A Hybrid Leakage Detection and Isolation Approach Based on Ensemble Multivariate Changepoint Detection Methods

Tuoyuan Cheng[1], Yuanzhe Li[2], Fouzi Harrou[3], Ying Sun[3], Jinliang Gao[2,*], TorOve Leiknes[1]
[1] Water Desalination and Reuse Center, Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal, 23955-6900 Saudi Arabia
[2] School of Environment, Harbin Institute of Technology, Harbin 150090, China
[3] Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
[*] *gjl@hit.edu.cn*

ABSTRACT

Early leakage detection and isolation are of paramount importance to the maintenance and resilience of the water distribution system (WDS). Efficient and accurate leakage detection and isolation algorithms could support practitioners to optimize WDS design and operation. Modern WDS, as equipped with supervisory control and data acquisition (SCADA), could record nodal measurements to facilitate training, validation, and selection of both mechanism-based and data-driven models.

In this study for the battle of the leakage detection and isolation methods, a hybrid approach based on benchmark simulation and the ensemble multivariate changepoint detection (EMCPD) is proposed to detect leakage occurrences. Bilinear bivariate spatial interpolation for irregular spaced data and two-sample one-sided Student's t-test is further invocated to isolate leakage sites.

First, a simulation was performed, assuming no leakage occurred to attain the benchmark working condition series of the L-town WDS. The pressure-dependent demand scheme was deployed via the toolkit, water network tool for resilience (WNTR), with a simulation time step of five minutes [1]. Upon simulation, the residuals of flowrates and pressures were calculated by contrasting the sensor measurements and the simulated results. The observed demand data were assimilated as the nodal pattern, and therefore the nodal demand residuals were omitted.

Based on the flowrate residuals, the EMCPD was performed to detect the rough time of leak events. Since the three flowrate sensors were installed either by the pump or by the reservoirs, the flowrate residuals should offer direct evidence of leakage events. At this stage, the flowrate residuals series with three variables were averaged to three observations per day in order to denoise short term fluctuation, keep diurnal and weekly cycles, and accelerate the computation.

Around each rough candidate, the EMCPDs were performed to distinguish accurate moment candidates. At this stage, both the flowrate residuals (three variables) and pressure residuals (34 variables, tank water level included) series were engaged at five minutes level, from 46 hours before the rough timing to 2 hours after the rough timing. With more covariates involved, the multivariate residuals time series would shrink the accurate candidates' pool.

Finally, around each accurate moment candidate, the pressure contours of the L-town were calculated by adopting bilinear bivariate spatial interpolation based on the monitored 34 pressure time series [2]. To be conservative, only the nodal pressures inside the convex hull of pressure sensors were estimated from 25 minutes before each accurate candidate moment to 25 minutes after each accurate candidate moment. For each node, its estimated pressures time series were firstly exponentially moving averaged and then grouped as before or after the candidate moment. The two-sample one-sided Student's t-tests with unknown variance were performed on the two groups to isolate the most likely sites. The sites would be characterized as showing the minimal p-values, or with the most significant negative mean pressure shift between two groups. Since all pressures were estimated based on 34

sparse sensors, bilinear spatial interpolation without extrapolation, as well as a p-value threshold of 0.01 were adopted, with the intention to avoid overfitting, to be conservative at un-monitored nodes, and to be parsimonious at raising alarms.

The EMCPD was performed by involving the results of six algorithms, including the most recent changepoint detection method [3], the non-parametric multiple changepoint analysis methods (with the E statistic or the Kolmogorov-Smirnov statistic) [4], the divisive hierarchical estimation algorithm, the kernel changepoint analysis, and the Bayesian estimator of abrupt change and trends [5]. The most recent changepoint detection method analyzed univariate time-series independently to obtain a profile-likelihood measures that summarizes the evidence of changepoints and then pooled information across time-series to give an optimal set. The non-parametric MCPD involved dynamic programming and pruning without underlying distributional assumptions. The Bayesian method quantified the relative usefulness of encompassed individual decomposition models, and then leveraged them via Bayesian model averaging. The votes from individual MCPDs were gathered to elect the most likely candidate per detection turn. Hyperparameters were tuned according to the 2018 dataset.

The proposed approach perceived 30 rough candidates and 55 accurate candidate moments, as tabulated in Tab.1. The vote counts from EMCPD was relatively scattered for the rough candidates, but highly concentrated for the accurate candidates. The flowrate residuals as well as rough candidates were delineated in Fig.1. Four instances of pressure changes around an accurate candidate were graphed in Fig.2. The Fig.2(a) presented an abrupt pressure drop prior and posterior to an accurate moment candidate, which could be due to a pipe burst event. The Fig.2(b) presented a minor pressure drop prior and posterior to an accurate moment candidate, which could be due to the downward pressure trend drifting lead by the development of a background leakage. The Fig.2(c) presented an abrupt pressure rise prior and posterior to an accurate moment candidate, which could be due to the repair of an existing pipe burst event. The Fig.2(d) presented a minor pressure rise prior and posterior to an accurate moment candidate, which could be due to the upward pressure trend drifting. The isolation outcomes together with one-sided Student's t-test results based on exponentially moving averaged short-term pressure time series around accurate moment candidates were listed in Tab. 2.

Table 1. Detected moment candidates.

Candidate type	Detected moment candidates
Rough candidates (30 in total number)	2019-01-15 23:55:00, 2019-01-24 15:55:00, 2019-01-24 23:55:00, 2019-02-01 15:55:00, 2019-02-07 15:55:00, 2019-02-15 07:55:00, 2019-02-26 07:55:00, 2019-03-20 23:55:00, 2019-03-22 07:55:00, 2019-04-02 15:55:00, 2019-04-19 15:55:00, 2019-05-05 23:55:00, 2019-05-23 23:55:00, 2019-06-13 07:55:00, 2019-06-13 15:55:00, 2019-07-01 15:55:00, 2019-07-17 07:55:00, 2019-07-17 15:55:00, 2019-07-17 23:55:00, 2019-08-22 07:55:00, 2019-08-27 07:55:00, 2019-08-27 15:55:00, 2019-09-03 15:55:00, 2019-09-16 07:55:00, 2019-10-01 23:55:00, 2019-10-25 07:55:00, 2019-10-25 15:55:00, 2019-11-13 23:55:00, 2019-11-27 15:55:00, 2019-12-22 23:55:00.
Accurate candidates (55 in total number)	2019-01-14 17:30:00, 2019-01-15 13:45:00, 2019-01-23 03:40:00, 2019-01-24 14:20:00, 2019-01-24 12:50:00, 2019-01-25 00:30:00, 2019-01-31 12:10:00, 2019-01-31 16:50:00, 2019-02-07 03:10:00, 2019-02-07 15:25:00, 2019-02-13 10:55:00, 2019-02-14 21:55:00, 2019-02-25 08:10:00, 2019-02-26 09:45:00, 2019-03-19 12:55:00, 2019-03-20 19:20:00, 2019-03-21 22:20:00, 2019-03-31 18:30:00, 2019-04-01 19:10:00, 2019-04-19 05:55:00, 2019-04-19 10:45:00, 2019-05-04 07:15:00, 2019-05-05 23:50:00, 2019-05-23 07:40:00, 2019-05-23 14:55:00, 2019-06-12 13:25:00, 2019-06-13 09:40:00, 2019-06-13 16:40:00, 2019-06-29 23:05:00, 2019-07-01 13:45:00, 2019-07-17 02:40:00, 2019-07-17 09:35:00, 2019-07-17 05:55:00, 2019-07-16 09:55:00, 2019-08-21 04:50:00, 2019-08-22 07:20:00, 2019-08-26 06:50:00, 2019-08-26 14:15:00, 2019-08-27 02:30:00,

2019-09-02 06:15:00, 2019-09-03 05:35:00, 2019-09-14 14:15:00,
2019-09-15 04:20:00, 2019-10-01 06:45:00, 2019-10-02 00:40:00,
2019-10-25 05:00:00, 2019-10-25 09:45:00, 2019-10-24 14:50:00,
2019-10-25 16:05:00, 2019-11-12 06:50:00, 2019-11-13 06:45:00,
2019-11-26 08:55:00, 2019-11-27 17:25:00, 2019-12-21 11:10:00,
2019-12-21 14:10:00

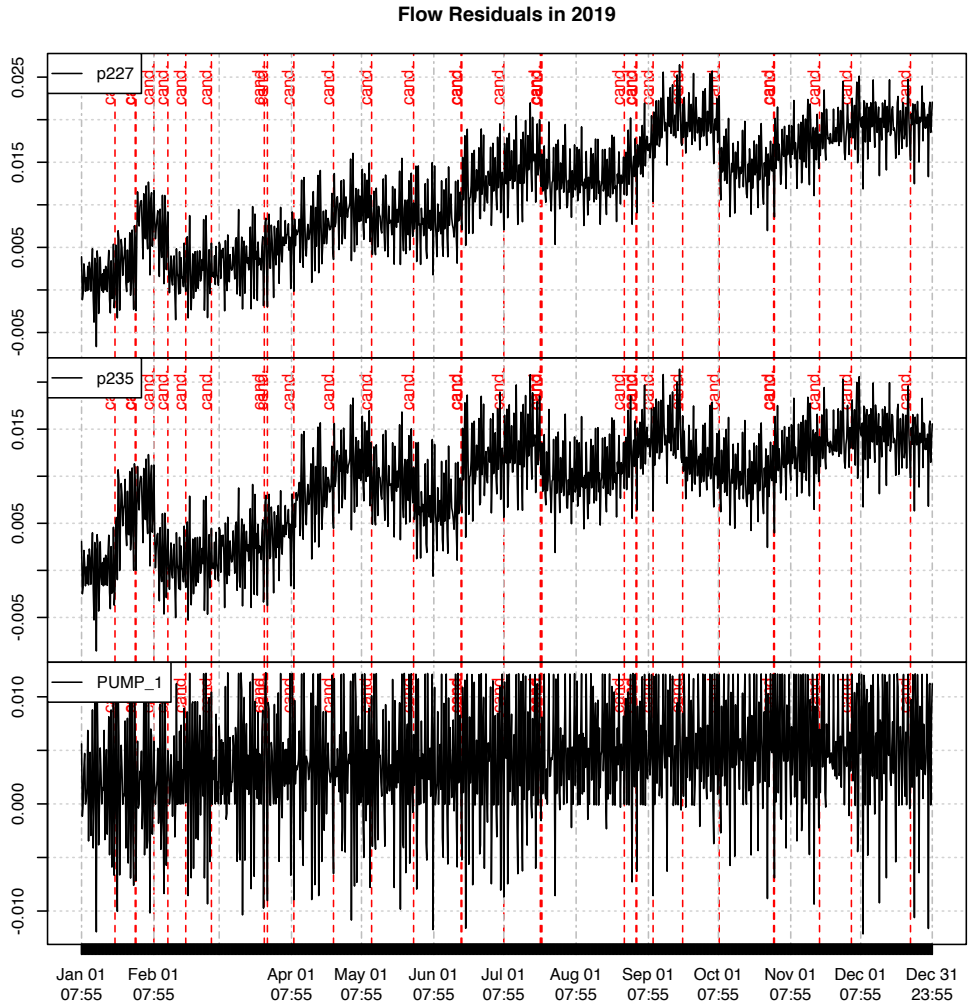
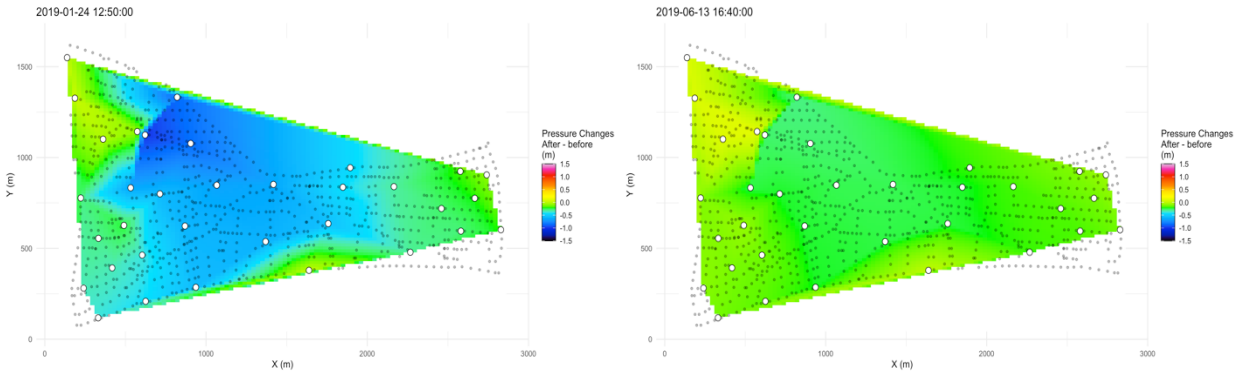
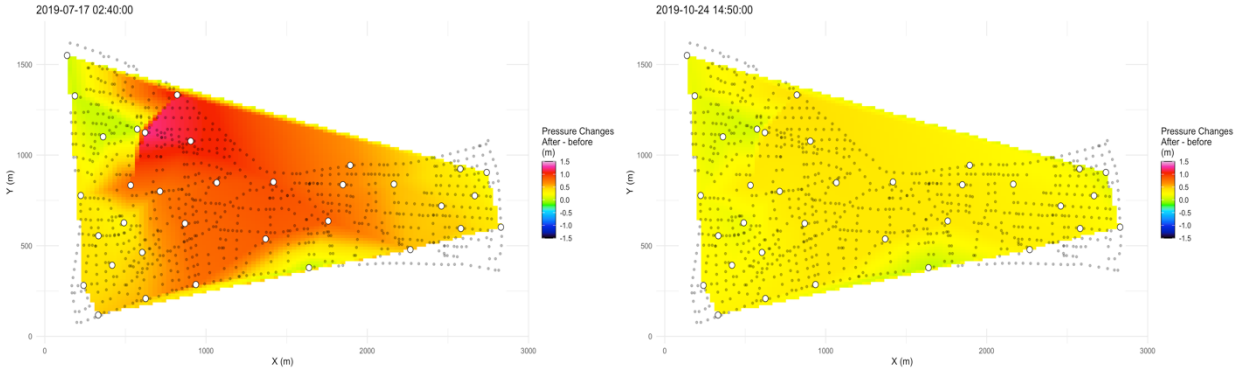


Figure 1. Time series of the flowrate residuals. Rough candidates from the ensemble multivariate changepoint detection algorithms were marked by red dash lines.



(a) (b)



(c) (d)

Figure 2. Pressure changes around (a) '2019-01-24 12:50:00', (b) '2019-06-13 16:40:00', (c) '2019-07-17 02:40:00', and (d) '2019-10-24 14:50:00'. Pipe midpoints were dotted in gray together with pressure sensors emphasized by circles. Temporally around the detected accurate moment candidates, bilinear interpolation was applied inside the convex hull of sensors. The changes in average levels before and after the changepoints were presented by the scale bar.

Table 2. Isolated leakage pipes. The two-sample one-sided Student's t-test with unknown variances was performed on the exponentially moving averaged short-term pressure time series of each pipe for groups that were prior and posterior to the accurate moment candidate.

Detected start time	Isolated pipe	X. Coord	Y. Coord	One-sided t-test p-value	Avg. pressure, before	Avg. pressure, after	Pressure shift
2019-01-14 17:35:00	p352	884.82	1295.43	0.0019	31.72	31.59	-0.14
2019-01-15 13:50:00	p346	647.59	1124.08	0.0000	37.01	36.20	-0.80
2019-01-24 12:55:00	p346	647.59	1124.08	0.0000	36.88	35.90	-0.98
2019-01-31 12:15:00	p346	647.59	1124.08	0.0002	36.63	35.74	-0.89
2019-02-25 08:15:00	p6	286.67	1281.28	0.0032	26.13	26.05	-0.08
2019-02-26 09:50:00	p239	560.92	1142.78	0.0006	5.43	5.40	-0.03
2019-03-21 22:25:00	p6	286.67	1281.28	0.0023	25.59	25.52	-0.08
2019-03-31 18:35:00	p239	560.92	1142.78	0.0004	5.37	5.30	-0.07
2019-04-19 06:00:00	p12	221.58	1318.85	0.0003	31.76	31.47	-0.28
2019-05-04 07:20:00	p352	884.82	1295.43	0.0001	31.42	31.06	-0.36
2019-05-05 23:55:00	p346	647.59	1124.08	0.0000	37.07	36.40	-0.67
2019-05-23 07:45:00	p12	221.58	1318.85	0.0002	31.94	31.69	-0.25
2019-06-13 09:45:00	p12	221.58	1318.85	0.0012	30.94	30.73	-0.22
2019-06-29 23:10:00	p346	647.59	1124.08	0.0000	36.43	35.50	-0.93
2019-07-17 09:40:00	p239	560.92	1142.78	0.0003	4.38	4.27	-0.11
2019-07-17 06:00:00	p352	884.82	1295.43	0.0021	32.48	32.14	-0.34
2019-08-21 04:55:00	p7	258.94	1361.98	0.0033	31.99	31.96	-0.04
2019-08-22 07:25:00	p352	884.82	1295.43	0.0000	30.58	29.98	-0.60
2019-08-26 06:55:00	p352	884.82	1295.43	0.0002	30.96	30.56	-0.40
2019-09-02 06:20:00	p352	884.82	1295.43	0.0001	31.38	30.88	-0.50
2019-09-03 05:40:00	p352	884.82	1295.43	0.0033	32.64	32.45	-0.19
2019-09-14 14:20:00	p352	884.82	1295.43	0.0083	29.78	29.64	-0.14
2019-10-01 06:50:00	p352	884.82	1295.43	0.0001	30.97	30.49	-0.47
2019-10-25 05:05:00	p12	221.58	1318.85	0.0005	31.77	31.53	-0.24
2019-10-25 09:50:00	p674	1693.87	646.09	0.0061	50.20	50.01	-0.19
2019-11-12 06:55:00	p352	884.82	1295.43	0.0000	30.56	30.17	-0.38

2019-11-13 06:50:00	p352	884.82	1295.43	0.0018	31.65	31.26	-0.39
2019-11-26 09:00:00	p352	884.82	1295.43	0.0005	30.65	30.31	-0.34
2019-12-21 14:15:00	p352	884.82	1295.43	0.0013	29.74	29.53	-0.20

The proposed hybrid approach employed both mechanistic model-based WNTR simulation and data-driven two-step EMCPD methods for leakage detection. Demand records data were assimilated to match the mechanistic WNTR model with the observations, and output residuals for detection. To decrease computation costs, the EMCPD was applied in two steps with different temporal resolutions, which avoided iteratively solving hydraulic simulations. During the isolation step, to be conservative and parsimonious at triggering alarms among the sparse sensor network, lower-order interpolation from observed pressure data with no extrapolation, as well as a low p-value threshold were adopted. The overall approach with limited assumptions on the distribution of the underlying dataset was both computationally efficient and conveniently transferable to other spatial-temporal datasets and mechanistic models. Future research could adopt spatial kriging methods to explore leakage isolation via interpolation among sparse sensors.

Keywords: Leakage detection and isolation; hybrid modeling; ensemble multivariate changepoint detection; bilinear spatial interpolation; hypothesis testing

References

[1] Klise, K.A., Murray, R., Haxton, T. (2018). An overview of the Water Network Tool for Resilience (WNTR), In Proceedings of the 1st International WDSA/CCWI Joint Conference, Kingston, Ontario, Canada, July 23-25, 075, 8p.
[2] Renka, R.J., 1996. Algorithm 751: TRIPACK: a constrained two-dimensional Delaunay triangulation package. *ACM Transactions on Mathematical Software (TOMS)*, *22*(1), pp.1-8.
[3] Bardwell, L., Fearnhead, P., Eckley, I.A., Smith, S. and Spott, M., 2019. Most recent changepoint detection in panel data. *Technometrics*, *61*(1), pp.88-98.
[4] James, N.A. and Matteson, D.S., 2013. ecp: An R package for nonparametric multiple change point analysis of multivariate data. *arXiv preprint arXiv:1309.3295*.
[5] Zhao, K., Wulder, M.A., Hu, T., Bright, R., Wu, Q., Qin, H., Li, Y., Toman, E., Mallick, B., Zhang, X. and Brown, M., 2019. Detecting change-point, trend, and seasonality in satellite time series data to track abrupt changes and nonlinear dynamics: a Bayesian ensemble algorithm. *Remote sensing of Environment*, *232*, p.111181.

SUMMARY

A hybrid approach based on benchmark simulation and the ensemble multivariate changepoint detection (EMCPD) is proposed to detect leakage occurrences. Bilinear bivariate spatial interpolation and Student's t-test is further invocated to isolate leakage sites.

First, a simulation was performed, assuming no leakage occurred to attain the benchmark working condition series of the L-town WDS. The pressure-dependent demand scheme was deployed with a simulation time step of five minutes. Upon simulation, the residuals of flowrates and pressures were calculated by contrasting the sensor measurements and the simulated results. The observed demand data were assimilated as the nodal pattern, and therefore the nodal demand residuals were omitted.

Based on the flowrate residuals, the EMCPD was performed to detect the rough time of leak events. At this stage, the flowrate residuals series were averaged to three observations per day to denoise short term fluctuation, keep diurnal and weekly cycles, and accelerate the computation.

Around each rough candidate, the EMCPDs were performed to distinguish accurate candidates. At this stage, both the flowrate residuals and pressure residuals series were engaged at five minutes level.

Finally, around each accurate candidate, the nodal pressures were calculated by using bilinear spatial interpolation based on monitored pressure time series. The two-sample one-sided Student's t-tests were performed to isolate the most likely sites. Since all pressures were estimated based on 34 sparse sensors, bilinear bivariate interpolation without extrapolation, as well as a p-value threshold of 0.01 were adopted, with the intention to avoid overfitting, to be conservative at un-monitored nodes, and to be parsimonious at raising alarms.

The overall approach with limited assumptions on the underlying dataset distribution was both computationally efficient and conveniently transferable to other spatial-temporal datasets and mechanistic models. Future research could adopt spatial kriging to explore leakage isolation via water pressure interpolation among sparse sensors.