Breaking Bad: robust Breakout detection based on E-Divisive with Medians (EDM) for modeling data quality control

Hao Zhang

January 31, 2018

The Philadelphia water department (PWD) has been actively monitoring flow data at over 400 sites over Philadelphia since the 2000s. Data is collected twice a month through contractors. Due to the high solid content in sewage, flow data at sewer pipes (level, velocity) suffered from breakouts (mean shift, ramp up) over the time caused by sensor ragging, pipe clogging, etc. A stringent Quality Control (QC) protocol is conducted before the data can be used for Hydrologic & Hydraulic modeling tasks. As one QC measure, the water level and velocity are examined to detect any potential breakout.

Since flow data fluctuates with rainfall-runoff events, the breakout detection algorithm must be robust to avoid the interference of runoff responses. Several breakout detection techniques were compared, and the E-Divisive with Medians (EDM) algorithm is adopted in this study. EDM recursively partitions a time series and uses a permutation test to determine change points. The EDM has following advantages: 1. EDM uses moving median as opposed to the mean, which is robust to the presence of anomalies; 2. EDM can detect both ‘mean shift’ (sudden change) and ‘ramping’ (gradual change) for multiple change points; 3. EDM takes a non-parametric approach, meaning the model will adapt to the data’s underlying distribution, and therefore can detect when distribution changes; 4. EDM is fast, as it uses interval trees to efficiently approximate the median.

The analysis is implemented in a program written in R, and the EDM algorithm is implemented via the ‘breakoutDetection’ package developed by Twitter.Inc. Non-trivial parameters of the EDM model are carefully tuned to match the expected outcome. This analysis provides an additional assurance to the data quality. Also, field crews (monitoring, Operation & Maintenance, etc.) can quickly respond to the issue once a breakout has been detected. This analysis is also applicable for other monitored data, such as the trunk and outfall levels at drainage system regulators.