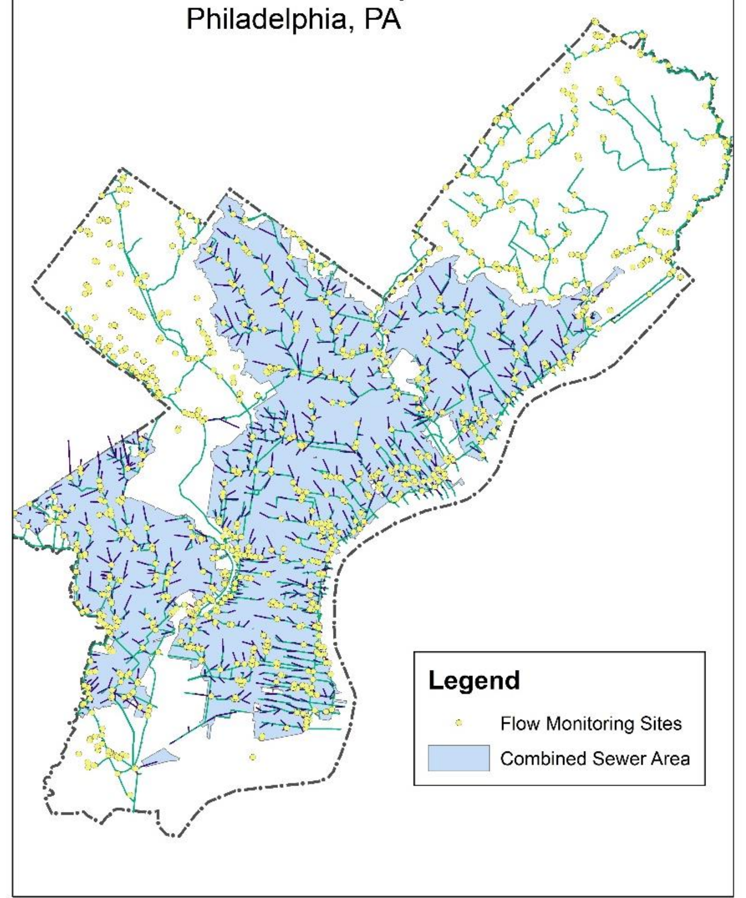
Breaking Bad: Robust Breakout Detection Based on E-Divisive with Medians (EDM) for Modeling Data Quality Control

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 H&H modeling tasks
- Due to the high solid content in sewage, monitored data may suffered from breakouts caused by
  - ragging, clogging
  - surcharging, etc.
- Visual breakout detection 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 crew (monitoring, Operation & Maintenance, etc.) can be notified for quickly responding to field issues

Combined sewer collection system model
Philadelphia, PA



OBJECTIVES

Monitored data quality determines 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/down, variance change, etc.)
- robust against the presence of anomalies (as the runoff component tends to be the interference)
- able to detect multiple breakouts in a time-series
- not rely on sample distribution (as it is usually unknown a-priori)
- fast enough for routine tasks

Next, tune the argument of selected breakout detection algorithm to optimize the outcome

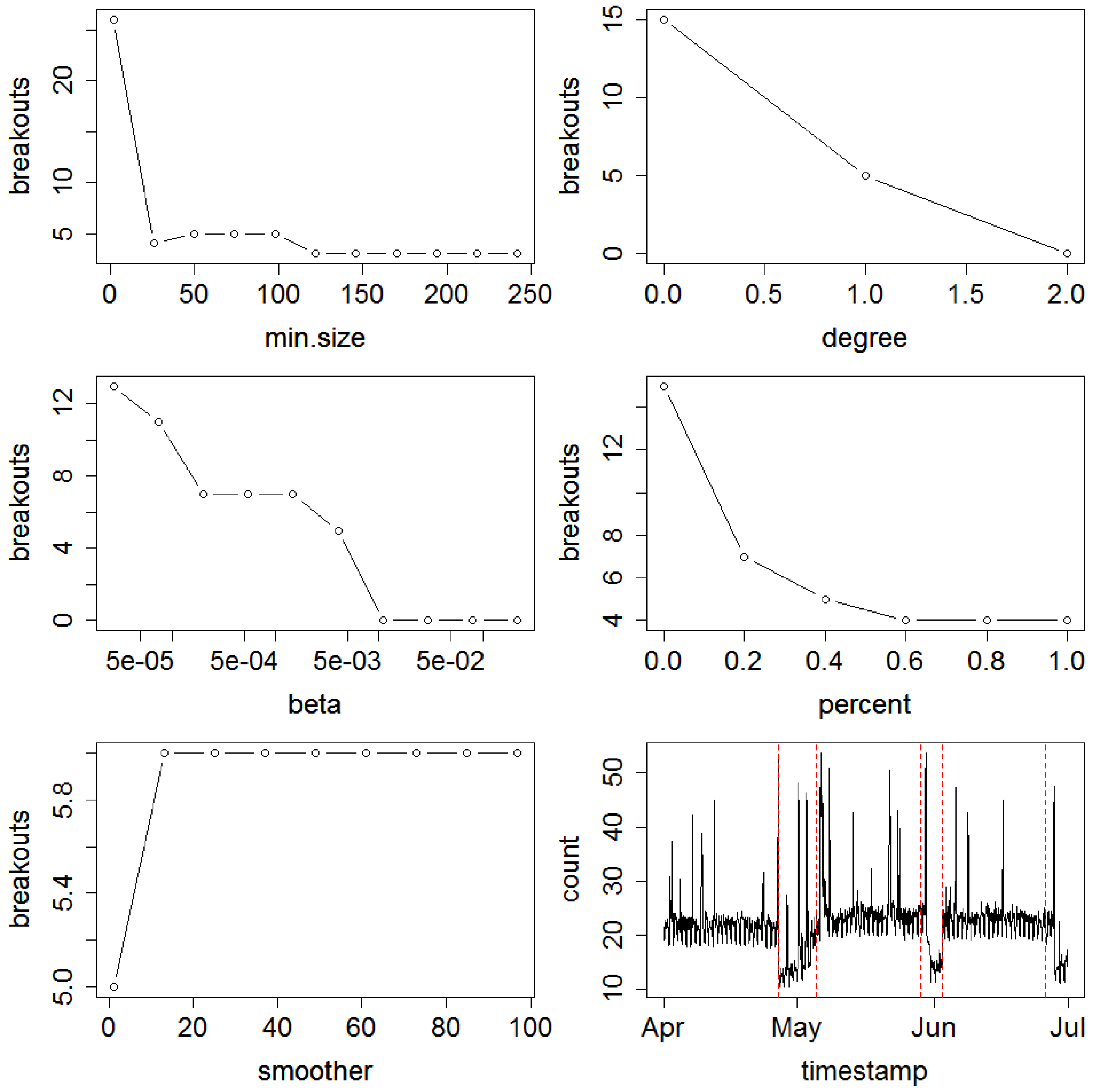
Finally, develop an application using the R statistical programming language to analyze flow monitoring data, and generate quarterly reports. Set up a routine workflow for this process.

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 against 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.
  3. EDM is non-parametric, which will adapt to the data's underlying distribution, and can detect when the distribution changes
  4. EDM is proven to have comparable or better efficacy, and it is 3.5x faster due to the usage of interval trees to approximate median.

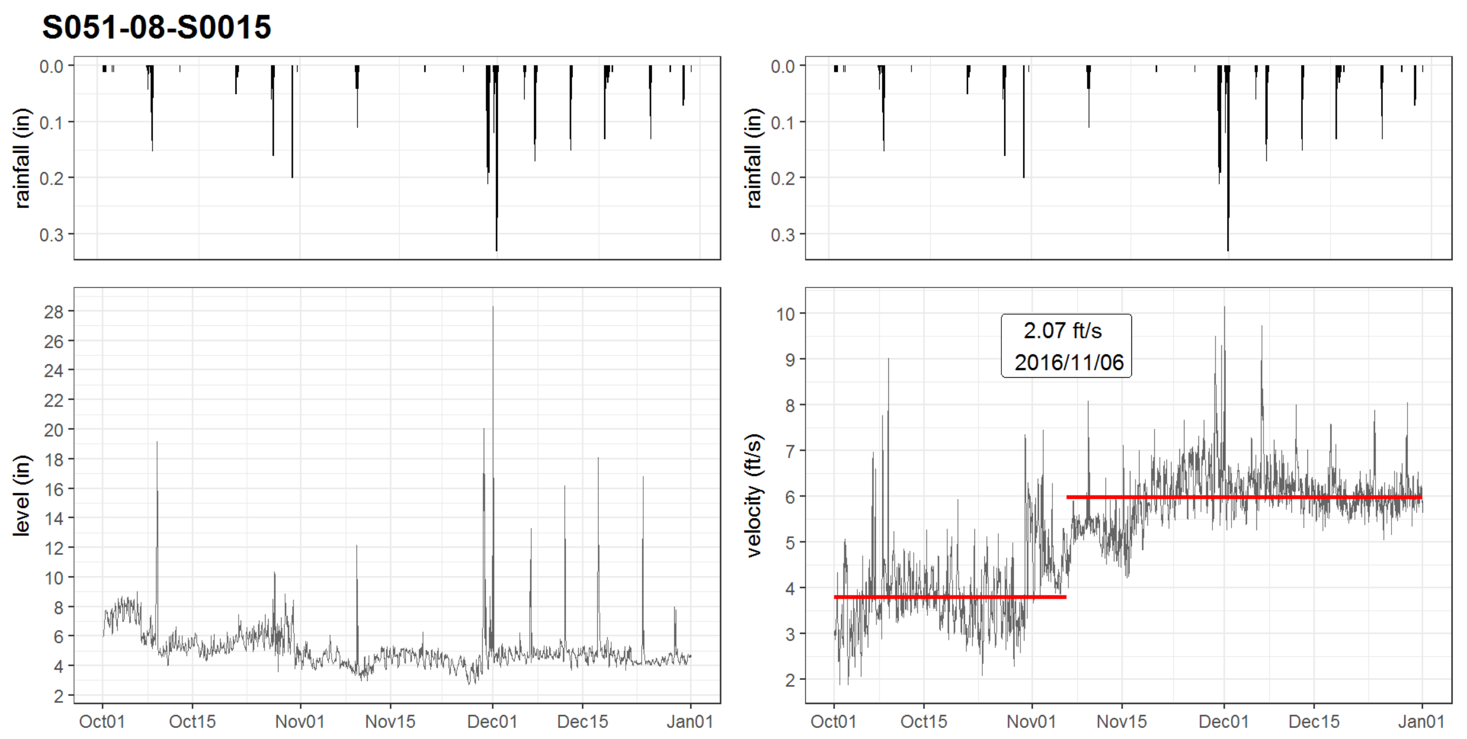
RESULTS

- The EDM algorithm is implemented by `BreakoutDetection`, an open-source R package developed by Twitter Engineers, which has been used for analyzing cloud data on a daily basis at Twitter
- `breakout()`: the detector function, includes several non-trivial argument:
  - `Z`: The input time series. In this study, `Z` = quarterly time-series @ 1 hour interval
  - `min.size`: The minimum number of observations between change points. In this study, `min.size = 120`, i.e., 5 days (5 x 24 = 120)
  - `method`: 'amoc' (At Most One Change) or 'multi' (Multiple Changes). In this study, `method = 'multi'`
  - `degree`: The degree (0, 1, or 2) of the penalization polynomial. In this study, `degree = 1`
  - `beta`: the default form of penalization. In this study, `beta = 0.008` for velocity, `0.002` for level
- The values of argument are determined through a series of trials with the assistance of elbow plots:

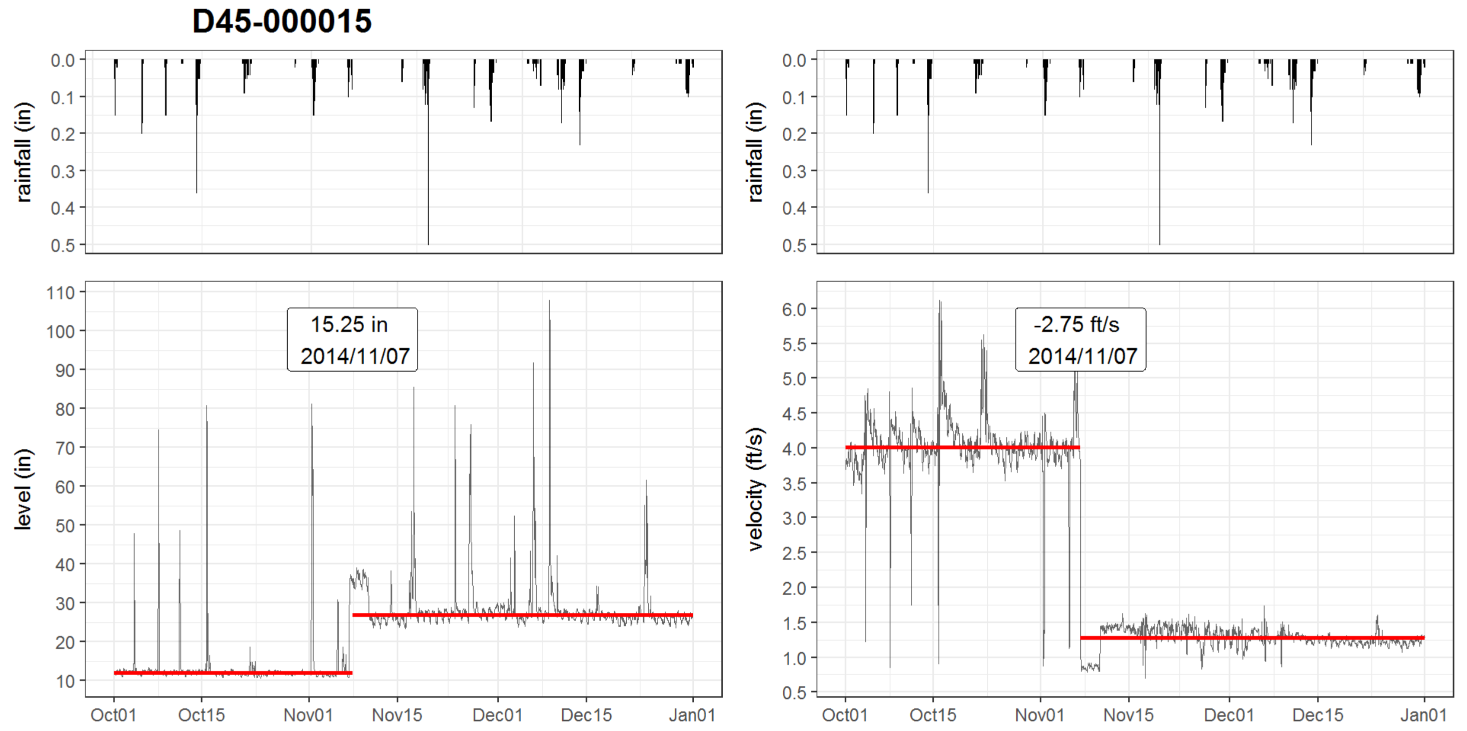


- Implementation:
  - Breakouts are detected by the `breakout()` function in the `BreakoutDetection` package in R
  - A custom function to plot breakouts with time-series is developed using the `ggplot2` package
  - A R markdown template is created for generating quarterly reports
  - The report is updated bi-weekly when new data becomes available

- Example: Ramp (up, down)



- Example: Mean shift



CONCLUSIONS

- With properly tuned argument, the E-divisive with Median (EDM) method can effectively detect multiple breakouts in sewage level and velocity time-series with known anomalies (runoff), and is expected to be applicable for other monitored time-series data.
- This application provides Quality Control (QC) to the modeling data, and can be used as an early warning system for field issues.

REFERENCES

- James, Nicholas A., Kejariwal, Arun, and Matteson, David S. 2016."Leveraging cloud data to mitigate user experience from 'Breaking Bad': The Twitter Approach." *2016 IEEE International Conference on Big Data* 3499-3508.
- Matteson, David S., and James, Nicholas A. 2014. "A non-parametric approach for multiple change point analysis of multivariate data." *Journal of the American Statistical Association* 109(505): 334-345.