Water supply network pollution source identification by random forest algorithm

Luka Grbčić, Ivana Lučin, Lado Kranjčević and Siniša Družeta

ABSTRACT

A novel approach for identifying the source of contamination in a water supply network based on the random forest classifying algorithm is presented in this paper. The proposed method is tested on two different water distribution benchmark networks with different sensor placements. For each considered network, a considerable amount of contamination scenarios with randomly selected contamination parameters were simulated and water quality time series of network sensors were obtained. Pollution scenarios were defined by randomly generated pollution source location, pollution starting time, duration of injection and the chemical intensity of the pollutant. Sensor layout's influence, demand uncertainty and imperfect sensor measurements were also investigated to verify the robustness of the method. The proposed approach shows high accuracy in localizing the potential sources of pollution, thus greatly reducing the complexity of the water supply network contamination detection problem.

Key words | machine learning, pipe network pollution, pollution source identification, random forest algorithm, water supply system contamination

Luka Grbčić (corresponding author)
Ivana Lučin
Lado Kranjčević
Siniša Družeta
Faculty of Engineering,
University of Rijeka,
Vukovarska 58, 51000, Rijeka,
Croatia
E-mail: *lgrbcic@riteh.hr*

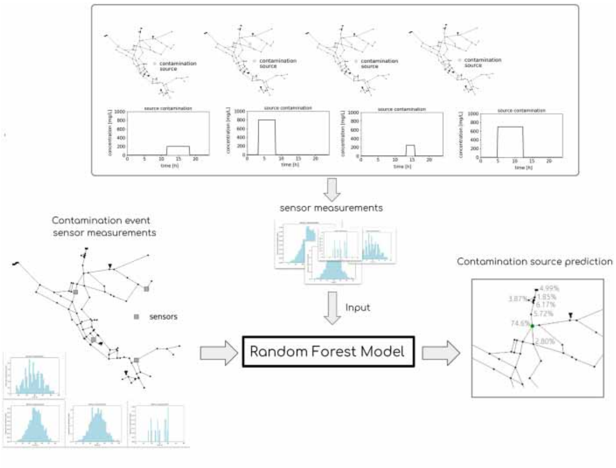
Luka Grbčić
Lado Kranjčević
Siniša Družeta
Center for Advanced Computing and Modelling,
University of Rijeka,
Radmile Matejčić 2, 51000, Rijeka,
Croatia

HIGHLIGHTS

- In case of a water network contamination event, the random forest machine learning algorithm trained on a great number of simulated contamination scenarios can provide rapid localization of contamination sources.
- With a greater number of inputs, greater model accuracy is achieved, and the only limit on the number of inputs are computational resources since the prediction model is prepared before a contamination event.
- It is shown that in most cases, around 10% of network nodes have a sum of 99.99% probability prediction, thus a considerable reduction of suspect nodes in the range of 90% can be achieved.
- Influence of different networks, sensor layouts and demand uncertainty showed that small variation in accuracy is present; however, in case of Boolean sensor imperfections, the prediction model significantly decreases in accuracy.

doi: 10.2166/hydro.2020.042

GRAPHICAL ABSTRACT



INTRODUCTION

The occurrence of contamination in water distribution networks presents a great concern due to potential major effects on human health and safety. Accidental or deliberate contamination of a water supply network with heavy metals or water-borne pathogenic organisms presents a danger which can cause serious health problems. A review of public health risks of such events is given in Besner *et al.* (2011). It is important to localize the potential source of pollution in case of such an event since the outcome could be dangerous for the human population. A number of recent studies investigated emergency reactions needed in case of such events (Shafiee *et al.* 2018; Strickling *et al.* 2020).

Contamination transport in water network distribution systems is a complex phenomenon due to diffusive processes of the contaminant and the eddying flow patterns created by various pipe network elements such as junctions and valves. Piazza *et al.* (2020) showed with experimental analysis that when fluid flow in the network is less turbulent, the contaminant diffusion is greatly enhanced and it was found that a transport model which incorporates these effects is more accurate than pure advection models. Grbčić *et al.* (2019) showed that complex contamination mixing which requires specific numerical models occurs in double-Tee junctions which are common building blocks of water distribution pipe networks. A solution to this problem was proposed in works by both Braun *et al.* (2015) and Grbčić *et al.* (2020a) where a combination of computational fluid dynamics with statistical and machine learning methods were applied for incomplete mixing modeling in double-Tee junctions.

An optimal placement of water quality sensors in a pipe network is essential for finding the source and dynamics of the water supply network pollution. Preis & Ostfeld (2008b) explored an optimal sensor placement in a water supply network in order to maximize the detection rate and minimize the detection time of contamination sources using Non-Dominated Sorted Genetic Algorithm-II (NSGA-II). In Ung *et al.* (2017), an adjoint source identification method, based on a backtracking algorithm, was used in conjunction with the Monte Carlo and optimization methods (greedy algorithm) in order to obtain the optimal sensor placement with maximizing the contamination source identification.

The data gathered by sensors coupled with optimization algorithms can be used to determine the most probable source of pollution, the pollution starting time, duration and the chemical concentration of the pollutant (Zechman & Ranjithan 2009; Kranjčević *et al.* 2010). A number of studies considered imperfect sensor measurements and uncertainty of water demand to more accurately mimic a realistic contamination scenario since the events are simulated by hydraulic and mixing models (Xuesong *et al.* 2017;

Yan *et al.* 2019). A detailed overview of optimization methods and general approaches for the identification of contamination sources is presented in Adedoja *et al.* (2018).

The approach where optimization algorithms are employed usually requires a significant amount of time to detect a contamination source, starting time and pollutant concentration, especially if a real size water distribution network is considered where a greater amount of suspect nodes increase the complexity of the optimization problem and in turn affect how fast the source is localized. This presents a problem since reaction time in case of contamination scenario is of the greatest importance.

An alternative to optimization algorithms for pollution source identification would be data mining. Data mining based search of contamination event source relies on a precompiled database of simulated contamination events results, unlike simulation-optimization methods. Optimization and statistical algorithms can be applied with data mining in order to search through the precompiled database for the most compatible contamination event parameters (source, concentration and duration), while simulation-optimization methods simultaneously run contamination event scenarios in conjunction with optimization algorithms and repeatedly evaluate a fitness function in order to detect the previously mentioned parameters. Huang & McBean (2009) developed a data mining approach in conjunction with a maximum likelihood method on a water distribution network with five sensors and showed that it is possible to reduce the search space of potential pollution sources. A drawback of this method would be that it is not a model and requires a large database of contamination scenario results. Similarly, in Shen & McBean (2011), an offline database was built by contamination simulation mining to correctly identify the pollution sources.

A model approach based on logistic regression was presented in Liu *et al.* (2012) and was coupled with an evolutionary algorithm for contaminant source detection. In Eliades *et al.* (2014), a parallel Monte Carlo based model in order to detect the source of pollution was developed. In studies by De Sanctis *et al.* (2008) and Perelman & Ostfeld (2013), a probabilistic approach based on Bayesian belief networks is explored for detecting the source of contamination in water supply networks.

A machine learning model presents another solution which would provide a fast prediction of possible contamination sources. In Kim *et al.* (2008), an artificial neural network (ANN) was trained on data from a small pipe network with five sampling locations to find the source of *E. coli* pollution. Even though the approach was tested on a small network where the complexity of the problem is reduced, a success rate of 87% was achieved. In Rutkowski & Prokopiuk (2018), a learning vector quantization (LVQ) neural network was used to find a sub-zone in a water supply network where a potential source of pollution would be located. It was found that the success of the LVQ data-driven model greatly depends on the number of monitoring stations in the water supply network. Applying machine learning tools and testing their efficiency have already been researched in groundwater pollution source identification (Singh & Datta 2007; Bashi-Azghadi *et al.* 2010; Rodriguez-Galiano *et al.* 2014); however, application in pollution source identification in water supply networks is open to be researched. In this study, we present a pollution source identification approach which utilizes the random forest (RF) classifying algorithm. The RF method belongs to the decision tree (DT) family of machine learning algorithms, and DTs were previously used in Eliades & Polycarpou (2011) for the purpose of contamination source area isolation. The RF model approach was explored in Lee *et al.* (2018) where it was applied for identifying the source of contamination based on sensor network observations in a river system. The contamination events were simulated, and with obtained data, an RF model was trained in order to detect the source location. In Rodriguez-Galiano *et al.* (2014), the RF method was applied (trained with measured data) for predictive modeling of vulnerable areas which are prone to groundwater nitrate pollution. Wang *et al.* (2015) also trained RF models with measured data for the purpose of identifying multi-source heavy metal pollution. Recently, Grbčić *et al.* (2020b) developed a dynamic learning algorithm for massively parallel systems which couples ANNs in a tournament style selection for search space reduction and the RF model regression analysis which ranks the potential contamination source nodes in water supply networks.

Data for the RF training were obtained by simulating randomly generated contamination scenarios on two benchmark networks in the pipe network hydraulic and water quality analysis software EPANET2. In this work, due to large distances between double-Tee junctions in the pipe networks, and due to dominantly turbulent flow in the pipes, the EPANET2 (Rossman 2000) complete mixing model is used for contaminant transport modeling. The complete mixing model calculates the chemical concentration at a pipe network junction with flow weighted chemical concentration at the junction inlet pipes. The calculated chemical concentration is then equally distributed at all junction outlet pipes. The randomized variables for each scenario were the pollution source location (network node), contamination start time, contamination duration and the chemical concentration value. The sensor measurements were the input data for the RF classifier, while the output was the location of the pollution source. For both benchmark networks, different sensor placements were considered to investigate the source identification ability of the proposed method on different network configurations. Additionally, the RF model was trained with input data that include demand uncertainty to investigate the robustness of the method in a more realistic case. Imperfect sensor measurements were also separately considered to assess the accuracy of the proposed method.

MATERIALS AND METHODS

Data generation

Two benchmark networks, Net3 EPANET2 example and Richmond network (Van Zyl 2001), are selected for testing the RF classifier for the purpose of water supply pollution source identification. Contamination scenarios are simulated in EPANET2 with a single contamination injection node. The flow paced method was used for contaminant injection, and the contaminant is modeled as non-reacting constituent. In order to train the RF model and achieve significant accuracy, a great number of contamination scenarios are simulated to cover all marginal possibilities.

The benchmark network Net3 consists of 92 nodes. Simulation time is set as 24 h with a hydraulic time step of 1 h, quality time step 5 min and pattern time step 10 min. Two different water quality sensor layouts in the Net3 benchmark network are considered. The first sensor layout was taken from the study by Preis & Ostfeld (2007) (sensors positioned in nodes 117, 143, 181 and 213), while the other from Zechman & Ranjithan (2009) (sensors at nodes 115, 119, 187 and 209). For both layouts, the investigation of different number and sensor placements was conducted. A summary of all investigated combinations is given in Table 1. The Net3 network layout with both sensor placements can be seen in Figure 1.

The Richmond network with 865 nodes is obtained from the Centre for Water Systems (CWS) at the University of Exeter CWS (CWS, U.o.E.). Simulation time is 72 h with a hydraulic time step of 1 h, quality time step 5 min and pattern time step 1 h. Two different sensor layouts with several different combinations are considered. The first considered sensor layout was taken from Preis & Ostfeld (2007) with five sensors which are positioned in nodes 123, 219, 305, 393 and 589. The second sensor layout was taken from Preis & Ostfeld (2008a) (sensors at nodes 93, 352, 428, 600 and 672). A summary of all investigated combinations of both sensor layouts can be found in Table 7. The Richmond network layout with a marked detail can be seen in Figure 2(a), and both sensor placements can be seen in Figure 2(b).

For each considered network, contamination scenarios are simulated with randomly chosen pollution source location, contamination start time, contamination duration and chemical concentration value. Pollution source location is randomly selected from the set of all network nodes, contamination start time is randomly defined from 0 to 24 h for both Net3 and Richmond network, and contamination duration is randomly chosen from 10 min to 24 h for Net3 and

Table 1 | Net3 network RF model results

		Model accuracy		
Sensors placement	**Number of model inputs**	**Top 3**	**Top 5**	**Top 10**
117, 143, 181, 213	3,422,803	91.70%	96.80%	99.60%
115, 119, 187, 209	2,687,695	91.00%	96.40%	99.0%
117, 181, 213	3,043,055	90.80%	96.20%	99.50%
117, 143, 213	106,188	88.00%	94.20%	98.50%
115, 187, 209	2,686,223	90.90%	95.70%	99.00%
115, 119, 209	2,681,045	90.50%	95.40%	98.90%
117, 213	2,726,280	86.40%	93.10%	98.00%
119, 209	2,666,258	90.20%	95.10%	98.80%

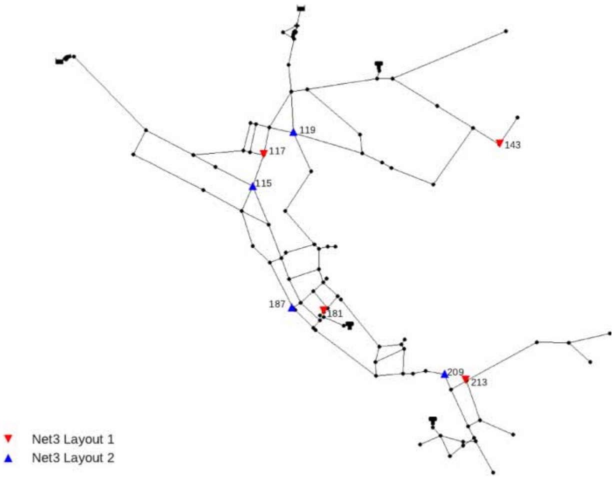


Figure 1 | Net3 network layout with sensor positioning. Net3 Layout 1 (layout by Preis & Ostfeld (2007)) and Net3 Layout 2 (layout by Zechman & Ranjithan (2009)).

10 min to 24 h for the Richmond network. The possible chemical concentration value of the contaminant is chosen from the range 0–1,000 mg/L for both networks and is kept constant during the contamination duration time.

For the Net3 benchmark network, the sensors recorded the quality of water every hour for a total of 24 h (25 measurements per sensor) which equals a total of 100 measurements per simulation for four sensors in the network. For the Richmond network simulation, the sensors recorded the quality every hour for 72 h (a total of 73 measurements per sensor) even though the contamination was only set to start in the period of the first 24 h of the pollution scenario, resulting in 365 measurements per simulation for five sensors in the network.

RF classifier

The RF algorithm was introduced by Breiman (2001), and it is an ensemble learning algorithm that uses multiple decision trees for training which are constructed with random subsets of features. Bootstrap aggregation or bagging technique was used for the RF model training procedure which greatly enhances the de-correlation of each randomly constructed DT.

A great feature of the RF algorithm is that with the introduced randomness, it manages to create decision trees with low variance, and hence, the possibility of model overfitting is reduced. Generally, the most important RF algorithm parameter is the number of trees used for model training. The bigger the number of trees is, a more robust prediction is achieved. Each DT constructed with random features

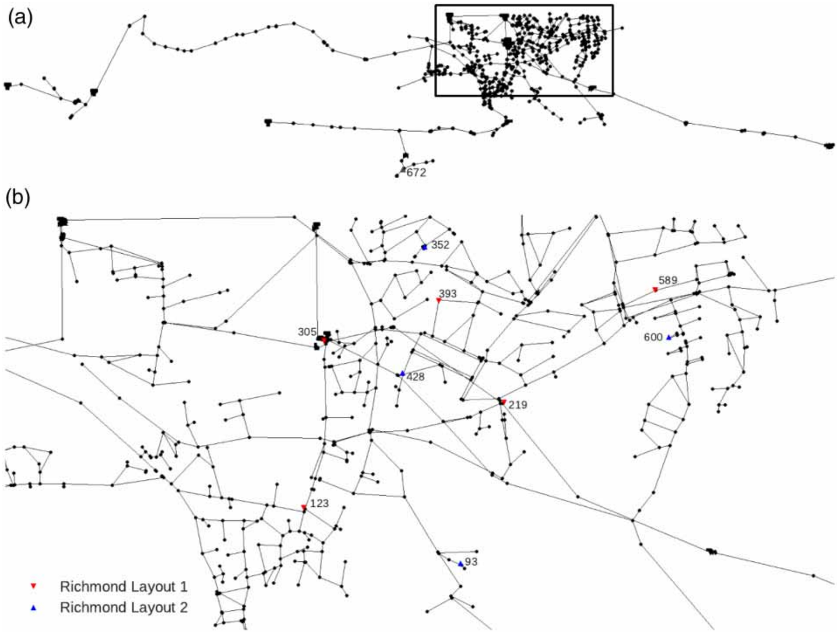


Figure 2 | Richmond network layout with Richmond Layout 1 (layout by Preis and Ostfeld (2007)) and Richmond Layout 2 (layout by Preis and Ostfeld (2008a)). (a) Richmond network with marked detail. (b) Richmond network detail with sensor positioning.

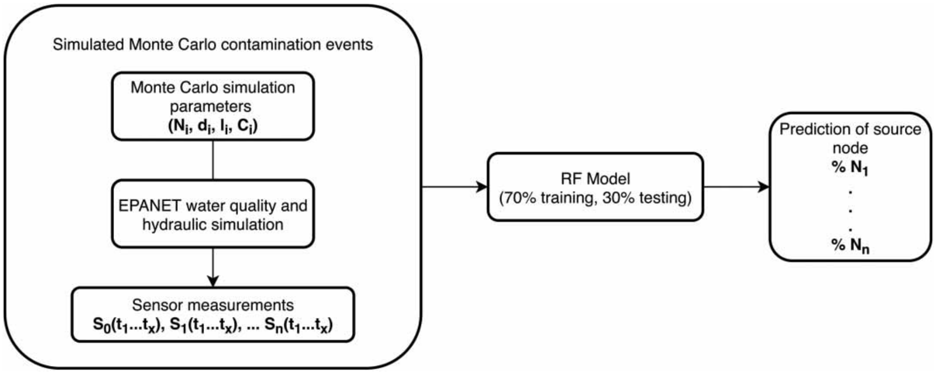

Simulated Monte Carlo contamination events
Monte Carlo simulation parameters
(N_i, d_i, I_i, C_i)
EPANET water quality and hydraulic simulation
Sensor measurements
$S_0(t_1...t_x), S_1(t_1...t_x), ... S_n(t_1...t_x)$
RF Model
(70% training, 30% testing)
Prediction of source node
% N_1
.
.
.
% N_n

Figure 3 | Flowchart of the RF model creation.

generates a prediction, and the final outcome or result of the RF model is the prediction that was achieved by the majority of the random decision trees. The RF implementation in the Python ML library Scikit-learn 0.23 (Pedregosa *et al.* 2011) was used, and it should be noted that in that implementation, the probabilistic prediction of each classifier (DT) is averaged instead of using a vote counting approach where each class is counted to form a final prediction.

For both networks, the number of estimators (trees) was set to 250, the minimum number of samples required to split an internal node was set to 20, and all other parameters were set to default. The maximum number of random features used for constructing a single DT was set to be as the square root of the total number of features or trees in the forest. All RF model parameters were defined with the grid search hyperparameter optimization. The selected number of trees was the most influential parameter in accurate prediction of the class. A bigger number of trees enhanced the success rate of the model, but after the tuned value of 250 trees, the success rate remained unchanged.

The input data for the RF classifier were the results from a large number of Monte Carlo simulations, where sensor water quality readings through a period of time described in the previous subsection are input features. For both networks, contamination scenarios where the sensors did not detect any pollution (0 mg/L of contaminant through the measurement period) were removed. The number of input features varied due to the number of sensors considered for the prediction model, where for the Net3 network it was 25 features per sensor and for the Richmond network 73 features per sensor.

In Figure 3, the steps of creating the RF model can be seen. After the randomized simulation parameters with pollution source node $N_i \in \{N_1, .., N_n\}$ ($N_n = 92$ for Net3 and 865 for Richmond), contamination duration $d_i \in \{0, 1, \ldots, 24\}$, contamination start time $I_i \in \{0, 1, \ldots, 24\}$ and a continuous contaminant chemical concentration $C_i \in \{0, \ldots, 1000\}$ were set, an EPANET2 network water quality and hydraulic analysis simulation was done. The sensor measurements $S_i \in \{S_1, .., S_n\}$ through time were recorded. From all of the obtained measured data, 70% was used for the RF model training and 30% for model testing, both testing and training data were randomly selected. The output of the RF model is the prediction where every network node has been assigned a probability of being the true contamination source node.

Demand uncertainties

To examine the proposed approach for a more realistic water supply network case, network nodes demand uncertainties have been taken into account. Node demand uncertainties have been introduced to this whole procedure through generating new machine learning (ML) model training input and output data for both studied networks.

In Figure 4, the flowchart of the algorithm for adding demand uncertainties into the Monte Carlo simulation process can be seen. For each network node, a random Boolean value (True, False) is generated. If the network node has been assigned a False Boolean value, the base demand of the original network for that certain node remains unchanged (original demand is used). If a True Boolean

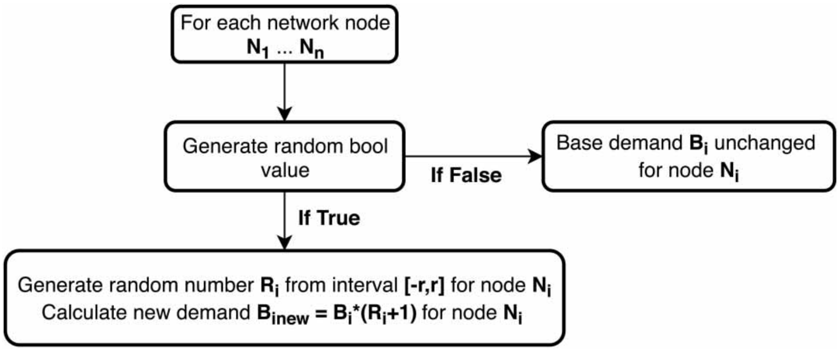

For each network node
N1 ... Nn
Generate random bool value
If False
Base demand Bi unchanged for node Ni
If True
Generate random number Ri from interval [-r,r] for node Ni
Calculate new demand Binew = Bi*(Ri+1) for node Ni

Figure 4 | Flowchart of contamination event hydraulic demand uncertainty generation.

value has been generated for a node, then for that node, a random number $R_i \in \{-r, .., r\}$ is generated and a new base demand $B_{i_{new}}$ is calculated based on the term $B_{i_{new}} = B_i \cdot (R_i + 1)$, where B_i is the original demand value. It can be observed that for the nodes with greater base demand, a greater demand variation will be achieved. This whole procedure was incorporated into the process of generating an EPANET input file along with random simulation parameters such as contamination source node, contamination duration and start time (as seen in Figure 3).

RF model accuracy was tested on for both networks with $r = 0.05$ which entails a possible maximum deviation of ±5% from the base demand. Additionally, a higher deviation value of $r = 0.2$ was investigated. For both values of r and networks, a total of 3 million input was generated for model training and testing.

Imperfect sensor measurements

A perfect sensor measurement of a chemical contaminant in a case of a water supply contamination event would mean acquiring an accurate value of concentration through a certain time interval. Boolean and fuzzy sensors which make imperfect measurements (Preis *et al.* 2007) are also considered in this study.

Boolean sensors set a value of 0 if there is no detected contamination or they set a value of 1 in case there is a registered contamination through a time interval. This entails that the input features (sensor measurement data) for the RF model would be a set of zeroes and ones.

Fuzzy sensors detect either low, medium or high contaminant concentration. Since the maximum set value of chemical concentration was set to 1,000 mg/L, a low contamination was considered if the measured chemical concentration value C_i was $0 < C_i < 100$ mg/L, medium contamination was $100 \leq C_i < 500$ mg/L and high if $C_i \geq 500$ mg/L. RF model input features for the fuzzy sensors were defined as 0 if no contaminant was detected, 1 for low measurements, 2 and 3 for medium and high measurements.

RESULTS AND DISCUSSION

Net3 network

The pollution source detection problem is a multimodal problem, where multiple different contamination parameters can give similar sensor measurements. Since it is critical to localize the source of pollution, it is necessary to assure that there is no possibility that a true contamination source node is eliminated from further examination. Due to this, multiple suspect contamination source nodes, with the greatest probabilities to be the true source nodes, are considered. For the Net3 network model with the most sensors (layout by Preis & Ostfeld (2007)), 73 out of 92 nodes were potential contamination sources (all tanks, rivers and lakes were included as sources) after scenarios where contamination was not detected were removed, which means that the RF classifier had to predict 1 of 73 classes. It is observed that for layouts with a smaller number of sensors, a smaller number of classes (potential contamination sources) remain after removing scenarios in which contamination was not detected. This is consistent with the fact that a greater number of sensors usually provide a greater detection rate of contamination events.

An analysis of the influence of a number of inputs on model accuracy is conducted for the Net3 network with four sensors (sensor layout by Preis & Ostfeld (2007)). In Figure 5, the relationship between the number of RF model inputs and the number of top suspect nodes needed for achieving a success rate or RF model accuracy of 99% is shown. For 10,000 simulation results or RF model inputs, 99% accuracy of true contamination source detection is achieved with the top 25 nodes which means that there is a 99% chance that one of the 25 suspect nodes is the true contamination source. Naturally, when the number of data inputs is increased, the accuracy of 99% is achieved with a smaller number of top suspect nodes. With 3 million inputs, a 99% accuracy is achieved with a top 8 set of suspect nodes, and for further model studies, this was a targeted number of inputs.

Model accuracy was assessed from 30% of the testing data, and the model prediction was deemed successful if the true contamination source node was in the top (3, 5 and 10) of the predicted contamination source nodes. To ensure that the RF training was stable and that the success rate does not greatly deviate after each trained model prediction run, the Net3 network model (with a randomly selected 70% testing and 30% training input data split) was repeatedly run 50 times and the success rate standard deviation for top 5 nodes prediction was 0.003

In Table 1, the Net3 network model results for different sensor layouts can be observed. The number of sensors and their placement are varied, and consequently, the number of model inputs is also different for each sensor layout. The total number of Monte Carlo obtained input and output data was 4.9 million for both sensor network layouts presented in Figure 1. The number of model inputs shown in Table 1 are the ones for which the placed sensors successfully detected the transported contaminant. With a reduced sensor number in the network, there is a smaller contamination detection rate, thus model inputs decrease since there is a greater number of scenarios that do not detect contamination and are removed. Naturally, it can be seen that with more data inputs, the model accuracy is higher with the true contamination source node being in the top 10 suspect nodes for 99.6% of the testing data for the four sensor layouts of Net3, while the two sensor layouts achieved the top 10 for 98% of the testing data. It can also be observed that the data obtained by the sensor layout (and all combinations) proposed by Zechman &

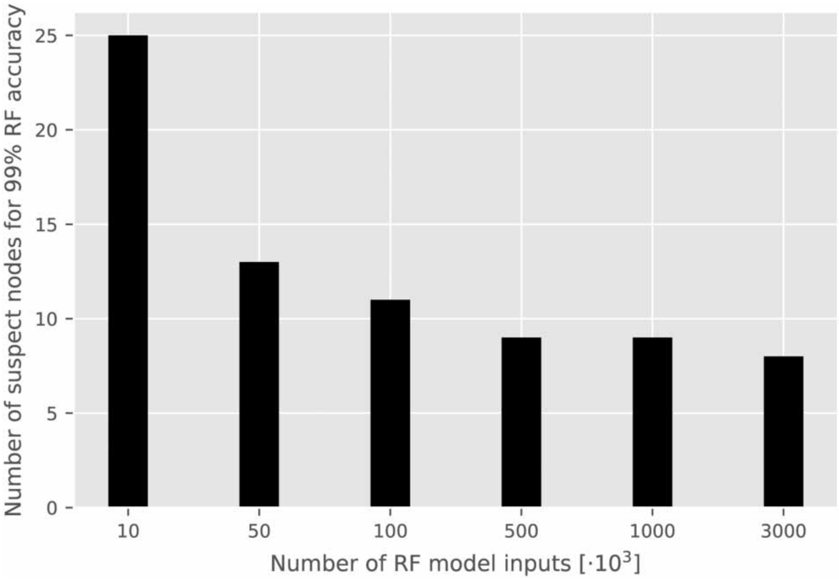


Figure 5 | Net3 network (sensor layout by Preis and Ostfeld (2007)) number of RF model inputs needed for 99% accuracy.

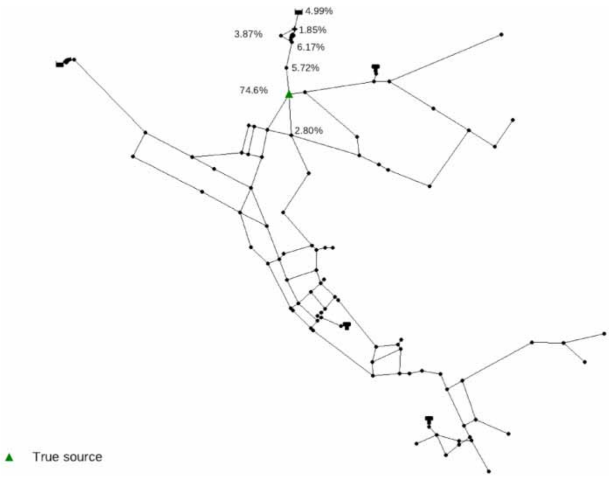


Figure 6 | Net3 network predicted suspect source nodes probabilities with the true source marked.

Ranjithan (2009) are similar in accuracy as the layout by Preis & Ostfeld (2007) even though the number of model inputs is smaller for the latter. Table 2 shows the number of top suspect nodes that contain the true contamination source with a certainty of 99% for each sensor placement. It can be observed that regardless of the sensor layout, the accuracy of the model is similar for all considered cases. To achieve 99% probability that the true source node is recognized, it is necessary to include top 8 suspect nodes for layout with 4 sensors (Preis & Ostfeld 2007) and top 14 suspect nodes for layout with 2 sensors. It is a reduction of 92 and 86% of the total number of nodes, respectively. This fact shows that a considerable reduction of search space can be achieved with the proposed method.

Table 2 | Net3 network RF model with the indicated number of potential source nodes for 99% accuracy and the total number of remaining suspect nodes

	Suspect nodes	
Sensors placement	**Top nodes for 99% accuracy**	**Total number**
117, 143, 181, 213	8	73
115, 119, 187, 209	9	57
117, 181, 213	8	64
117, 143, 213	12	67
115, 187, 209	10	57
115, 119, 209	11	57
117, 213	14	58
119, 209	11	57

It must be noted that with different sensor layouts (Preis & Ostfeld 2007), different number of classes or total number of suspect nodes is distinguished, i.e. for some source nodes, no contamination scenario is detected whatsoever, therefore the number of remaining classes is smaller than the total number of network nodes. With a greater number of sensors, a greater number of model output classes is present (for 2 sensors 58 classes and for 4 sensors 64 classes). Besides, a sensor placement also greatly contributes to the detection rate, where 1 layout with 3 sensors (117, 143 and 213) detects 67 classes, which is 3 more than the other layout with 3 sensors (117, 181 and 213). This does not influence the RF model efficiency, but it shows that the RF model can indicate sensor layout detection rate and also completely eliminate some nodes as possible sources of pollution. All combinations of the sensor layout proposed by Zechman & Ranjithan (2009) create an RF model with the same number of total output classes (57) which is ultimately less than all values that were obtained by Preis & Ostfeld (2007).

An example of the chosen contamination scenario with four sensors with a probability assigned to suspect nodes can be seen in Figure 6. This prediction was made with the RF model with 3.4 million inputs with data from all four sensors of Net3. The true source node was assigned a 74.6% probability of being the true source node by the RF model. Other suspect nodes with the greatest probabilities are topologically in the vicinity of the true source node which shows that the RF model correctly localizes the broader contamination source area. This is especially important when several suspect nodes are assigned equal probabilities for being the true contamination source. The RF model indicates all probable source locations which need to be further considered narrowing down the suspect area.

An analysis was done on the Net3 network with the four sensor layouts by Preis & Ostfeld (2007) with a total of 3,422,803 RF model inputs in order to show the ranking of the true source node for the 30% (1,026,841) of the RF model testing data. Table 3 shows that the true source node is dominantly ranked first for the RF model testing data and it can be observed that for 88.38% of all the training data, the true source node is ranked in the top 4 of the predicted source nodes.

Table 3 | Net3 network prediction of true source node for RF testing data

Rank	Number of times	%
1	729,283	71.02
2	147,482	14.36
3	64,768	6.31
4	30,837	3.00
5 +	54,471	5.31

Percentage indicates the number of times the true source node obtained the given rank.

Table 5 | Net3 network model predictions for both perfect and imperfect input data (Boolean and fuzzy)

	Model accuracy			
Sensor type	Top 3	Top 5	Top 8	Top 10
Perfect	91.70%	96.80%	99.00%	99.60%
Fuzzy	81.90%	92.30%	97.90%	99.20%
Boolean	68.57%	84.40%	94.50%	97.50%

The number of model inputs was 3.4 million for all three sensor types.

The investigation of demand uncertainty for the Net3 network with four sensors (layout by Preis & Ostfeld (2007)) can be found in Table 4. It can be observed that the model accuracy decreases with the increase of demand uncertainty, which is expected. However, for demand uncertainty of ±5%, the model accuracy only slightly decreases and for ±20%, which represents a variation range of 40%, accuracy decreases by only several percent. Also, it must be observed that when a greater number of top nodes is considered, the difference in model accuracy decreases. This shows that the prediction model can be used in contamination events when demand uncertainty is present to indicate suspect nodes.

The influence of sensor imperfection for the Net3 network with four sensors (layout by Preis & Ostfeld (2007)) can be found in Table 5. The number of inputs was 3.4 million. It can be observed that the influence of sensor imperfection greatly influences model accuracy. For the Boolean type of sensors, accuracy decreases by 23% and for the fuzzy type of sensors, accuracy decreases by 10%. However, same as in the case of demand uncertainty, when a greater number of top nodes is considered, the difference in model accuracy decreases, which indicates that even with the Boolean type of sensors, the proposed machine learning approach can successfully narrow down the number of potential source nodes.

Table 4 | Net3 network model predictions for input data with demand uncertainty

	Model accuracy			
Demand uncertainty	Top 3	Top 5	Top 8	Top 10
±0%	90.70%	96.40%	99.00%	99.57%
±5%	88.00%	95.50%	98.70%	99.30%
±20%	86.50%	94.28%	98.10%	99.00%

The number of model inputs was 2 million (out of 3 million Monte Carlo data) for all three cases.

Richmond network

For the Richmond network (layout by Preis & Ostfeld (2007)), after scenarios where pollution is not detected were removed, the number of output classes was 163 (out of 865) but also varied for different sensor placements. The relationship between the number of inputs and the number of top sources for 99% accuracy is shown in Figure 7. For the Richmond network with an increase of input data, the number of nodes to achieve 99% accuracy is reduced from 60 nodes for 10,000 inputs to 15 nodes for 1.5 million inputs. The number of needed RF model inputs indicates (both for Net3 and Richmond) that this is big data manipulation and it is quite computationally demanding.

In Table 6, the number of top suspect nodes for 99% accuracy is shown and it can be observed that all three sensor placements had a total number of 163 (sensor layout by Preis & Ostfeld (2007)) suspect nodes (or output classes of the RF model). The potential source node search space reduction is around 91% for both five and four sensor layouts and 88% for the three sensor layouts which is a significant reduction of problem complexity. The sensor layout proposed in Preis & Ostfeld (2008a) is superior to the one proposed in Preis & Ostfeld (2007) due to the fact that the total number of output classes of the RF model is much higher (the maximum being 352 as seen in Table 6).

The results of the RF model prediction for the Richmond network with different sensor placements are presented in Table 7. The total number of Monte Carlo input data for

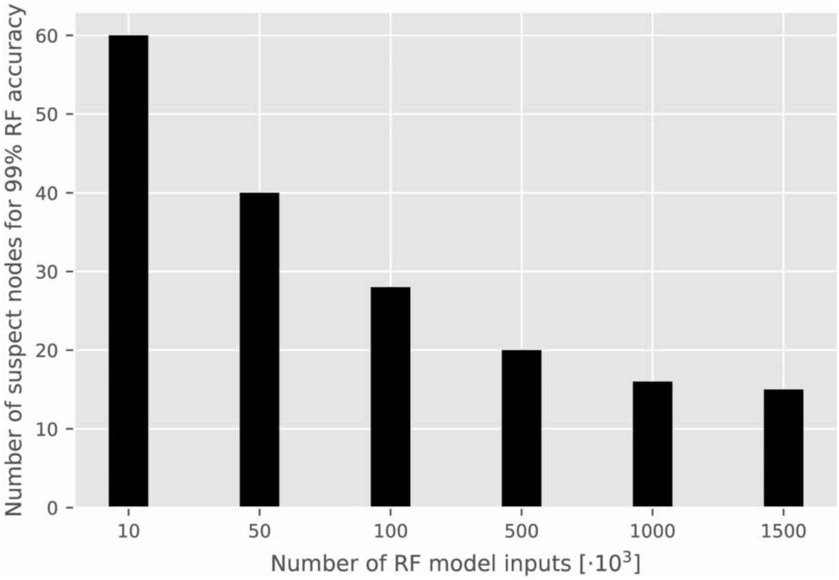


Figure 7 | Richmond network (sensor layout by Preis and Ostfeld (2007)) number of RF model inputs needed for 99% accuracy.

Table 6 | Richmond network RF model with the indicated number of potential source nodes for 99% accuracy and the total number of remaining suspect nodes

	Suspect nodes	
Sensors placement	**Top nodes for 99% accuracy**	**Total number**
123, 219, 305, 393, 589	15	163
93, 352, 428, 600, 672	18	352
123, 219, 393, 589	15	163
93, 428, 600, 672	25	301
123, 393, 589	20	163
352, 600, 672	25	312

Table 7 | Richmond network RF model results

		Model accuracy		
Sensors placement	**Number of model inputs**	**Top 3**	**Top 5**	**Top 10**
123, 219, 305, 393, 589	1,734,507	74.00%	84.00%	94.00%
93, 352, 428, 600, 672	3,785,650	80.20%	88.90%	96.30%
123, 219, 393, 589	1,729,783	73.00%	83.00%	94.00%
93, 428, 600, 672	3,205,674	79.60%	88.40%	95.70%
123, 393, 589	1,720,815	72.80%	83.00%	93.00%
352, 600, 672	3,299,763	79.80%	87.10%	94.40%

both investigated layouts was 12 million, and the number of model inputs is smaller than the one for the Net3 network. The sensor layout proposed by Preis & Ostfeld (2008a) is much more efficient in detecting contamination events since the number of model inputs is greatly higher than the one which was obtained by the sensor layout presented in Preis & Ostfeld (2007). The RF model accuracy is increased with the number of sensors and inputs even though the discrepancy is not large for all investigated layouts (around 8% maximum difference in prediction for a different number of sensors and model inputs). The top suspect nodes RF model accuracy varies greatly depending on how many top nodes are selected, and this can be attributed to the higher complexity of the network. For comparison, in the Net3 model, the accuracy of the true source node being in the set of the top 10 suspect nodes is 98% for the most sparse layout, while in the Richmond model, the top 10 suspect nodes include the true source node with an accuracy of 96% for the best possible sensor layout. Same as for Net3, the RF model training and testing was repeated 50 times (top 5 suspect nodes prediction) and it was found that the standard deviation of the prediction was 0.0056, showing that the model is stable.

The true source node ranking for the Richmond network with five sensor layouts by Preis & Ostfeld (2007) with a total of 1,734,507 RF model inputs (30% was used for testing) is given in Table 8. It can be seen that the true source node is also mostly ranked first but not as dominantly as for Net3 (Table 3). For 80.49% of the training data, the true source node achieved the top 4 node ranking.

The results for the Richmond network with four sensors (layout by Preis & Ostfeld (2007)) with added demand uncertainty can be found in Table 9. Similar to the case for the Net3 network, the model accuracy decreases with the increase of demand uncertainty. When a greater number of top nodes is considered, the influence of demand uncertainty decreases, i.e. model accuracy is slightly decreased.

The results for imperfect sensors for the Richmond network with five sensors (layout by Preis & Ostfeld (2008a)) can be found in Table 10. The number of inputs was 3.4 million. Same as in the case for Net3 network, it can be observed that the sensor type greatly influences model accuracy. When the top 3 nodes are considered, model accuracy decreases by 50% when sensor type changes from perfect to Boolean. When a greater number of top nodes is considered, model accuracy increases. The fuzzy type of the sensor performs better than the Boolean type; however, a considerable reduction in model accuracy can be observed when comparing with perfect sensors. Results indicate that with greater sensor imperfections, a greater number of suspect nodes must be further considered to assure that the true source node is included in the list of suspect nodes.

Table 8 | Richmond network prediction of true source node for RF testing data

Rank	Number of times	%
1	249,987	48.00
2	85,557	16.44
3	49,355	9.48
4	34,191	6.57
5 +	101,263	19.51

Percentage indicates the number of times the true source node obtained the given rank.

Table 9 | Richmond network model predictions for input data with demand uncertainty

	Model accuracy			
Demand uncertainty	Top 3	Top 5	Top 8	Top 10
±0%	64.36%	77.00%	87.05%	91.38%
±5%	60.40%	73.97%	84.00%	89.29%
±20%	58.30%	72.30%	83.60%	88.45%

Number of model inputs was 412,000 (out of 3 million Monte Carlo simulations) for all three cases.

Table 10 | Richmond network model predictions for both perfect and imperfect input data (Boolean and fuzzy)

	Model accuracy			
Sensor type	Top 3	Top 5	Top 8	Top 10
Perfect	80.20%	88.90%	94.50%	96.30%
Fuzzy	52.00%	65.40%	77.91%	81.20%
Boolean	30.85%	42.40%	54.80%	60.00%

The number of model inputs was 3.7 million for all three sensor types.

Efficiency of the RF classifier

The benefit of the proposed method is that the RF model can be continually updated with the new simulation data. Although model training requires a substantial amount of time and computer resources, it must be noted that model training processes are conducted prior to the contamination event. When contamination occurs, the prediction of the source node can be obtained in 0.1 s on average for both the Net3 and the Richmond network. The Net3 RF model training was done using the supercomputing resources at the Center for Advanced Computing and Modelling, University of Rijeka. Training the RF model for the Net3 network with around 3.4 million inputs took 1 h on 1 Intel E7 fat node with 6 TB of RAM, while the Richmond network RF model (with 1.7 million inputs) training took 2 h. The Richmond RF model took longer due to higher complexity (365 input features and 163 output classes). It was found that the major requirement for both RF models was available RAM capacity, meaning that training must be done in a high-performance computing environment, while the RF model run can be performed on a PC and the top suspect nodes can be quickly generated.

Limitation and extent of the method

It is shown that the proposed method has great robustness since it can be applied for different networks and different sensor layouts. Also, a great accuracy is achieved even when hydraulic demand uncertainties are taken into account. However, uncertainties in the hydraulic model such as pipe roughness, diameters and unknown valve states should be explored in further work. The proposed approach assumes a single contamination injection scenario, thus a formulation that considers multiple contamination injections nodes should be explored in future work.

In the study for the Net3 benchmark network (92 nodes), 3.4 million inputs were used for model training, while for the Richmond network (865 nodes), 1.7 million inputs, which is around 37,000 and 2,000 times greater number of inputs than the number of water supply network nodes, respectively. While both RF models provided good accuracy even for a smaller number of inputs, due to easy parallelization of Monte Carlo simulations and usage of high-performance computing, it is possible to obtain a greater number of inputs to achieve greater accuracy. In realistic cases, water supply networks can have many more nodes than the benchmark network presented in this paper, which could require a greater amount of model training input data. However, it must be noted that due to a sparse sensor placement, the number of model output classes could be significantly smaller than the number of network nodes in those cases. Therefore, the methodology for obtaining the Monte Carlo simulation results needs to be adjusted for such large water distribution networks. Also, in such networks, a greater number of remaining suspect nodes could be found due to very dense node placement in some areas as the problem is of multimodal nature.

CONCLUSION

In this paper, a method for identifying the source of contamination in a water supply network with a machine learning prediction model was presented. The proposed method was tested on two different water distribution benchmark networks with different sensor placements. For each considered network, a considerable number of contamination scenarios with randomly selected contamination parameters were simulated and water quality time series of network sensors were obtained. The generated big data were used as the input for the RF classifier to predict the contamination source node. It is shown that the RF model presents good accuracy in true contamination source prediction, with the accuracy increase tendency as the amount of input data grows. Since the RF model inputs are simulation results, additional training data can always be obtained with the only limiting factor being computational resources. In this manner, the RF model prediction accuracy can be improved.

The model is tested on two realistic complex benchmark cases. Due to the multimodal nature of the problem, all possible contamination sources are sorted with the top suspect nodes given the greatest probabilities. Although the training of the model requires substantial computational resources, in case of a contamination event, predicting the possible contamination sources is rapid and for every investigated case, 99% accuracy is obtained with a reasonable number of top contamination source node candidates, thus greatly reducing the overall time of the contamination source identification problem in the water supply network in urgent situations. The proposed approach was tested with input data which includes network nodes demand uncertainty and it was shown that the RF method achieved good accuracy for both water supply networks when compared to data with no demand uncertainty. Also, imperfect sensor measurements were investigated and it was found that an RF model trained with input data based on fuzzy sensors can also achieve a significant reduction in search space for both networks while input data based on Boolean sensors require significantly more data for larger networks in order to achieve a good reduction.

The proposed method must be coupled with an optimization algorithm to identify other contamination parameters – contamination start time, contamination duration and contamination chemical concentration. With the massively reduced search space when only top suspect nodes are considered, independent optimization procedures can be conducted for each of the suspect nodes to ensure that the true source node is detected.

ACKNOWLEDGEMENT

This work was supported by the Center for Advanced Computing and Modelling, University of Rijeka.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

COMPLIANCE WITH ETHICS REQUIREMENTS

This article does not contain any studies with human or animal subjects.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Adedoja, O., Hamam, Y., Khalaf, B. & Sadiku, R. 2018 Towards development of an optimization model to identify contamination source in a water distribution network. *Water* **10**, 579.

Bashi-Azghadi, S. N., Kerachian, R., Bazargan-Lari, M. R. & Solouki, K. 2010 Characterizing an unknown pollution source in groundwater resources systems using PSVM and PNN. *Expert Systems with Applications* **37**, 7154–7161.

Besner, M. C., Prévost, M. & Regli, S. 2011 Assessing the public health risk of microbial intrusion events in distribution systems: conceptual model, available data, and challenges. *Water Research* **45**, 961–979.

Braun, M., Bernard, T., Ung, H., Piller, O. & Gilbert, D. 2015 Computational fluid dynamics modeling of contaminant mixing at junctions for an online security management toolkit in water distribution networks. *Journal of Water Supply: Research and Technology – AQUA* **64**, 504–515.

Breiman, L. 2001 Random forests. *Machine Learning* **45**, 5–32.

CWS, U.o.E. CWS benchmarks. Available from: http://emps.exeter.ac.uk/engineering/research/cws/downloads/benchmarks/ (accessed 6 November 2019).

De Sanctis, A., Boccelli, D., Shang, F. & Uber, J. 2008 Probabilistic approach to characterize contamination sources with imperfect sensors. In *World Environmental and Water Resources Congress 2008*, Ahupua'A, pp. 1–10.

Eliades, D. G. & Polycarpou, M. M. 2011 Water contamination impact evaluation and source-area isolation using decision trees. *Journal of Water Resources Planning and Management* **138**, 562–570.

Eliades, D., Lambrou, T., Panayiotou, C. G. & Polycarpou, M. M. 2014 Contamination event detection in water distribution systems using a model-based approach. *Procedia Engineering* **89**, 1089–1096.

Grbčić, L., Kranjčević, L., Lučin, I. & Čarija, Z. 2019 Experimental and numerical investigation of mixing phenomena in double-Tee junctions. *Water* **11**, 1198.

Grbčić, L., Kranjčević, L., Družeta, S. & Lučin, I. 2020a Efficient double-Tee junction mixing assessment by machine learning. *Water* **12**, 238.

Grbčić, L., Lučin, I., Kranjčević, L. & Družeta, S. 2020b A machine learning-based algorithm for water network contamination source localization. *Sensors* **20**, 2613.

Huang, J. J. & McBean, E. A. 2009 Data mining to identify contaminant event locations in water distribution systems. *Journal of Water Resources Planning and Management* **135**, 466–474.

Kim, M., Choi, C. Y. & Gerba, C. P. 2008 Source tracking of microbial intrusion in water systems using artificial neural networks. *Water Research* **42**, 1308–1314.

Kranjčević, L., Čavrak, M. & Šestan, M. 2010 Contamination source detection in water distribution networks. *Engineering Review* **30**, 11–25.

Lee, Y. J., Park, C. & Lee, M. L. 2018 Identification of a contaminant source location in a river system using random forest models. *Water* **10**, 391.

Liu, L., Zechman, E. M., Mahinthakumar, G. & Ranjithan, S. R. 2012 Coupling of logistic regression analysis and local search methods for characterization of water distribution system contaminant source. *Engineering Applications of Artificial Intelligence* **25**, 309–316.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. 2011 Scikit-learn: machine learning in python. *Journal of Machine Learning Research* **12**, 2825–2830.

Perelman, L. & Ostfeld, A. 2013 Bayesian networks for source intrusion detection. *Journal of Water Resources Planning and Management* **139**, 426–432.

Piazza, S., Blokker, E. M., Freni, G., Puleo, V. & Sambito, M. 2020 Impact of diffusion and dispersion of contaminants in water distribution networks modelling and monitoring. *Water Supply* **20**, 46–58.

Preis, A. & Ostfeld, A. 2007 A contamination source identification model for water distribution system security. *Engineering Optimization* **39**, 941–947.

Preis, A. & Ostfeld, A. 2008a Genetic algorithm for contaminant source characterization using imperfect sensors. *Civil Engineering and Environmental Systems* **25**, 29–39.

Preis, A. & Ostfeld, A. 2008b Multiobjective sensor design for water distribution systems security. In: *Water Distribution Systems Analysis Symposium 2006* (S. G. Buchberger, R. M. Clark, W. M. Grayman & J. G. Uber, eds). American Society of Civil Engineers (ASCE), Cincinnati, OH, pp. 1–17.

Preis, A., Ostfeld, A. & Perelman, L. 2007 Contamination source detection with fuzzy sensors data. In: *World Environmental* *and Water Resources Congress 2007. Restoring Our Natural Habitat* (K. C. Kabbes, ed.). American Society of Civil Engineers (ASCE), Tampa, FL, pp. 1–13.

Rodriguez-Galiano, V., Mendes, M. P., Garcia-Soldado, M. J., Chica-Olmo, M. & Ribeiro, L. 2014 Predictive modeling of groundwater nitrate pollution using random forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (southern Spain). *Science of the Total Environment* **476**, 189–206.

Rossman, L. A. 2000 EPANET 2: Users Manual., U.S. Environmental Protection Agency, Cincinnati, OH.

Rutkowski, T. & Prokopiuk, F. 2018 Identification of the contamination source location in the drinking water distribution system based on the neural network classifier. *IFAC-PapersOnLine* **51**, 15–22.

Shafiee, M. E., Berglund, E. Z. & Lindell, M. K. 2018 An agent-based modeling framework for assessing the public health protection of water advisories. *Water Resources Management* **32**, 2033–2059.

Shen, H. & McBean, E. 2011 False negative/positive issues in contaminant source identification for water-distribution systems. *Journal of Water Resources Planning and Management* **138**, 230–236.

Singh, R. M. & Datta, B. 2007 Artificial neural network modeling for identification of unknown pollution sources in groundwater with partially missing concentration observation data. *Water Resources Management* **21**, 557–572.

Strickling, H., DiCarlo, M. F., Shafiee, M. E. & Berglund, E. 2020 Simulation of containment and wireless emergency alerts within targeted pressure zones for water contamination management. *Sustainable Cities and Society* **52**, 101820.

Ung, H., Piller, O., Gilbert, D. & Mortazavi, I. 2017 Accurate and optimal sensor placement for source identification of water distribution networks. *Journal of Water Resources Planning and Management* **143**, 04017032.

Van Zyl, J. E. 2001 *A Methodology for Improved Operational Optimization of Water Distribution Systems*. Ph.D. Thesis, University of Exeter, UK.

Wang, Q., Xie, Z. & Li, F. 2015 Using ensemble models to identify and apportion heavy metal pollution sources in agricultural soils on a local scale. *Environmental Pollution* **206**, 227–235.

Xuesong, Y., Jie, S. & Chengyu, H. 2017 Research on contaminant sources identification of uncertainty water demand using genetic algorithm. *Cluster Computing* **20**, 1007–1016.

Yan, X., Zhu, Z. & Li, T. 2019 Pollution source localization in an urban water supply network based on dynamic water demand. *Environmental Science and Pollution Research* **26**, 17901–17910.

Zechman, E. M. & Ranjithan, S. R. 2009 Evolutionary computation-based methods for characterizing contaminant sources in a water distribution system. *Journal of Water Resources Planning and Management* **135**, 334–343.

First received 3 March 2020; accepted in revised form 6 August 2020. Available online 15 September 2020