier on a plot of one constituent versus another may be due to an error in either constituent, or both. However, a check against other constituents will typically reveal which was in error (Denton 2015; Hem 1985).

These tests may provide a framework for deciding whether a data point is an outlier but will still rely on the attention and expertise of a SME. While most of the outliers identified in Group 1 can be explained by equipment errors, outliers flagged by the Group 2 and 3 tests should be treated with respect and accepted or rejected only after investigation.

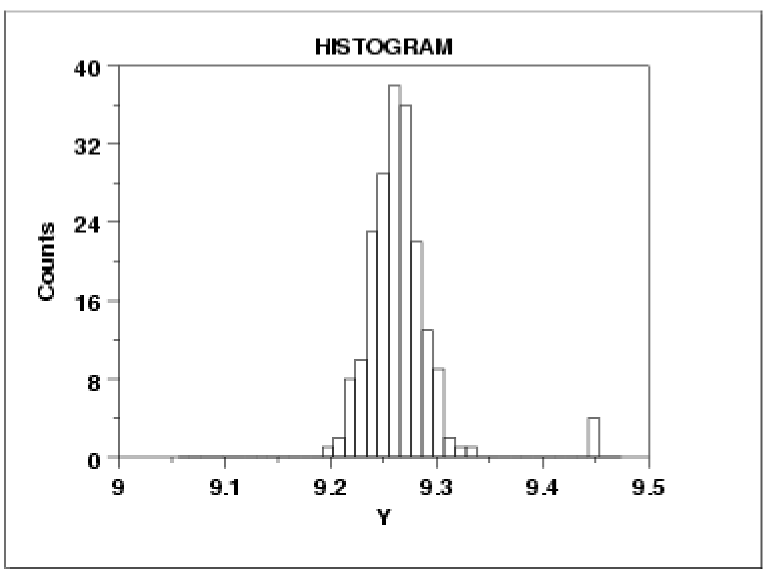
Boxplots and histograms can be useful for summarizing data and identifying outliers (Figure 2); the boxplot effectively presents Tukey's interquartile range (IQR) as the boundaries of the box. Boxplots can often be configured to plot extreme values as individual points outside the "box and whisker" parts of the plot.

Figure 2. Box plot including extreme values plotted as individual points (https://www.dummies.com/article/technology/information-technology/data-science/big-data/graphical-tests-of-data-outliers-141267/)



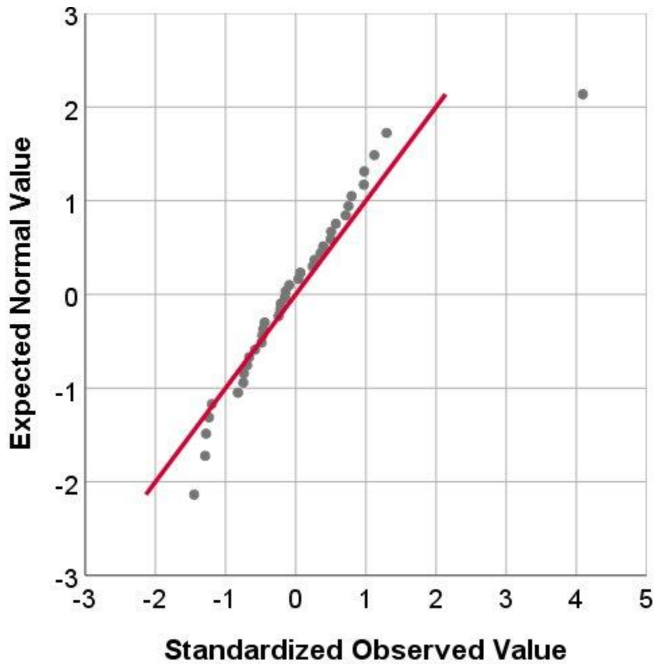
Histograms can highlight data that fall outside the expected distribution (Figure 3). In this illustration, the extreme value is almost certainly not "part of the main population distribution." However, it may represent a sample from a different but valid population, such as an unusual environmental condition.

Figure 3. A histogram of normally-distributed data and an outlier (https://doi.org/10.18434/M32189, www.itl.nist.gov/div898/handbook/eda/section3/histogr8.htm)



A quantile-quantile or "Q-Q" Plot (Figure 4) plots the z-score normalized observed data against a normal distribution. If the data fall approximately on a straight diagonal line, the data are normally distributed. Often, data in environmental sciences will be log-normally distributed (mostly clustered at lower values, with a few high extremes), and the Q-Q plot straight line will instead be a curve. However, in either case, a Q-Q plot can quickly highlight values that don't behave like the other data, as with the top right datapoint in the figure.

Figure 4. Example Q-Q Plot (Elkhrachy 2020)



References

Ahmad S, Lavin A, Purdy S, & Agha Z. 2017. "Unsupervised real-time anomaly detection for streaming data." Neurocomputing, 262, 134–147. [Website.] Viewed online at: https://www.sciencedirect.com/science/article/pii/S0925231217309864.

Ateljevich Eli. 2023. Personal communication, discussion at Outlier Detection Working Group meeting.

Basu S, & Meckesheimer M. 2007. "Automatic outlier detection for time series: An application to sensor data." Knowledge and Information Systems, 11(2), 137–154. [Website.] Viewed online at: https://link.springer.com/article/10.1007/s10115-006-0026-6.

Cho HY, Oh JH, Kim KO, & Shim JS. 2013. "Outlier detection and missing data filling methods for coastal water temperature data." In D.C. Conley, G. Masselink, P.E. Russell, & T.J. O'Hare (Eds.), Proceedings 12th International Coastal Symposium (Plymouth, England). Journal of Coastal Research, Special Issue No. 65, 1898-1903. [Website.] Viewed online at: https://www.jstor.org/stable/26491068.

Denton RA. 2015. *Delta salinity constituent analysis*. State Water Project Contractors Authority. [Website.] Viewed online at: https://emp.baydeltalive.com/assets/588ee18bdb51ef1619ac6fd28b97f694/application/pdf/Denton_2015_Delta_Salinity_Constituents_Report.pdf.

Elkhrachy, I. 2020. "Modeling and Visualization of Three-Dimensional Objects Using Low-Cost Terrestrial Photogrammetry." International Journal of Architectural Heritage. VOL. 14, NO. 10, 1456–1467. [Website.] Viewed online at: https://doi.org/10.1080/15583058.2019.1613454.

Giannoni F, Mancini M, & Marinelli F. 2018. *Anomaly detection models for IoT time series data*. [Website.] Viewed online at: arXiv preprint arXiv:1812.00890.

Gorrie C. 2016. *Three ways to detect outliers*. [Website.] Viewed online at: http://colingorrie.github.io/outlier-detection.html.

Guivetchi K. 1986. Salinity unit conversion equations. DWR Interoffice Memo.

Gupta M, Gao J, Aggarwal C, & Han J. 2014. *Outlier detection for temporal data: A survey*. IEEE Trans. Knowl. Data Eng., 26(9), 2250–2267. [Website.] Viewed online at: https://ieeexplore.ieee.org/document/6684530.

Gupta M, Gao J, Aggarwal C, & Han J. 2014. “Outlier detection for temporal data.” Synthesis Lectures on Data Mining and Knowledge Discovery, 5(1), 1–129. [Website.] Viewed online at: https://link.springer.com/book/10.1007/978-3-031-01905-0.

Hem JD. 1985. Study and interpretation of the chemical characteristics of natural water. U.S. Geological Survey Water-Supply Paper 2254. [Website.] Viewed online at: https://pubs.usgs.gov/wsp/wsp2254/html/pdf.html.

Hodge V, & Austin J. 2004. A survey of outlier detection methodologies. Artif. Intell. Rev., 22(2), 85–126. [Website.] Viewed online at: https://www.researchgate.net/publication/220638052_A_Survey_of_Outlier_Detection_Methodologies.

Integrated Ocean Observing System (IOOS). 2016. *Manual for real-time quality control of in-situ temperature and salinity data version 2.0: A guide to quality control and quality assurance of in-situ temperature and salinity observations*. [Website.] Viewed online at: https://www.researchgate.net/publication/321979246_Manual_for_Real-Time_Quality_Control_of_In-situ_Temperature_and_Salinity_Data_A_Guide_to_Quality_Control_and_Quality_Assurance_for_In-situ_Temperature_and_Salinity_Observations.

Komsta L. 2006. “Processing data for outliers.” R News, 6(2), 10–13. [Website.] Viewed online at: https://journal.r-project.org/articles/RN-2006-009/RN-2006-009.pdf.

Leys C, Ley C, Klein O, Bernard P, & Licata L. 2013. “Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median.” Journal of Experimental Social Psychology, 49(4), 764–766. [Website.] Viewed online at: https://doi.org/10.1016/j.jesp.2013.03.013.

Li S. 2019. *Time series of price anomaly detection: Anomaly detection detects data points in data that does not fit well with the rest of the data*. [Website.] Viewed online at: https://towardsdatascience.com/time-series-of-price-anomaly-detection-13586cd5ff46.

Montoya B. 2004. *Factors affecting the composition and salinity of exports from the south Sacramento-San Joaquin Delta*. DWR.

NIST/SEMATECH *e-Handbook of Statistical Methods*. [Website.] Viewed online at: http://www.itl.nist.gov/div898/handbook/. https://doi.org/10.18434/M32189.

Peterson Thomas C, R Vose, R Schmoyer, and V Razuvaev. 1998. “Global historical climatology network (GHCN) quality control of monthly temperature data.” International Journal of Climatology, 18, 1169-1179. [Website.] Viewed online at: https://rmets.onlinelibrary.wiley.com/doi/10.1002/%28SICI%291097-0088%28199809%2918%3A11%3C1169%3A%3AAID-JOC309%3E3.0.CO%3B2-U.

Pien Catarina. 2020. *Outlier Detection Tests*. PowerPoint presentation to DWR QA Outlier Detection Working Group.

Rebbapragada U, Protopapas P, Brodley CE, & Alcock C. 2009. “Finding anomalous periodic time series: An application to catalogs of periodic variable stars.” Machine Learning, 74(3), 281–313. [Website.] Viewed online at: https://link.springer.com/article/10.1007/s10994-008-5093-3.

Suits B. 2001. *DWR office memo, relationships between EC, chloride, and bromide at Delta export locations*.

Suits B. 2002. “Chapter 5, Relationships between Delta water quality constituents as derived from grab samples.” In DWR’s *Methodology for flow and salinity estimates in the Sacramento-San Joaquin Delta and Suisun Marsh*. 23rd Annual Progress Report, June 2002. [Website.] Viewed online at: https://data.cnra.ca.gov/dataset/methodology-for-flow-and-salinity-estimates-in-the-sacramento-san-joaquin-delta-and-suisun-marsh/resource/59aad468-6823-4d31-aea5-b1846ddb4ae0.

Tripathy SS, Saxena RK, & Gupta PK. 2013. “Comparison of statistical methods for outlier detection in proficiency testing data on analysis of lead in aqueous solution.” American Journal of Theoretical and Applied Statistics, 2(6), 233-242. [Website.] Viewed online at: https://sciencepublishinggroup.com/article/10.11648/j.ajtas.20130206.21.

Yu Y, Zhu Y, Li S, & Wan D. 2014. *Time series outlier detection based on sliding window prediction*. Mathematical problems in Engineering. [Website.] Viewed online at: https://doi.org/10.1155/2014/879736.

Appendix 1. Literature Review

Introduction

Outlier detection is a broad field driven by the fact that outliers are caused by a wide variety of mechanisms. An outlier is an observation that lies an abnormal distance from other values in the dataset (NIST). Thus, the search for a single "best" algorithm for outlier detection leads to the conclusion that no single approach will catch everything, and that a layered approach is needed. This document contains summaries of some pertinent literature and resources that we reviewed and includes pros and cons of the outlier detection methods discussed in each reference.

Outliers can be investigated from the standpoint of several different "dimensions": They can be investigated through comparisons in time (individual measurements in a time series), in space (comparing measurements from neighboring locations), or comparing constituents that are expected to co-vary in a well-understood way. Outliers can be spatial, where an individual data instance can be considered anomalous with respect to data from geographically nearby locations. An outlier can also be temporal, or contextual, if the temporal sequence of data is relevant, i.e., a data point is anomalous with respect to data before and/or after it, nearby in time. A data point can also be considered anomalous with respect to other constituents that usually co-vary, such as chemical ions, temperature, and dissolved oxygen, etc. If one has the luxury of having data available from geographically nearby stations, these can potentially be used to detect instrument malfunction, drift, etc.

Within temporal outlier detection, one can attempt to detect outliers in real-time, or within a "historical" time series. Detecting outliers in real time is more challenging (IOOS QARTOD 2016), since one is missing the temporal context provided by later data points. The data analyst can't tell whether the latest data point is an outlier (bad data), or the first valid data point of a rapidly changing variable (e.g., a pulse of increased concentration during a "first flush" rainfall event).

Summaries of Literature

Below, organized alphabetically by first author, are journal articles identified as being applicable to environmental monitoring outlier detection, and briefly summarized.

Ahmad Subutai, Lavin A, Purdy S, & Agha Z. 2017. Unsupervised real-time anomaly detection for streaming data. Neurocomputing, 262, 134–147.

Defines the ideal characteristics of a real-world anomaly detection algorithm. Describes a fully automated, on-line, machine learning based anomaly detector using a sequence memory algorithm called Hierarchical Temporal Memory. Provides an open-source implementation and a novel benchmark for performance evaluation. Seeks to work with real-time, rather than batch data.

Basu S, & Meckesheimer M. 2007. Automatic outlier detection for time series: An application to sensor data. Knowledge and Information Systems, 11(2), 137–154.

Focuses on a method using the median from nearby data values in a time series to detect possible outliers. This is a good paper, since the method seems simple, easy to implement, and works well on data sets that are difficult to model.

Eli Ateljevich and Amanda Maguire evaluated this paper. Eli wrote 3 Nov 2022, in part:

“I'll start small and specific about the Basu technique. My use of Basu is extensive, and it’s kind of a lead catch-all technique but I have modified the way the thresholding is set in order to make it a bit more automatic (in Basu I think they use judgment). It does as well as anything I've tested but is merely the least imperfect choice. I've long since integrated it with other checks so I'm not doing a lot of isolated experiments at the moment. [The modeling group’s] suite of tools also includes thresholding (based on regional reasonability bounds), Basu detection, some stuff based on tides/wavelets, tests for repeats on things that shouldn't repeat like water levels, etc.). There are also a lot of things I've tested that are more frequency-based but they tend to be limited to water levels and flow. We have tried many other approaches as well.

We have a big repository. Twice a week, we obtain a large collection of surface water in the Bay-Delta, including the period of record from NCRO, all the USGS stations, NOAA, the DICE stations (still labeled "des" though) from their sources and USBR through CDEC. We store them in a big csv repository. We have fast python/dask methods of retrieval, and the total size is about 1000 time series. There is a reconciliation of the various agency quirks and concepts ... simple things like "station" and "variable" and "sublocation." We have pretty robust readers for people’s data. We tend to apply agency flags and treat them as nans (not-a-number), so in some cases like NCRO that might cull some values.

This is our "raw" download. We save it for consistency on our own, but it goes immediately into a "formatted" mode that further reformats it all, so it is faster to read and attaches metadata using a particular set of formats. The end result is a file that at least in Python can be trivially parsed with default reader settings.

After that, it enters the "screen" stage which remains in development and is the major reason for my interest in the outlier detection effort. Any flagging we do is considered a flag and doesn't throw away data.

In any event, models break if data is too weird so it may be quite a while before we'd think we can avoid hand checking certain data and making sure they make sense in the context of the models."

Cho HY, Oh JH, Kim KO and Shim JS. 2013. Outlier detection and missing data filling methods for coastal water temperature data, In: Conley DC, Masselink G, Russell PE, and O'Hare TJ. (eds.). Proceedings 12th International Coastal Symposium (Plymouth, England), Journal of Coastal Research, Special Issue No. 65, pp. 1898-1903, ISSN 0749-0208.

Relied on the z-score approach applied to residuals after removal of periodic components identified by harmonic analysis. This may be very useful in the Bay-Delta, though developing the tidal periodic signal will take effort. There is also the risk that interesting valid signals will be treated as "wrong" data outliers.

Denton Richard A. 2015. Delta Salinity Constituent Analysis. Prepared for the State Water Project Contractors Authority. February 2015.

Dr Richard Denton is a water quality modeler, formerly with the Contra Costa Water Agency. As a consultant working for the State Water Project Contractors Authority through DWR's Municipal Water Quality Investigations, he performed an exhaustive regression analysis of salinity constituents (ion data) in the Delta, building on earlier work by Kamyar Guivetchi of DWR and several other studies. Very useful in untangling variations due to source water and seasonal patterns in Delta water quality data.

In this report, historic grab samples collected by DWR and USBR were used to develop regression relationships between the key indicators of salinity and water quality in the Delta, i.e., specific conductance (EC) and total dissolved solids (TDS), as well as chloride, bromide, sodium, calcium, sulfate, magnesium, potassium, hardness, and alkalinity. Denton's own QA/QC procedures for the data in this report followed the suggestions of Hem (1985). The relationships established by Denton allow readers to implement the accuracy/outlier checks defined in Hem (1985) to a more precise degree for Sacramento-San Joaquin River Delta waters.

Denton implemented a series of checks:

Tests 1-4 can all be completed by importing lab results and field readings into an anion/cation balance calculator and defining criteria for flagging suspect data. As a UC Davis professor of geochemistry once lectured, “The anions and cations had better sum to zero, or you would have lightning bolts shooting out of your water sample.”

Test 5 (Comparison to a Pre-Defined Relationship) — Input a field reading for specific conductance into the established regression to calculate a predicted result for the corresponding water quality sample, then compare this to the actual lab result. Flagging criteria can be established based on model accuracy compared to historic results or known accuracy of analyses. This approach is used by Denton (2015) but, as he points out, one needs to investigate further whether the error is in one lab result, the other, or both.

Test 5 is especially important to several DWR monitoring activities. Many of our stations have very wide ranges for the water quality constituents/parameters we monitor. Water year type, tidal, seasonal, and operational variations all have sizeable effects on the water quality in the Interior and Western Delta. Often the datasets have a wide range and are heavily skewed with long tails (log-normally distributed). This is due to the somewhat regular occurrence of short-lived, extreme events in this region. With a proven model, the residuals between modeled and determined results should fall into a normal distribution with a small(er) range, which allows us to flag erroneous data that wouldn’t be considered an outlier when compared only to historic results.

Although this literature is focused on conductivity and ions, it may also be useful for other parameters that experience similar distributions, e.g., Total Suspended Solids (TSS) compared to turbidity, or Chlorophyll-a compared to a Total Algae Sensor, Dissolved organic Carbon (DOC or TOC) to an fDOM sensor. Those regressions would have to be developed and tested.

There are also several DWR studies that have examined the relationship between TDS, conductivity, and ionic composition in the Delta. Notably, in chronological order:

Guivetchi Kamyar. 1986. DWR Interoffice Memo, Salinity unit conversion equations.

Suits Bob. 2001. DWR Office Memo, Relationships between EC, chloride, and bromide at Delta export locations.

Suits Bob. 2002. Chapter 5, Relationships between Delta Water Quality Constituents as derived from Grab Samples. In DWR’s “Methodology for Flow and Salinity Estimates in the Sacramento-San Joaquin Delta and Suisun Marsh." 23rd Annual Progress Report, June 2002

Montoya Barry. 2004. DWR Report, Factors affecting the composition and salinity of exports from the south Sacramento-San Joaquin Delta.

Giannoni F, Mancini M, and Marinelli F. 2018. Anomaly detection models for IoT time series data. arXiv preprint: 1812.00890. https://doi.org/10.48550/arXiv.1812.00890.

Explored multiple methods for outlier detection in ammonia concentration time series data for the purpose of detecting any potential anomaly and then using the event frequency as a flag to detect sensor failure. Tested methods included Running Average Low-High Pass Filters, Univariate Gaussian Predictors, Seasonal Extreme Studentized Deviate (S-ESD) algorithms, and Local Density Cluster-Based Outlier Factor (LDCOF) algorithms. Pseudo-code was provided for each algorithm. The low-high pass filter is highly customizable and required several tests to understand which configuration worked best; large window sizes were necessary to identify anomalous trends, but small window sizes were needed to identify accurately sporadic outliers. The authors concluded that this method was sub-optimal due to the tuning requirements. The univariate Gaussian predictor has limited application as it requires mono-dimensional data that fits a Gaussian distribution. The authors noted that tuning the algorithm to identify smaller synthetic anomalies generated many more false positives, which can be explained by the fact that the probability of each measurement is computed in a context-less fashion, which potentially is too great a simplifying assumption. The Seasonal-ESD algorithm proved to be the best algorithm to identify sporadic outliers in the tested datasets but performed poorly when used to identify anomalous trends. The authors concluded that S-ESD is incapable of detecting slow sensor drift, and thus is not adequate for identifying anomalous trends. LDCOF was the only algorithm examined that could work with multivariate data, and the numerous hyper-parameters required makes it highly customizable and very difficult to tune. Once tuned, LCDOF proved to be able to identify both anomalous trends and single outliers with average performance. The authors concluded that LCDOF is sub-optimal due to the significant tuning required to obtain reliable performance. The authors concluded that no single method could reliably detect both sporadic outliers and anomalous trends and recommended applying multiple algorithms; results could then be combined using either a majority vote or weights-based approach to classify anomalous data.

Gupta M, J Gao, C Aggarwal, and J Han. 2014. Outlier Detection for Temporal Data: A Survey. IEEE Trans. Knowl. Data Eng. 26, 9 (2014), 2250–2267 and **Gupta, M., J. Gao, C. Aggarwal, and J. Han. 2014. Outlier Detection for Temporal Data. Synthesis Lectures on Data Mining and Knowledge Discovery, March 2014, Vol. 5, No. 1, Pages 1-129. Morgan & Claypool Publishers. 1–129 pages.**

A good review paper of common and modern outlier detection algorithms. Needs further review. The second reference is a longer version of the journal article from the same year.

Hem John D. 1985. Study and interpretation of the chemical characteristics of natural water: U.S. Geological Survey Water-Supply Paper 2254, 264 p. https://pubs.usgs.gov/wsp/wsp2254/html/pdf.html.

A thorough evaluation on detecting outliers using anion-cation balance budgeting. Referenced by Denton (2015). This is a dense paper that explores the chemical characteristics and composition of natural water. The Section "Evaluation of Water Analysis" (pg. 163) includes tests for evaluating water quality samples for accuracy and outliers, specifically in regard to ions. The following are the recommended tests and checks for sample accuracy:

- Mass Balances: The sum of anions and cations should approximately equal TDS (in mg/L). Another procedure for checking analytical accuracy that is sometimes useful is to compare determined and calculated values for dissolved solids. The two values should agree within a few milligrams or tens of milligrams per liter unless the water is of exceptional composition. The comparison is often helpful in analytical or transcribing errors.

- Charge balances: The sum of anions approximately equals sum of cations (in meq/L). The difference between the two sums will generally not exceed 1 or 2 percent of the total of cations and anions in waters of moderate concentration (250–1,000 mg/L). If the total of anions and cations is less than about 5.00 meq/L, a somewhat larger percentage difference can be tolerated. Calculated values would be the sum of individual ion results and determined is the lab result.

Caveat: Water having dissolved-solids concentrations much greater than 1,000 mg/L tends to have large concentrations of a few constituents. In such water, the test of anion-cation balance does not adequately evaluate the accuracy of the values of the lesser constituents.

- EC vs TDS: An approximate accuracy check is possible using the conductivity and dissolved-solids determinations. The dissolved-solids value in milligrams per liter should generally be from 0.55 to 0.75 times the specific conductance in micromhos per centimeter for waters of ordinary composition, up to dissolved-solids concentrations as high as a few thousand milligrams per liter. Water in which anions are mostly bicarbonate and chloride will have a factor near the lower end of this range, and waters high in sulfate may reach or even exceed the upper end.

- Sum of cations (in meq/L) times 100 approximately equals EC (in µS/cm): The total of milliequivalents per liter for either anions or cations multiplied by 100 usually agrees approximately with the conductivity in microSiemens per centimeter.

- Comparison to a Pre-Defined Relationship: For repeated samples from the same source, a well-defined relationship of conductivity to dissolved solids often can be established, and this can afford a good general accuracy check for analyses of these samples.

Caveat: The relationship of dissolved solids to conductance becomes more poorly defined for waters high in dissolved solids (those exceeding about 50,000 mg/L) and also for very dilute solutions, such as rainwater. For solutions of well-defined composition such as seawater, however, conductivity is a useful indicator of ionic concentration.

Hodge V and J Austin. 2004. A Survey of Outlier Detection Methodologies. Artif. Intell. Rev. 22, 2 (2004), 85–126.

An extensive review paper that provides a broad sample of modern approaches. This paper is from 2004, so some of the specific software examples may be out of date, while the underlying principle still stands. They take a very broad view of where outliers might occur, from financial fraud to finding novel features in satellite images. They confirm that there “is no single universally applicable or generic outlier detection approach”. They recommend that the evaluator consider which of three fundamental approaches is most suitable for their data set: an (unsupervised) clustering approach (e.g., excluding physically impossible values), a classification approach that involves supervised classification with labeled outliers, or a novelty approach, which also relies on training to define what is “normal” versus “novel.” The choice will depend on the data type and what the evaluator intends to do with the outliers: Remove them from any further processing or keep them as flagged data to train a further classification system. The paper is divided into (1) an introduction, (2) statistical models (e.g., Grubb’s Z-score, though they move into higher-dimensional approaches such as PCA), (3) Neural networks (supervised and unsupervised), (4) Machine Learning, (5) Hybrid systems, and (6) Conclusions. Overall, this paper is pitched at a relatively theoretical level that provides good principles (e.g., a classifier should obey Occam’s Razon and be as simple as possible with minimal redundancy) but might be challenging to implement in practice.

IOOS. Integrated Ocean Observing System. 2016. Manual for Real-Time Quality Control of In-situ Temperature and Salinity Data Version 2.0: A Guide to Quality Control and Quality Assurance of In-situ Temperature and Salinity Observations.

This technical report describes a series of tests designed for real-time quality of continuous temperature and salinity data. Tests result in Pass, Suspect, and Fail flags with the intent that a subject matter expert will review them. However, these same techniques can be applied to virtually any continuous signal. The authors outline the 13 real-time QC tests that are required, strongly recommended, or suggested for real-time measurements of temperature and salinity. Note that the accompanying manuals on

- Dissolved Oxygen data (https://doi.org/10.25923/q0m1-d488),
- Dissolved Nutrients data (https://doi.org/10.7289/V5TT4P7R),
- Phytoplankton data (https://doi.org/10.7289/V56D5R6S), and
- Water Level data (https://doi.org/10.25923/vpsx-dc82)

propose the same set of tests.

Group 1 “Traditional Approaches” (IOOS required)

1. Gap Test: determines that the most recent data point has been measured and received within the expected time window and has the correct time stamp.
2. Syntax Test: checks that received data message contains the proper structure without any indicators of flawed transmission such as parity errors (example: the expected number of characters for fixed-length messages equals the number of characters received).
3. Location Test: checks that the reported present physical location (latitude/longitude) is within operator-determined limits.
4. Gross Range Test: checks if data point exceeds sensor or operator-selected min/max. All sensors have a limited output range, and this can form the most rudimentary gross range check. Additionally, the operator can select a smaller span based upon local knowledge or a desire to draw attention to extreme values.

5. Climatological Test: a variation on the gross range check, where the thresholds are adjusted monthly, seasonally, or at some other operator-selected time period. Expertise of the operator is required to determine reasonable seasonal averages.

Group 2 (IOOS strongly recommended)

6. Spike Test: check for single value spikes. Adjacent data points are averaged to form a spike reference. The absolute value of the spike is tested to capture positive and negative spikes. Multiple thresholds can be used to differentiate between large and small spikes.

7. Rate of Change Test: inspects the time series for a time rate of change that exceeds a threshold value identified by the operator.

8. Flat Line Test: This test compares the present observation to several previous observations to identify a continuously repeated observation of the same value (a common result when sensors and/or data collection platforms fail).

Group 3 (IOOS suggested)

9. Multivariate Test: pairs rate of change tests as described in Test 7. The rate of change test is conducted for one analyte with a more restrictive threshold; if this test fails, a second rate of change test operating on a second analyte is conducted.

10. Attenuated Signal Test: inspects for a standard deviation or range variation that fails to exceed threshold values over a selected time period. Designed to identify a sensor failure mode where a data series is nearly but not exactly a flat line (e.g., if the sensor head were to become wrapped in debris).

11. Neighbor Test: This test is the same as Test 9, but a second sensor is used rather than a second analyte. Ideally, redundant sensors using different technology would be co-located and alternately serviced at different intervals, but in some cases a nearby (but not co-located) sensor can provide a useful QC check as well. The selected thresholds depend entirely upon the relationship between the two sensors as determined by the local knowledge of the operator.

12. TS Curve/Space Test: Specific to temperature-salinity tests in open ocean below the thermocline. The authors note the test will probably not be useful in estuaries or ocean surface waters.

13. Density Inversion Test: checks that potential water density increases with increasing pressure. Similar to Test 7 but substitutes depth (vertical space)

for time. Useful only for vertical profiles where thermal or chemical-driven density variations are expected.

The authors suggest defining multiple thresholds for each test where possible to differentiate failure values from "suspect" values which require expert review for final flagging.

Komsta L. 2006. Processing data for outliers. R News, 6(2), 10–13.

Summary of various outlier tests and their R implementations in the R "outliers" package. Includes Grubbs test, Dixon test and Cochran test.

Leys C, et al. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median, Journal of Experimental Social Psychology.

This is a short paper that seems to be more of an opinion piece advocating for using the median absolute deviation (MAD) statistic rather than the mean and standard deviation for detecting outliers. Used in the USGS sensorQC R package.

Li, Susan. 2019. Time Series of Price Anomaly Detection: Anomaly detection detects data points in data that does not fit well with the rest of the data.

https://medium.com/data-science/time-series-of-price-anomaly-detection-13586cd5ff46 . Accessed Jan 2025.

This particular blog post provides a detailed R-based analysis of a time series data set using k-means clustering algorithm, principal components analysis, and several other innovative approaches.

Peterson Thomas C, R Vose, R Schmoyer, and V. Razuvaev. 1998. Global historical climatology network (GHCN) quality control of monthly temperature data. International Journal of Climatology, 18, 1169-1179.

A case study using the global historical climatology network (GHCN) to evaluate various traditional outlier detection methods. While the topic is meteorological data, the recommendations are very similar to the IOOS framework. A very practical, accessible paper.

Rebbapragada U, P Protopapas, CE Brodley, and C Alcock. 2009. Finding anomalous periodic time series: An application to catalogs of periodic variable stars. Machine Learning 74, 3 (2009), 281–313.

This paper reviews PCAD, an unsupervised outlier detection method for "unsynchronized periodic" time-series data. The abstract makes the point that this data type is unlike single continuous time series data or multiple time series with "aligned periods."

Tripathy, Sushree Swarupa, Rajiv Kumar Saxena, Prabhat Kumar Gupta. Comparison of Statistical Methods for Outlier Detection in Proficiency Testing Data on Analysis of Lead in Aqueous Solution. American Journal of Theoretical and Applied Statistics. Vol. 2, No. 6, 2013, pp. 233-242.

A case study using data from proficiency testing of various laboratories looking at lead concentrations in water. This paper appears to review and compare various traditional outlier detection methods being employed on this laboratory data. It is a shorter paper and reviews traditional methods used on discretely-collected laboratory data.

Yu Y, Zhu Y, Li S, & Wan D. 2014. Time series outlier detection based on sliding window prediction. Mathematical problems in Engineering.
https://doi.org/10.1155/2014/879736.

This paper proposes a k-nearest-neighbor method based on moving windows for outlier detection in a time series and applied to a hydrologic time series. The method first builds a forecasting model on the historical data in the moving window and then used it to predict future values. Anomalies are assumed to take place if the observed values fall outside a given prediction confidence interval (PCI), which can be calculated by the predicted value and confidence coefficient. The use of PCI as threshold is advantageous because it calculated based on forecasting model itself rather than requiring operators to define thresholds a priori. This can be applied in a real-time, automated setting without supervision and it does not require predefined thresholds, and thus could be applied to any variety of analytes and locations without customization or tuning.