

Data Quality Assessment Using a Sliding Window Cumulative Sum Control Chart

Jesse Njuguna¹, Julia L. Sharp², Sam Esswein², Aravindh Sampath², Jason Hallstrom², Gene Eidson²

¹Washington State University, ²Clemson University

Abstract

When assessing the quality of streaming data, as is the case for the Intelligent River®, a landscape-scale observation instrument covering a large geographic area, challenges arise. Statistical process control procedures aid in monitoring and analyzing variability in process performance and allow for the identification of extreme observations. Traditional quality control charts are limiting because of the quantity of data across time and from many sensors in a streaming environmental data collection domain.

We suggest the use of a sliding-window cumulative sum (cusum) quality control chart. In the sliding-window approach, data (e.g., turbidity measurements from a single sensor) in a fixed window of number of observations is used to compute unique statistical control limits for analyzing the quality of observations moving through time. Our results show that the sliding-window approach may detect shifts and outlying observations; however, other techniques like the cusum control chart using a trimmed mean or the Exponentially Weighted Moving Average (EWMA) control chart may be more appropriate for identifying “out-of-control” observations.

What is a Cusum Control Chart?

Cusum chart, an acronym for the Cumulative Sum control chart is a statistical method that is used to monitor data variation during quality control processes.

Tabular Cusum

- The tabular cusum works by accumulating deviations from

$\mu \downarrow 0$, the target value for the quality characteristic of an individual observation (Montgomery, 2012).

- Deviations from the target value are accumulated using

the statistic $C \uparrow +$ (upper cusum) if the value

is above target or statistic $C \uparrow -$ (lower

cusum) for those below target. The statistics

Sliding Window Approach

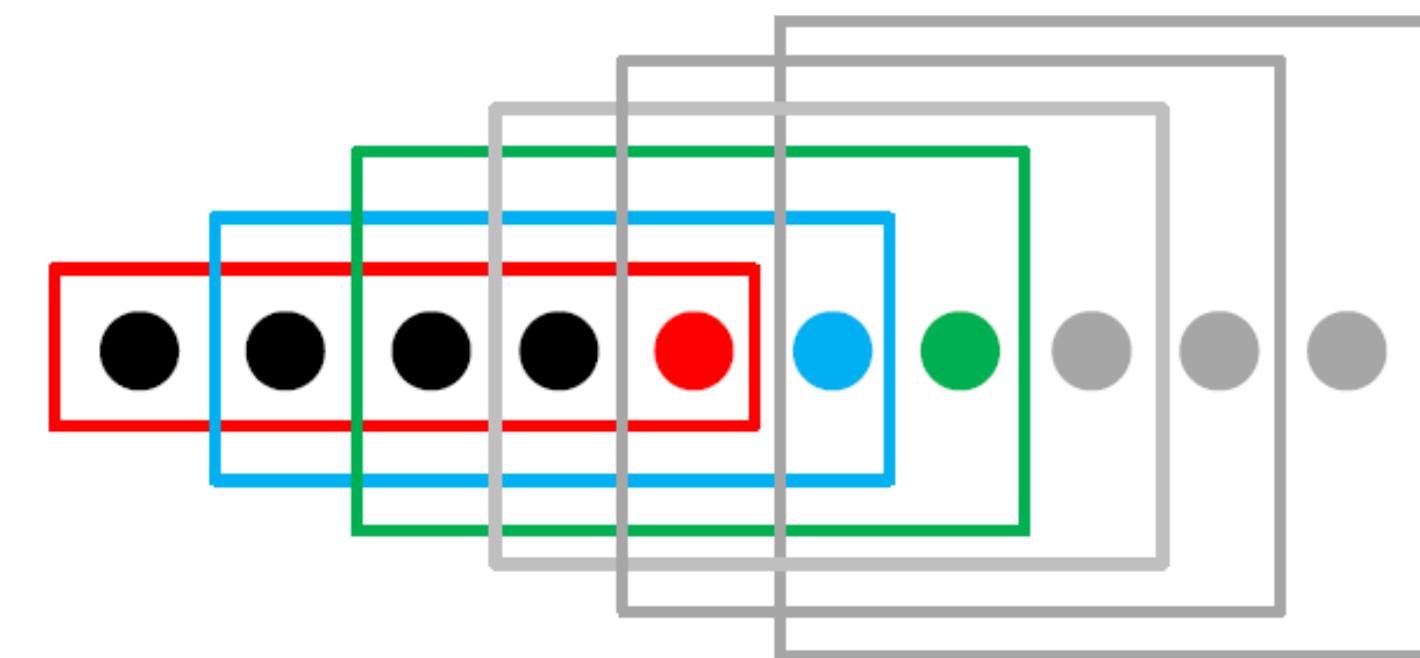


Figure 1: Illustration of the sliding window approach.

- Consider the i^{th} observation within a data stream and a window size of five observations (Figure 1). The four previous observations as well as the i^{th} observation are used to compute the cusum and assess the state of control for the i^{th} observation.

Sliding Window Tabular Cusum Function

- The sliding window tabular cusum function takes as its arguments a dataset, and the mean of the previous window of observations (

$x \downarrow i-1$) to represent the target value for the tabular cusum computation of the current window of observations

- To detect unit.

- The deci: of the cu

- The func

Figure 3: Turbidity sample cusum control chart with a window size of 5 observations. No observations are flagged as “out-of-control”.

of the current window of observations and if either

statistic exceeds the decision interval ($\pm 5\sigma$), the observation is considered to be in an “out-of-control” state.

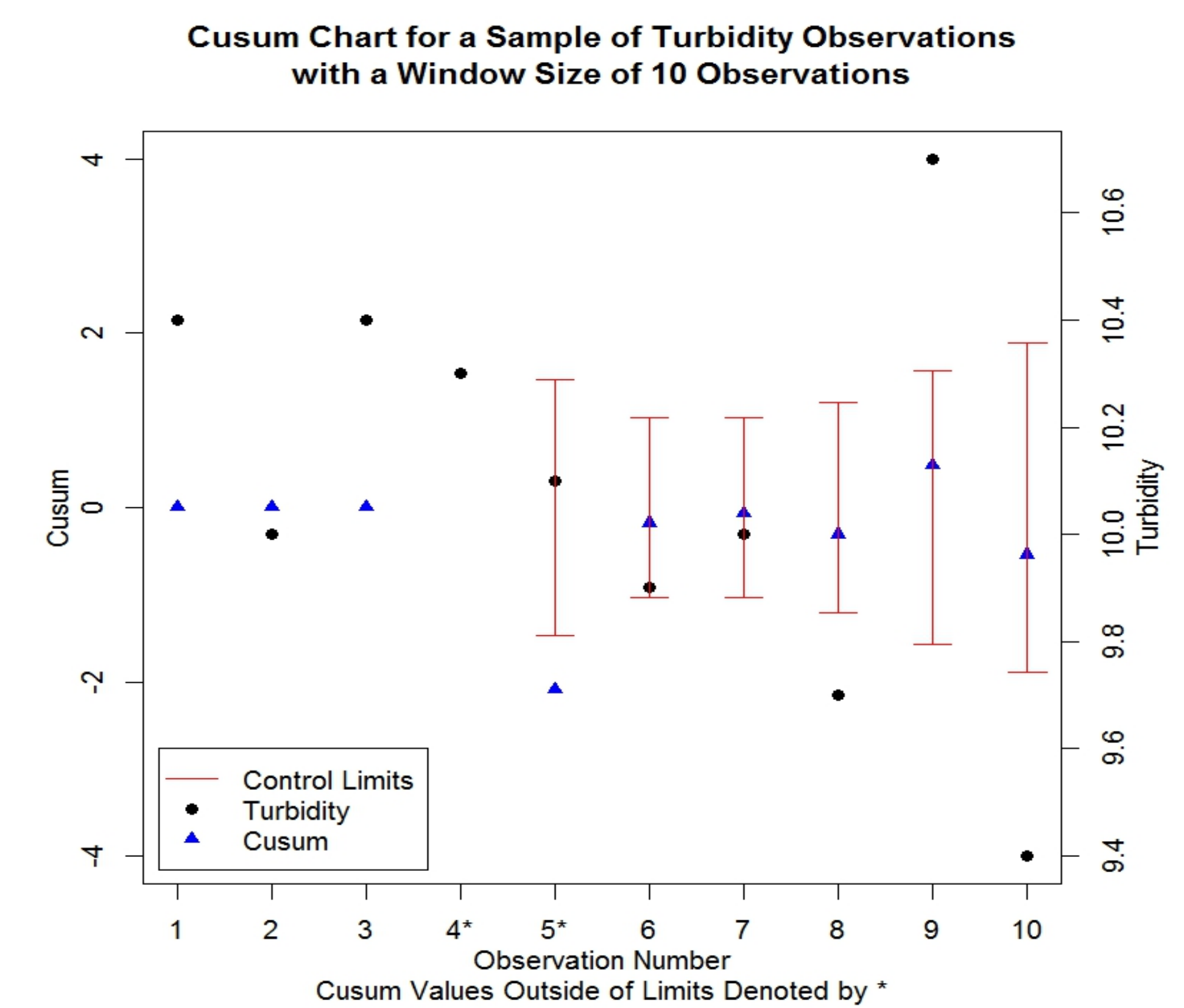


Figure 2: Turbidity sample cusum control chart with a window size of 10 observations. Note: Control limits and/or cusum values for some observations exceed the dimensions of the figure and are not plotted. Note that observations that do not appear to be out-of-control are flagged as such due to the variation in the window of observations.

Discussion and Conclusions

- It is observed that extreme observations in the current and previous window of observations cause shifts in the cusum value, depending on the window size.
- Relative to the size of the shift to monitor, a quality control procedure with a larger window size has higher sensitivity to detect shifts as demonstrated by magnitude of the decision interval. However, these shifts may not be of practical importance.
- The cusum control chart based on a sliding window illustrates how statistical control charts can be used to detect shifts in a streaming data environment such as from the data collection in the Intelligent River®.
- Future work includes using an exponentially weighted moving average (EWMA) control chart to detect truly extreme observations in a streaming environment.

Acknowledgement

This work was supported by grants from the National Science Foundation (CNS-1126344), National Park Service – Great Smoky Mountains National Park Grant #2097821, and the City of Aiken, SC.

References

Montgomery, Douglas C. *Introduction to Statistical Quality Control*. Wiley, 2012.