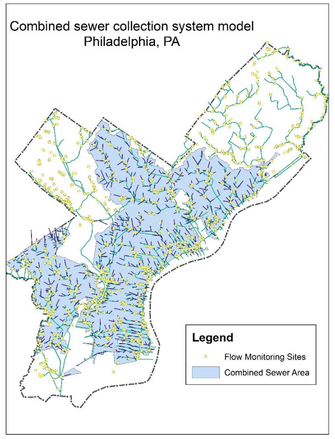
Breaking Bad: Robust Breakout Detection Based on E-Divisive with Medians (EDM) for Modeling Data Quality Control

Hao Zhang, Ph.D., Philadelphia Water Department, Philadelphia, PA, USA

INTRODUCTION

- The Philadelphia Water Department (PWD) maintains hydrologic and hydraulic (H&H) models of the combined sewer collection system for planning, management and compliance purposes
- For model calibration/validation, sewerage level and velocity at over 400 manholes have been monitored since the 2000s, with a monitoring period of at least one year
- A stringent Quality Control (QC) protocol is conducted before the data can be used for Hydrologic & Hydraulic modeling tasks
- Due to the high solid content in sewage, monitored data may suffered from breakouts caused by
  - sensor ragging, clogging
  - pipe surcharging, etc.
- Visual detection of breakouts may not be feasible as some breakouts are not obvious. Thus, a programmatic approach that can automatically detect breakouts is imperative for modeling data quality control. Also, field crews (monitoring, Operation & Maintenance, etc.) can quickly respond to the issue



OBJECTIVES

Overall, monitored data quality determines the model quality. This study aims to develop a workflow as a quality control (QC) measure for detecting various types of breakouts in flow monitoring data by utilizing a sound breakout detection algorithm.

First, select a breakout detection algorithm that is:

- able to detect various types of change (mean shift, ramp up, variance change, etc.)
- robust against the presence of anomalies (as runoff response tends to be the interference)
- able to detect multiple breakouts
- not assuming sample distribution (as it is usually unknown a-priori)
- fast enough for routine tasks

Next, tune the parameters of the breakout detection algorithm to optimize the outcome

Finally, develop an application using the R statistical programming language to analyze flow monitoring data, and generate quarterly reports. Set up a routine workflow for this process.

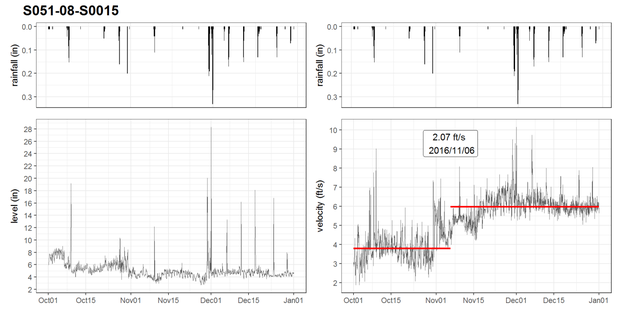
METHODS

- A novel statistical technique, E-divisive with medians (EDM), is utilized for this study.
- As compared to the other algorithms, EDM has the following advantages:
  1. EDM utilizes a local smoother (rolling median) to raw data, and therefore is robust to the presence of anomalies;
  2. EDM employs energy statistics (E-divisive) to detect divergence of means that can detect 'mean shift' (sudden change), 'ramping' (gradual change), and distribution changes at multiple change points. This method is proven to have comparable or better efficacy
  3. EDM is non-parametric, the method will adapt to the data's underlying distribution, and can detect when distribution changes
  4. EDM is 3.5x faster due to the usage of interval trees to approximate median. Since the median is approximated, the breakout location may be not exact.

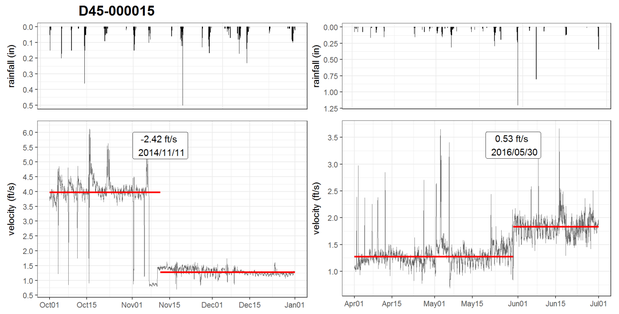
RESULTS

- The EDM algorithm is implemented by the `BreakoutDetection`, which is an open-source R package developed by Twitter Engineers and has been used for analyzing network breakouts on a daily basis at Twitter.
- The `breakout()` function contains several non-trivial parameters:
  - `Z`: The input time series. In this study, the hourly averaged data is used for efficiency
  - `min.size`: The minimum number of observations between change points. E.g., in this study, it is set to 240, i.e., the steady state must be at least 10 days (10 x 24 = 240)
  - `method`: '`amoc`' (At Most One Change) or '`multi`' (Multiple Changes). In this study, '`multi`' is chosen as multiple breakouts are desired
  - `degree`: The degree of the penalization polynomial; can be 0, 1, or 2. In this study, `degree = 1`
  - `beta`: the default form of penalization, In this study, `0.008` is determined via elbow plot
  - `percent`: the minimum percent change in the goodness of fit statistic to consider adding an additional change point. In this study, `0.10` (10%) is used
- Implementation:
  - Breakouts are detected by the `breakout()` function in the `BreakoutDetection` package in R
  - A custom function to plot breakouts with time-series is developed using the `ggplot2` package
  - A R markdown template is created for generating quarterly reports

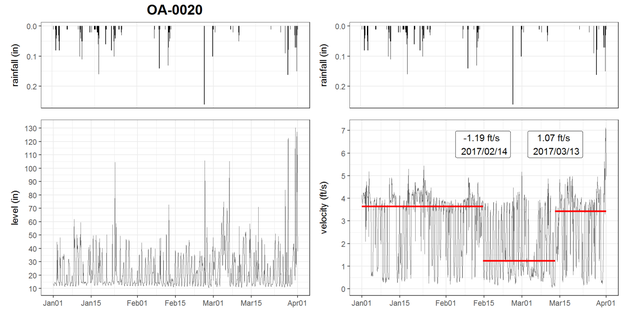
- Ramping



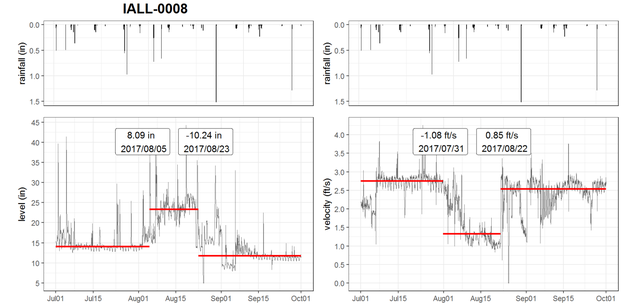
- Mean shift (up, down)



- Velocity meter clogging



- pipe-surcharging



CONCLUSIONS

- EDM can effectively detect multiple breakouts in monitored time-series with the presence of anomalies (runoff response)
  - Velocity measurements tends to show more breakouts than level measurements
- EDM has a few limitations:
  - Breakouts at both ends of the time series could not be identified
  - large runoff response may be recognized as breakouts

REFERENCES

- James, Nicholas A., Arun Kejariwal, and David S. Matteson. "Leveraging cloud data to mitigate user experience from 'Breaking Bad'." In Big Data (Big Data), 2016 IEEE International Conference on, pp. 3499-3508. IEEE, 2016.
- Matteson, David S., and James, Nicholas A. "A nonparametric approach for multiple change point analysis of multivariate data." Journal of the American Statistical Association 109, no. 505 (2014): 334-345.
- Rebecca Killick, Paul Fearnhead, and IA Eckley. "Optimal detection of changepoints with a linear computational cost." Journal of the American Statistical Association, 107(500):1590–1598, 2012
- Rodionov, S. N. "A brief overview of the regime shift detection methods." Large-scale disturbances (regime shifts) and recovery in aquatic ecosystems: challenges for management toward sustainability (2005): 17-24.
- Pohlert, Thorsten. "Non-parametric trend tests and change-point detection." CC BY-ND 4 (2018).