Type of the Paper (Article, Review, Communication, etc.)

On the distributional characterization of graph models of water distribution networks in Wasserstein spaces

Antonio Candelieri [1,*], Andrea Ponti [1,3,4] and Francesco Archetti [2,4]

[1] University of Milano-Bicocca, Department of Economics, Management and Statistics, Italy;
[2] University of Milano-Bicocca, Department of computer Science, Systems and Communication, Italy;
[3] Oaks s.r.l.;
[4] Consorzio Milano Ricerche;
* Correspondence: antonio.candelieri@unimib.it

Abstract: This paper is focused on two topics very relevant in water distribution networks (WDNs): vulnerability assessment and the optimal placement of water quality sensors. The main novelty element of this paper is to represent the data of the problem, in this case all objects in a graph underlying a water distribution network, as discrete probability distributions. For vulnerability (and the related issue of resilience) the metrics from network theory, widely studied and largely adopted in the water research community, reflect connectivity expressed as closeness centrality or, betweenness centrality based on the average values of shortest paths between all pairs of nodes. Also network efficiency and the related vulnerability measures are related to average of inverse distances. In this paper we propose a different approach based on the discrete probability distribution, for each node, of the node-to-node distances. For the optimal sensor placement, the elements to be represented as discrete probability distributions are sub-graphs given by the locations of water quality sensors. The objective functions, detection time and its variance as a proxy of risk, are accordingly represented as a discrete e probability distribution over contamination events. This problem is usually dealt with by EA algorithm. We'll show that a probabilistic distance, specifically the Wasserstein (WST) distance, can naturally allow an effective formulation of genetic operators. Usually, each node is associated to a scalar real number, in the optimal sensor placement considered in the literature, average detection time, but in many applications, node labels are more naturally expressed as histograms or probability distributions: the water demand at each node is naturally seen as a histogram over the 24 hours cycle. The main aim of this paper is twofold: first to show how different problems in WDNs can take advantage of the representational flexibility inherent in WST spaces. Second how this flexibility translates into computational procedures.

Keywords: multi-objective; evolutionary algorithms; Pareto optimality; Wasserstein distance; network vulnerability; resilience; sensor placement.

1. Introduction

In this paper we address 2 problems: the assessment of network vulnerability for which we introduce a new index which naturally captures the

impact of the removal of nodes/edges of the network and the optimal placement of water quality sensors for the early detection of contamination events. The main contribution of this paper is the proposal of a unifying mathematical framework in which the graph underlying the WDN, and its elements are represented as discrete probability distributions. Thanks to this representation the elements of both problems can be seen as points in a space endowed with distance between distributions.

Let's first consider the vulnerability assessment. A first set of measures, widely used in the water research community, is derived from the topology of the network as captured by the analysis of the distance matrix whose entries are the shortest paths between all pair of nodes. The analysis of this matrix yields centrality measures like closeness centrality, betweenness centrality expressed as average values of the node to node distances. Also network efficiency and the related vulnerability measures are related to average of inverse distances. These measures are widely studied and largely adopted in the water research community. Other metrics originate from the analysis of the adjacency matrix and its Laplacian. The spectral analysis of these matrices gives thru the spectral radius and the smallest nonzero eigenvalue respectively another a widely used set of connectivity measures.

This paper stems from the observation that there is a lot more information hidden in these matrices than is captured by the average values used to compute them. In this paper we represent the data of the problem, in this case all objects in the graph underlying a water distribution network, as discrete probability distributions and measure their similarity thru a distributional distance. The distance between probability distributions is a central topic in statistical learning and more generally in Machine learning: there are many distances, as Kullback-Leibler and Jensen-Shannon. In this paper we focus on the Wasserstein distance, and we embed input data as probability distributions in a Wasserstein space. To turn this observation into a tool for network analysis able to capture the difference between nodes and net-works through their distances we need a sound theoretical and computational framework. A main contribution of this paper is to propose one such framework to characterize its mathematical features and the computational solutions and to show its representational flexibility and computational efficiency on two different paradigmatical problems in the analysis of WDNs. The same mathematical structure is generalized to the multi-objective optimal sensor problem. The model set-up assumes that the introduction of a contaminant can be associated to any node of the network. Each contamination event can be considered as a scenario in which the objective functions are evaluated. In order to get a distributional representation, instead of taking the average over scenarios, we consider, for a sensor placement the sample of detection times over all scenarios and the associated discrete probability distributions. The distance between two sensor placements can then be computed as the Wasserstein distance between the associate distributions.

The Wasserstein distance , first introduced in [1], has received its modern formulation in [2]. WST has a very rich mathematical structure whose complexity and flexibility are analyzed in a landmark volume [3]. A difficulty with WST is its computational cost which has hampered its diffusion outside the computer vision community. Recently a number of specialized computational approaches have drastically reduced the computational hurdles [4].

1.1 Related Works

The literature on vulnerability and resilience in WDNs is extremely large. The analysis of water distribution systems using tools from complex networks theory has been introduced in [5] and extended in [6] to weighted and directed network models. Approaches base on graph theory have been the subject of several papers [7,8]. Other approaches have been focused on the general issue of resilience [9–11]. Giustolisi et al. [12] proposes a new statistical distribution of the node degree. Diao [13] proposes a global resilience analysis in order to assess the resilience with respect to pipe failures, excess demand and substance intrusion. Candelieri et al. [14] integrates network analysis and hydraulic simulation [15,16].

A new approach proposed in [17] uses a representation of the network as a discrete probability distribution over the domain of node-to-node distances. The difference in vulnerability between two networks is expressed thru the WST distance between distributions based on the Wasserstein distance. As shown in ANS this provides also a link criticality index [18].

The problem of optimal sensor placement has been widely analyzed at least in the last 2 decades. The seminal contributions are [19,20]. The problem has been largely modelled as a multi-objective optimization problem and addressed by evolutionary algorithms [21,22]. An early contribution for risk-based sensor placement for contaminant detection is [23]. Naserizade et al. [24] introduces a risk-based multi-objective model, which considers quantile analysis in the evaluation of the objective functions for optimal sensor placement. Candelieri et al. [25] proposes optimal sensor placement to minimize detection and its standard deviation. Zhang et al. [26] proposes a new metric in order to identify the relative importance of each sensor in maintaining the detection performance.

The problem of optimal sensor placement has been addressed in [27,28], proposing a Wasserstein based multi-objective evolutionary algorithm for the risk aware optimization of sensor placement.

1.2 Our Contributions

The main novelty in this paper is to represent the data of the problem as discrete probability distributions and measure their similarity thru a distributional distance more general and flexible than the Euclidean distance. In this paper we focus on the Wasserstein distance, and we embed input data as probability distributions in a Wasserstein space. This new theoretical and computational framework is suitable to two problems. For vulnerability assessment this distributional representation is shown to capture effectively the topological in-formation embedded in network related matrices and yield more meaningful metrics than centrality and vulnerability measures based on average values. The computational outcomes show that the Wasserstein distance probabilistic distance measures have a good capacity to measure the dissimilarity between different networks not only at network but also at edge level. In this sense it can be said that they generalize the information given by clustering methods.

In this framework not only, we can compute the difference in vulnerability between any two networks but also the contribution of each component

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2021 doi:10.20944/preprints202112.0506.v1

to the increase of vulnerability induced by the removal of that component. For the other target problem, optimal sensor placement, the contribution of the paper is the proposal of a new multi-objective evolutionary optimization algorithm: the WST distance enables a new genetic operator which synthetizes the distance between two placements. The performance of the resulting MOES/WST algorithm is computationally shown to be significantly superior than the standard NSGA-II. Both in terms of hypervolume and of cover-age of the approximate Pareto set.

1.3 Organization of the paper

Sect. 2 is devoted to the probabilistic representation of the problem's vulnerability and optimal sensor placement. as discrete probability distribution. Sect. 3 contains the basic notions of the Wasserstein distance and a summary of the computational issue connected with its evaluation. Sect. 4 proposes the new vulnerability index based on the Wasserstein distance. Sect. 5 introduces the problem of sensor placement and explains how the "fitness" of a sensor placement can be captured by a histogram. Sect. 6 introduces the archiving structure for the output of the hydraulic simulation. Sect. 7 gives the new selection and cross over operators based on the Wasserstein distance, the structure of the resulting multi-objective evolutionary algorithm and the computational results about optimal sensor placement. Sect. 8 contains the conclusions and perspectives.

2. Probabilistic Representation of Graphs

In general terms a histogram is a function m_k that counts the elements in a sample of n observations of a random variable that fall into each of the disjoint categories (known as bins). Therefore, if k is the number of bins, the histogram m_k satisfies the condition:

$$n = \sum_{k=1}^{K} m_k \tag{1}$$

To construct a histogram, the first step is to "bin" the range of values – that is, divide the support of the random variable into a number of intervals – and then compute the "weight" of the bin counting how many sampled values fall into each interval. The bins are usually specified as adjacent, consecutive, non-overlapping intervals and are usually of the same size.

In this section a new analysis is performed in terms of node–node discrete distance distributions where the weights m_k are $P_i(k)$ the percentage of nodes which are connected to i at a distance k with each node $i = 1, \dots, n$ of the graph $G(V, E)$.

$$P_i(k) = \frac{n_{i,k}}{n-1} \tag{2}$$

The distance distribution over the whole graph is given by the average over all nodes of the node-to-node distance distributions:

$$P_G(k) = \mu_k = \frac{1}{n}\sum_{i=1}^{n} \frac{n_{i,k}}{n-1} = \frac{1}{n}\sum_{i=1}^{n} P_i(k)\,. \tag{3}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2021 doi:10.20944/preprints202112.0506.v1

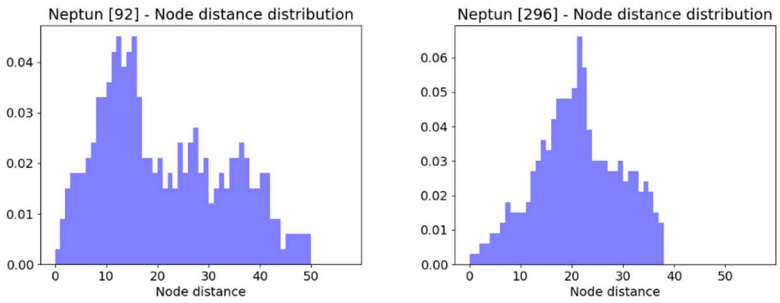

Figure 1. The node-to-node distance distribution for node 92 (left) and node 296 (right) of the Neptun network.

3. The Wasserstein Distance

Measuring the distance between distributions can be accomplished by many alternative models including Kullback-Leibler (and its symmetrized version Jensen-Shannon), Hellinger, total variation and χ-square divergence. In this paper we focus on the Wasserstein (WST) distance whose basic notions are given in Sect. 3.1 while Sect. 3.2 is devoted to the computation of the barycenter between distributions and the extension of k-means clustering to the Wasserstein space. It is important to remark that the presentation is quite basic omitting any mathematical characterization of WST for which the reader is referred to [3,4].

The Wasserstein distance can be traced back to the works of Gaspard Monge [1] and Lev Kantorovich [2] and is based on the solution of an optimal transport problems. WST enables to synthetizes the comparison between two multi-dimensional distributions through a single metric using all information in the distributions.

3.1 Basic Notion

The WST distance between continuous probability distributions is:

$$W_p\left(P^{(1)}, P^{(2)}\right) = \left(\inf_{\gamma \in \Gamma\left(P^{(1)}, P^{(2)}\right)} \int_{X \times X} d\left(x^{(1)}, x^{(2)}\right)^p d\gamma\left(x^{(1)}, x^{(2)}\right)\right)^{\frac{1}{p}} \quad (4)$$

where $d(x^{(1)}, x^{(2)})$ is also called *ground distance* (usually it is the Euclidean norm), $\Gamma\left(P^{(1)}, P^{(2)}\right)$ is the set of joint distributions $\gamma(x^{(1)}, x^{(2)})$ whose marginals are respectively $P^{(1)}$ and $P^{(2)}$, and $p > 1$ is an index.

In some cases, WST can be written in an explicit form. Let $\hat{P}^{(1)}$ and $\hat{P}^{(2)}$ be the cumulative distribution for one-dimensional distributions $P^{(1)}$ and $P^{(2)}$ on the real line and $\left(\hat{P}^{(1)}\right)^{-1}$ and $\left(\hat{P}^{(2)}\right)^{-1}$ be their quantile functions.

$$W_p\left(P^{(1)}, P^{(2)}\right) = \left(\int_0^1 \left|\left(\hat{P}^{(1)}\right)^{-1}\left(x^{(1)}\right) - \left(\hat{P}^{(2)}\right)^{-1}\left(x^{(2)}\right)\right|^p dx\right)^{\frac{1}{p}} \quad (5)$$

3.2 The Wasserstein Distance for Discrete Distributions

Let P denote a discrete distribution with support points x_i with $i = 1, \ldots, m$ and their associated probabilities w_i such that $\sum_{i=1}^m w_i = 1$ with $w_i \geq 0$ and $x_i \in M$ for $i = 1, \ldots, m$. Usually, $M = \mathbb{R}^d$ is the d-dimensional Euclidean space with the l_p norm and x_i are called the support vectors. M

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2021 doi:10.20944/preprints202112.0506.v1

can also be a symbolic set provided with a symbol-to-symbol similarity. P can also be written using the notation:

$$P(x) = \sum_{i=1}^{m} w_i \delta(x - x_i) \tag{6}$$

where $\delta(\cdot)$ is the Kronecker delta.

The WST distance between two distributions $P^{(1)} = \left\{w_i^{(1)}, x_i^{(1)}\right\}$ with $i = 1, \dots, m_1$ and $P^{(2)} = \left\{w_i^{(2)}, x_i^{(2)}\right\}$ with $i = 1, \dots, m_2$ is obtained by solving the following linear program:

$$W\left(P^{(1)}, P^{(2)}\right) = \min_{\gamma_{ij} \in \mathbb{R}+} \sum_{i \in I_1, j \in I_2} \gamma_{ij}\, d\left(x_i^{(1)}, x_j^{(2)}\right) \tag{7}$$

The cost of transport between $x_i^{(1)}$ and $x_j^{(2)}$, $d\left(x_i^{(1)}, x_j^{(2)}\right)$, is defined by the p-th power of the norm $\left\|x_i^{(1)}, x_j^{(2)}\right\|$ (usually the Euclidean distance). We define two index sets $I_1 = \{1, \dots, m_1\}$ and I_2 likewise, such that:

$$\sum_{i \in I_1} \gamma_{ij} = w_j^{(2)}, \forall j \in I_2 \tag{8}$$

$$\sum_{j \in I_2} \gamma_{ij} = w_i^{(1)}, \forall i \in I_1 \tag{9}$$

Equations 8 and 9 represent the in-flow and out-flow constraint, respectively. The terms γ_{ij} are called matching weights between support points $x_i^{(1)}$ and $x_j^{(2)}$ or the optimal coupling for $P^{(1)}$ and $P^{(2)}$. The discrete version of the WST distance is usually called Earth Mover Distance (EMD). For instance, when measuring the distance between grey scale images, the histogram weights are given by the pixel values and the coordinates by the pixel positions. Another way to look at the computation of the EMD is as a network flow problem. In the specific case of histograms, the entries γ_{ij} denote how much of the bin i has to be moved to bin j.

In the case of one-dimensional histograms, the computation of WST can be performed by a simple sorting and the application of the following equation.

$$W_p\left(P^{(1)}, P^{(2)}\right) = \left(\frac{1}{n} \sum_{i}^{n} \left|x_i^{(1)*} - x_i^{(2)*}\right|^p\right)^{\frac{1}{p}} \tag{10}$$

where $x_i^{(1)*}$ and $x_i^{(2)*}$ are the sorted samples.

3.3 A Toy Example

Two graphs G and G' are considered with their distributions $P_G(k)$ and $P_{G'}(k)$ that will be referred to as p and p'. In order to give an instance of the computation of the node-to-node distributions, a small synthetic water distribution network, Anytown (Figure 2), is considered in [29]. The associated graph G consists of 25 nodes and 44 edges. G' is the graph without the red edge.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2021 doi:10.20944/preprints202112.0506.v1

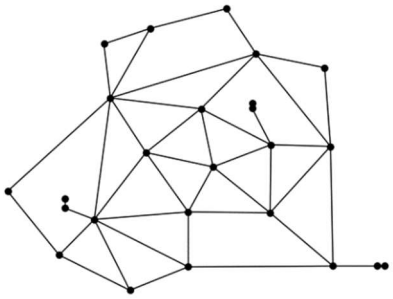

Figure 2. A schematich representation of Anytown WDN.

$$P_G = [0.147, 0.263, 0.297, 0.177, 0.083, 0.030, 0.003, 0]$$
$$P_{G'} = [0.133, 0.237, 0.290, 0.183, 0.100, 0.043, 0.010, 0.003]$$

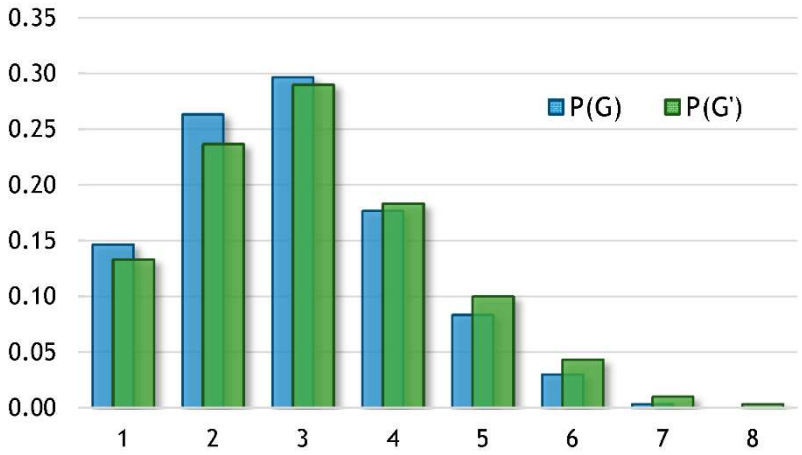

Figure 3. Node-to-node distance distributions.

The support of $P_G(k)$ and $P_{G'}(k)$ are respectively the integers $k = 1, \dots, D(G)$ (analougsly for G'). When G' is derived from G removing some edges then $D(G') \geq D(G)$. Since the distributions are represented by histograms (Figure 2) one can extend to G the support of G' setting $\mu_G(k) = 0$ for $k = D(G) + 1, \dots, D(G')$. In this toy examples Equation (7) translate into:

$$\delta_k = \delta_{k-1} + P_G(k) - P_{G'}(k) \; k = 1, \dots, \max\big(D(G), D(G')\big) \quad (11)$$

$$W(G, G') = \sum |\delta_i| = 0.1767 \quad (12)$$

4. Data and software resources

Anytown is a small synthetic water distribution network [29] whose associated graph has 25 nodes and 44 edges. Neptun is the WDN of the Romanian city of Timisoara, with an associated graph of 333 nodes and 339 edges [30].

The multi-objective evolutionary algorithm used for the optimal sensor placement is based on the Python framework Pymoo [31].

The Water Network Tool for Resilience (WNTR) [32] is a Python package based on EPANET 2.0.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2021 doi:10.20944/preprints202112.0506.v1

5. Resilience

In this section we show how the Wasserstein distance enables a new index of vulnerability which can be applied both network wise and as a criticality index of each network component.

5.1 Wasserstein distance as a measure of network vulnerability

This remarkable result is displayed in Figures 4 and 5 for the Neptun network. Given the graph $G = (V, E)$ associated to the network, each edge is represented by a pair of adjacent nodes (i, j). The removal of (i, j) yields $G\backslash\{(i, j)\}$ for which we compute the aggregate node–node distance distribution $p(G\backslash\{(i, j)\})$ and the Wasserstein distance $W(p(G), p(G\backslash\{(i, j)\})$, whose value is represented by the the color associated to each edge (i, j) by the Wasserstein distance. Usual measures of network vulnerability are given by the loss of efficiency generated removing one or more network elements (nodes and graphs). The distributional approach taken here is different: we consider the original network and that obtained removing components, and we consider the WST distance between the two network as an index of vulnerability.

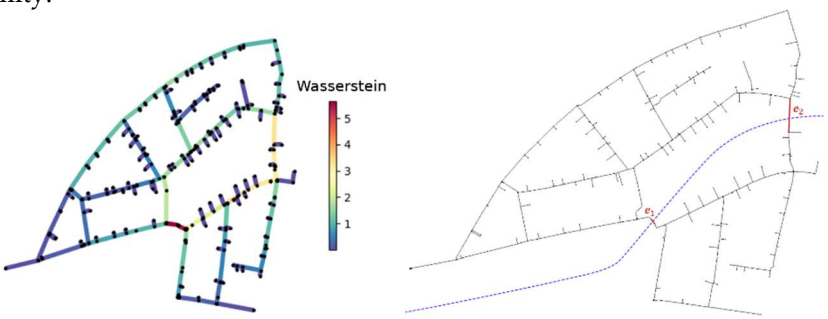

Figure 4. The heat map representation of the Wasserstein based vulnerability index for Neptun (left). Critical edges (red) highlight a cut-set in Neptun consisting of 2 edges (bridges) whose simultaneous removal generates a disconnection (right).

It is important to note that breakages must occur, at the same time, on all the different red edges to imply a hydraulic disconnection. Breakages affecting only one pipe may imply a reduction in the efficiency of the network and an increase in vulnerability. It is important to remark that the edge Wasserstein criticality index agrees, with the result of the spectral clustering in Figure 4 (right).

The computational results show that probabilistic distance measures have a good capacity to discriminate between different networks not only at a network level but also for different edges. This assessment of link criticality could assess important tasks WDN management by just using topological and geometric information. The Wasserstein metric can be regarded as a natural extension of the Euclidean distance to statistical distributions via a single metric while still exploiting all the information present in the distributions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2021 doi:10.20944/preprints202112.0506.v1

6. Sensor Placement

6.1 Problem Formulation

We consider a graph $G = (V, E)$ We assume a set of possible locations for placing p sensors, that is $L \subseteq V$. Thus, a sensor placement (SP) is a subset of sensor locations, with the subset's size less or equal to p depending on the available budget. An SP is represented by a binary vector $s \in \{0,1\}^{|L|}$ whose components are $s_i = 1$ if a sensor is located at node i, $s_i = 0$ otherwise. Thus, an SP is given by the nonzero components of s. For a WDN the vertices in V represent junctions, tanks, reservoirs or consumption points, and edges in E represent pipes, pumps, and valves. Let $A \subseteq V$ denote the set of contamination events $a \in A$ which must be detected by a sensor placement s, and d_{ai} is the detection time of a sensor placed in node I for a contamination event in node a. A probability distribution is placed over possible contamination events associated to the nodes. In the computations we assume – as usual in the literature – a uniform distribution, but in general discrete distributions are also possible. In this paper we consider as objective functions the detection time and its standard deviation. We consider a general model of sensor placement:

$$P = \begin{cases} \min f_1(s) = \sum_{a \in A} \alpha_a \sum_{i=1,..,|L|} d_{ai} x_{ai} \\ s.t. \\ \sum_{i=1,..,|L|} s_i \leq p \\ s_i \in \{0,1\} \end{cases} \quad (13)$$

- α_a is the probability for the contaminant to enter the network at node a.
- $x_{ai} = 1$ if $s_i = 1$, where i is the first sensor detecting the contaminant injected at node a; 0 otherwise.

In our study we assume that $\alpha_a = 1/|A|$, therefore $f_1(s)$ is:

$$f_1(s) = \frac{1}{|A|} \sum_{a \in A} \hat{t}_a \quad (14)$$

where $\hat{t}_a = \sum_{i=1,..,|L|} d_{ai} x_{ai}$ is the MDT of event a. f_1 is the average over the contamination events of the detection time for each event. For each event a and sensor placement s the Minimum Detection Time is defined as $MDT_a = \min_{i:\, s_i=1} d_{ai}$ with $\hat{t}_a$ the minimum time step at which concentration reaches or exceeds a given threshold τ for the event a. As a measure of risk, we consider f_2 as the standard deviation of the sample average approximation of f_1.

$$f_2(s) = STD_{f_1}(s) = \sqrt{\frac{1}{|A|} \sum_{a \in A} \left(\hat{t}_a - f_1(s)\right)^2} \quad (15)$$

6.2 Representation of Sensor Placements as Histograms

If bins are the time subintervals of the simulation horizon and the weights are the number of events detected in each time subinterval (or their relative frequency) the histogram obtained is a representation of the information acquired in the hydraulic simulation. Denote the time steps in the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2021 doi:10.20944/preprints202112.0506.v1

simulation $\Delta t_i = t_i - t_{i-1}$ where $i = 1, \dots, k$ are equidistanced in the simulation time horizon $(0, T_{MAX} = 86400)$. $T_{MAX} = k\Delta t,\ \Delta t = 1,\ k = 24$. We consider the discrete random variable $|A_i|$ where $A_i = \{a \in A: \hat{t}_a \in \Delta t_i\}.n = |A|$ cardinality is the number of contamination events, the bins are Δt_i and $m_i = |A_i|$. To each sensor placement s we can associate not only the placement matrix $H^{(s)}$ but also the histogram $h^{(s)}$ whose bins are Δt_i and weights are $|A_i|$.

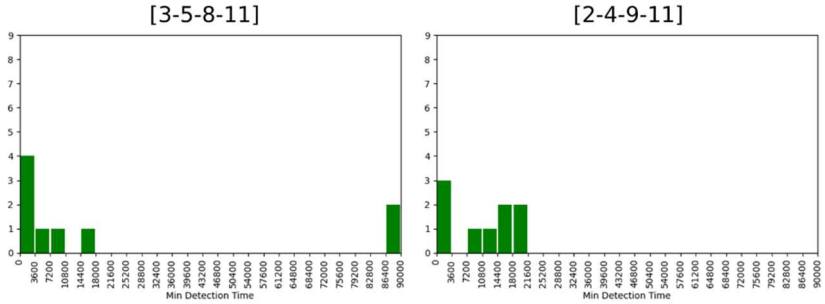

Figure 5. Examples of histograms associated to two different sensor placements considering Net1 WDN.

We have added an "extra" bin (86400 to 90000) whose weight $|A_{k+1}|$ represents number of contamination events which were undetected during the simulation up to 86400. In this way $\sum_{i=1}^{k+1}|A_i| = 1$. The "ideal" placement is that in which $|A_1| = |A|$.Intuitively a "good" sensor placement has a relatively large mass in lower Δt_i), the larger the probability mass in the higher Δt the worse is sensor placement (Figure 5 left). The worst SPs are those in the extra bin.

7. Hydraulic Simulation and Data Structure

7.1 WNTR

The Water Network Tool for Resilience (WNTR) [32] is a Python package based on EPANET 2.0. The simulation is computationally costly as we need one execution for each contamination event.

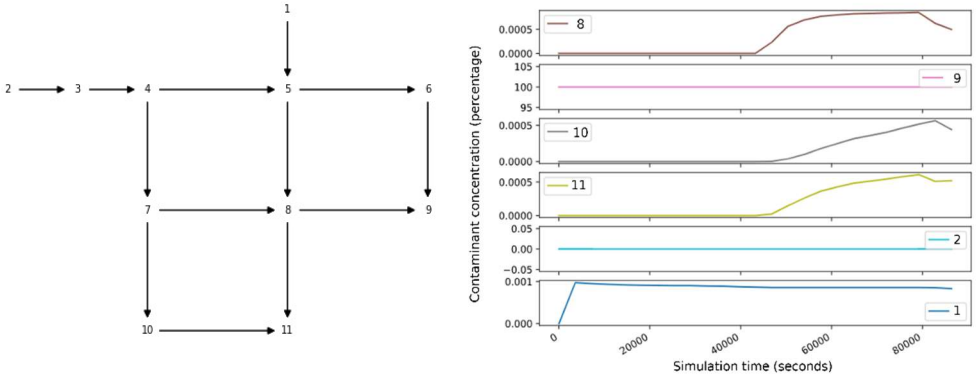

Figure 6. A schematic representation of the Net1 synthetic example (left). Concentrations of the contaminant introduced at node 9 for node 1, 2, 8, 9, 10, 11 (right).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2021 doi:10.20944/preprints202112.0506.v1

The hydraulic simulation has been performed for 24 hours, in steps of 1 hour. Assuming $L = V$ and $A = V$ (i.e., the most computationally demanding problem configuration) the time required to run a simulation for the synthetic example called Net1 (available with EPANET and WNTR, and whose associated graph is depicted in Figure 1) is 2 seconds. The detection threshold is 10% of the initial concentration.

7.2 Single Sensor Placement and Placement Matrix

Denote with $S^{\ell} \in \mathbb{R}^{(K+1)\times|A|}$ the so-called "sensor matrix", with $\ell = 1, \dots, |L|$ an index identifying the location where the sensor is deployed at. Each entry of $S^{(\ell)}$, z_{ta}^{ℓ} represents the concentration of the contaminant for the event $a \in A$ at the simulation step $t = 0, .., K$, with $T_{max} = K\Delta t$. Thus, in our study $T_{max} = 24$, $\Delta t = 1$ and $K = 24$. Without loss of generality, we assume that the contaminant is injected at the beginning of the simulation (i.e., $t = 0$).

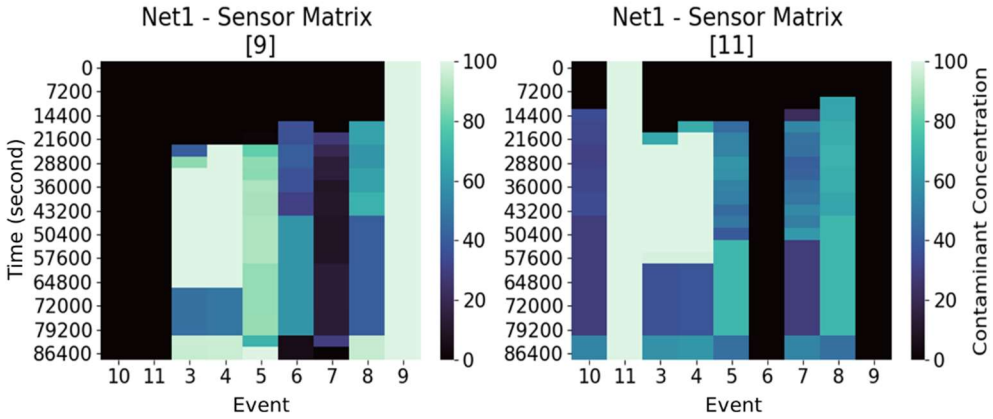

Figure 7. Sensor matrices $S^{(9)}$ and $S^{(11)}$ for sensors deployed respectively at locations (i.e., nodes) 9 (left) and node 11 (right).

A "sensor placement matrix", $H^{(s)} \in \mathbb{R}^{(K+1)\times|A|}$ can be also defined whose entry h_{ta} is the maximum concentration over those detected by the sensors in s, for the event a and at time step t. Suppose to have a sensor placement s consisting of m sensors with associated sensor matrices $S^1, \dots, S^m$, then $H^{(s)}$ is the matrix with entries $h_{ta} = \max_{j=1,\dots,m} z_{ta}^{j}$ $\forall a \in A$. The columns of $H^{(s)}$ having maximum concentration at row $t = 0$ (i.e., injection time) are those associated to events with injection occurring at the deployment locations of the sensors in s. Indeed, we can now explicit the computation of $\hat{t}_a$ in $f_1(s)$ and $f_2(s)$: $\hat{t}_a$ is the minimum time step at which concentration reaches or exceeds a given threshold τ for the scenario a, that is $\hat{t}_a = \min_{t=1,\dots,K} \{h_{ta} \geq \tau\}$. Each bin of the histogram $h^{(s)}$ represents the number of events that are detected in a specific time range by s. These values can be extracted from the placement matrix $H^{(s)}$ Indeed, each column of this matrix represents an event, and the detection time of this event is given by the row in which the contaminant concentration exceed a given threshold.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2021 doi:10.20944/preprints202112.0506.v1

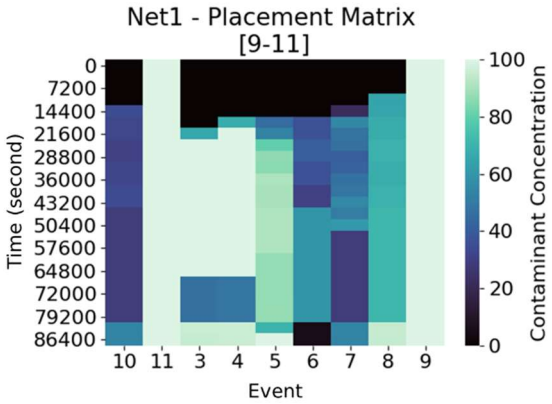

Figure 8. Sensor placement matrix for a sensor placement consisting of two sensors deployed at locations 9 and 11.

A metric in the information space has been introduced in Sect. 3 using histograms and the Wasserstein distance.

8. The Algorithm MOEA/WST

MOEA/WST is instantiated in the above framework according to the following steps:

1. The individuals of the population, that are sensor placements, in the evolutionary algorithm are represented as discrete probability distributions, namely histograms.
2. The space of histograms is endowed with a metric given by the WST distance between them.
3. The results of WST based computations are mapped back into the search space.

This section is focused on analyzing how the mathematics presented in the previous sections enables the construction of two new genetic operators.

8.1 The Selection Operator

For the crossover operation, a problem specific selection method has been developed. First, we randomly sample from the actual Pareto set two pairs of individuals (F_1, M_1) and (F_2, M_2). Then we choose the pair (F_i, M_i) as the parents of the new offspring, where $i = \arg\max_{i\in\{1,2\}} D(F_i, M_i)$. This favors exploration and diversification. In this paper we used for $D(F_i, M_i)$ the Wasserstein distance between the histograms corresponding to the sensor placement F_i and M_i.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2021 doi:10.20944/preprints202112.0506.v1

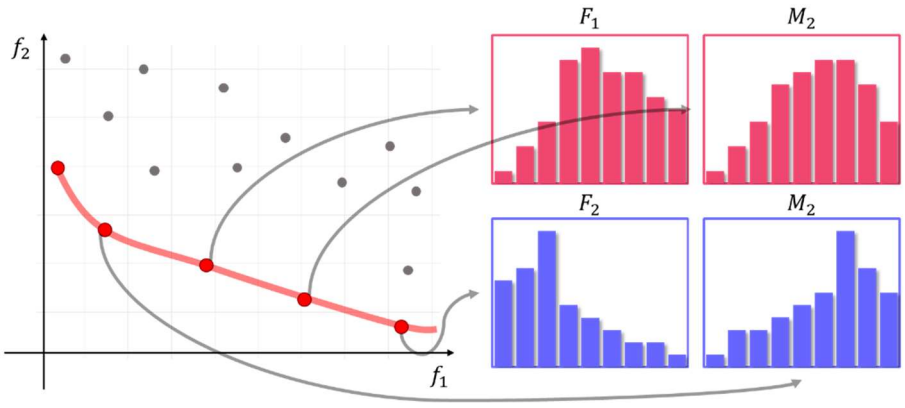

Figure 9. Visualization of the selection mechanism.

If at least one individual of the pair of parents is not feasible (i.e., the placement contains more sensors than the budget p) the Constraint Violation (CV) is considered instead. Let's $c = [c_i]$ be a generic individual and p the budget, the Constraint Violation is defined as follow

$$CV(c) = \max\left(0, \sum_i c - p\right) \tag{17}$$

Then we choose the pair of parents (F_i, M_i) with $i = \arg\min_{i\in\{1,2\}}\big(CV(F_i) + CV(M_i)\big)$.

8.2 The Crossover Operator

Denote with $x, x' \in \Omega$ two feasible parents and with J (FatherPool) and J' (MotherPool) the two associated sets $J = \{i: x_i = 1\}$ and $J' = \{i: x'_i = 1\}$. To obtain two feasible children, c and c' are initialized as $[0, \dots, 0]$. In turn, c and c' samples an index from J and from J', respectively, without replacement. Therefore, the new operator rules out children with more than p non-zero components.

In Figure 10, an example comparing the behaviour of our crossover compared to a typical 1-point crossover.

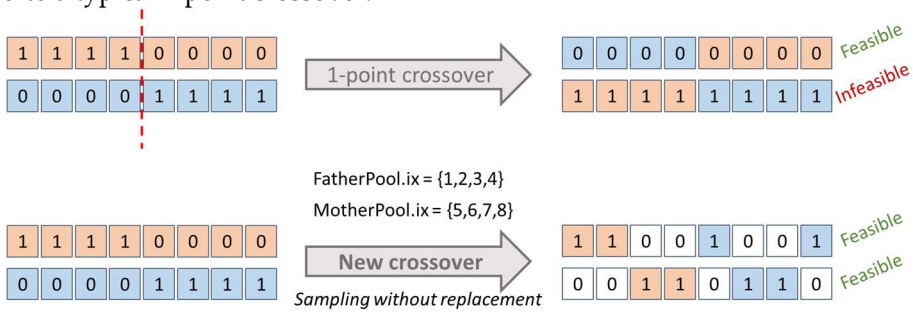

Figure 10. A schematic representation of the new crossover operator.

This problem specific crossover generates two "feasible-by-design" children from two feasible parents.

8.3 Computational Results

Figure 11 displays the comparative computational results on the Netpun network of MOEA/WST and the benchmark NSGA-II. It is apparent that MOEA/WST reaches a large value of the dominated hypervolume and therefore a much better approximation of the Pareto set way before NSGA-II.

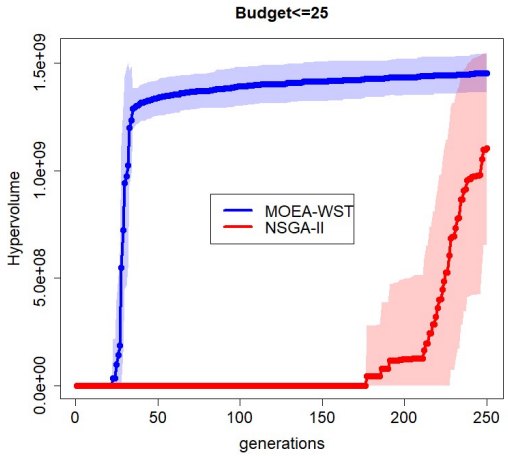

Figure 11. Neptun computational results.

9. Conclusion and Perspectives

The key objective of this work is to show that a distributional framework is suitable to solve two key problems in the design of WDNs: resilience assessment and optimal sensor placement. In particular the Wasserstein distance and the associated analytical methods offer significant advantages over comparing distributions using a set of parametric values such as mean, variance and higher order moments. Indeed, the analysis of these parameters only does not take the whole distribution into account. Even if the roots of WST are in abstract spaces of probability distributions, WST and the associated optimal transport map offer a visually intuitive representation of the similarity between distributions.

The approach proposed is extremely flexible and can be generalized to different failure modes (pipe failures, excess demand and substance intrusion) and different probability functions to account for spatially varied contamination probabilities. A novel data structure has been also proposed for archiving the results of the simulation. This distributional approach has been shown to enable more effective genetic operators, able to capture the specific structure of the optimal sensor placement problem and give a much better computational performance than standard evolutionary algorithms. This distributional approach could be extended to other problems in the design and management of WDNs in the wider context of multitask learning (Ponti, A., (2021 d). This new approach can be naturally extended to different domains of networked infrastructure, as transport and energy and also informational networks for problems as fake-news detection and collaborative filtering in recommender system.

Author Contributions: All authors contributed equally to the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This study has been partially supported by the Italian project "PERFORM-WATER 2030"—programma POR (Programma Operativo Regionale) FESR (Fondo Europeo di Sviluppo Regionale) 2014–2020, innovation call "Accordi per la Ricerca e l'Innovazione" ("Agreements for Research and Innovation") of Regione Lombardia, (DGR N. 5245/2016—AZIONE I.1.B.1.3—ASSE I POR FESR 2014–2020)—CUP E46D17000120009. This study has also been partially supported by the Italian project ENERGIDRICA co-financed by MIUR.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data related to the network Anytown are available in the literature. The data of Neptun and Abbiategrasso are available from the authors on demand.

Acknowledgments: The contribution of DEMS Data Science Lab for supporting this work by providing computational resources (DEMS—Department of Economics, Management and Statistics) is gratefully acknowledged.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Monge, G. Mémoire Sur La Théorie Des Déblais et Des Remblais. *Histoire de l'Académie Royale des Sciences de Paris* **1781**.
2. Kantorovitch, L. On the Translocation of Masses. *Management Science* **1958**, *5*, 1–4, doi:10.1287/mnsc.5.1.1.
3. Villani, C. *Optimal Transport: Old and New*; Springer Science & Business Media, 2008; Vol. 338;.
4. Peyré, G.; Cuturi, M. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning* **2019**, *11*, 355–607.
5. Yazdani, A.; Jeffrey, P. Complex Network Analysis of Water Distribution Systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **2011**, *21*, 016111.
6. Yazdani, A.; Jeffrey, P. Water Distribution System Vulnerability Analysis Using Weighted and Directed Network Models. *Water Resources Research* **2012**, *48*.
7. Herrera, M.; Abraham, E.; Stoianov, I. A Graph-Theoretic Framework for Assessing the Resilience of Sectorised Water Distribution Networks. *Water Resources Management* **2016**, *30*, 1685–1699.
8. Di Nardo, A.; Giudicianni, C.; Greco, R.; Herrera, M.; Santonastaso, G.F. Applications of Graph Spectral Techniques to Water Distribution Network Management. *Water* **2018**, *10*, 45.
9. Jung, D.; Kim, J.H. *Emerging Issues and Methodologies for Resilient and Robust Water Distribution Systems*; Multidisciplinary Digital Publishing Institute, 2020;
10. Assad, A.; Moselhi, O.; Zayed, T. A New Metric for Assessing Resilience of Water Distribution Networks. *Water* **2019**, *11*, 1701.
11. Diao, K.; Sweetapple, C.; Farmani, R.; Fu, G.; Ward, S.; Butler, D. Global Resilience Analysis of Water Distribution Systems. *Water research* **2016**, *106*, 383–393.
12. Giustolisi, O.; Simone, A.; Ridolfi, L. Network Structure Classification and Features of Water Distribution Systems. *Water Resources Research* **2017**, *53*, 3407–3423.
13. Diao, K. Multiscale Resilience in Water Distribution and Drainage Systems. *Water* **2020**, *12*, 1521.
14. Candelieri, A.; Giordani, I.; Archetti, F. Supporting Resilience Management of Water Distribution Networks through Network Analysis and Hydraulic Simulation. In Proceedings of the 2017 21st International Conference on Control Systems and Computer Science (CSCS); IEEE, 2017; pp. 599–605.
15. Soldi, D.; Candelieri, A.; Archetti, F. Resilience and Vulnerability in Urban Water Distribution Networks through Network Theory and Hydraulic Simulation. *Procedia Engineering* **2015**, *119*, 1259–1268.

16. Ulusoy, A.-J.; Stoianov, I.; Chazerain, A. Hydraulically Informed Graph Theoretic Measure of Link Criticality for the Resilience Analysis of Water Distribution Networks. *Applied network science* **2018**, *3*, 1–22.
17. Ponti, A.; Candelieri, A.; Giordani, I.; Archetti, F. A Novel Graph-Based Vulnerability Metric in Urban Network Infrastructures: The Case of Water Distribution Networks. *Water* **2021**, *13*, 1502.
18. Ponti, A.; Candelieri, A.; Giordani, I.; Archetti, F. Probabilistic Measures of Edge Criticality in Graphs: A Study in Water Distribution Networks. *Applied Network Science* **2021**, *6*, 1–17.
19. Ostfeld, A.; Uber, J.G.; Salomons, E.; Berry, J.W.; Hart, W.E.; Phillips, C.A.; Watson, J.-P.; Dorini, G.; Jonkergouw, P.; Kapelan, Z. The Battle of the Water Sensor Networks (BWSN): A Design Challenge for Engineers and Algorithms. *Journal of Water Resources Planning and Management* **2008**, *134*, 556–568.
20. Krause, A.; Leskovec, J.; Guestrin, C.; VanBriesen, J.; Faloutsos, C. Efficient Sensor Placement Optimization for Securing Large Water Distribution Networks. *Journal of Water Resources Planning and Management* **2008**, *134*, 516–526.
21. Margarida, D.; Antunes, C.H. Multi-Objective Optimization of Sensor Placement to Detect Contamination in Water Distribution Networks. In Proceedings of the Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation; 2015; pp. 1423–1424.
22. Dorini, G.; Jonkergouw, P.; Kapelan, Z.; Di Pierro, F.; Khu, S.T.; Savic, D. An Efficient Algorithm for Sensor Placement in Water Distribution Systems. In Proceedings of the Water Distribution Systems Analysis Symposium 2006; 2008; pp. 1–13.
23. Weickgenannt, M.; Kapelan, Z.; Blokker, M.; Savic, D.A. Risk-Based Sensor Placement for Contaminant Detection in Water Distribution Systems. *Journal of Water Resources Planning and Management* **2010**, *136*, 629–636.
24. Naserizade, S.S.; Nikoo, M.R.; Montaseri, H. A Risk-Based Multi-Objective Model for Optimal Placement of Sensors in Water Distribution System. *Journal of Hydrology* **2018**, *557*, 147–159.
25. Candelieri, A.; Ponti, A.; Archetti, F. Risk Aware Optimization of Water Sensor Placement. *arXiv preprint arXiv:2103.04862* **2021**.
26. Zhang, Q.; Zheng, F.; Kapelan, Z.; Savic, D.; He, G.; Ma, Y. Assessing the Global Resilience of Water Quality Sensor Placement Strategies within Water Distribution Systems. *Water research* **2020**, *172*, 115527.
27. Ponti, A.; Candelieri, A.; Archetti, F. A New Evolutionary Approach to Optimal Sensor Placement in Water Distribution Networks. *Water* **2021**, *13*, 1625.
28. Ponti, A.; Candelieri, A.; Archetti, F. A Wasserstein Distance Based Multiobjective Evolutionary Algorithm for the Risk Aware Optimization of Sensor Placement. *Intelligent Systems with Applications* **2021**, *10*, 200047.
29. Farmani, R.; Walters, G.A.; Savic, D.A. Trade-off between Total Cost and Reliability for Anytown Water Distribution Network. *Journal of water resources planning and management* **2005**, *131*, 161–171.
30. Fantozzi, M.; Popescu, I.; Farnham, T.; Archetti, F.; Mogre, P.; Tsouchnika, E.; Chiesa, C.; Tsertou, A.; Gama, M.C.; Bimpas, M. ICT for Efficient Water Resources Management: The ICeWater Energy Management and Control Approach. *Procedia Engineering* **2014**, *70*, 633–640.
31. Blank, J.; Deb, K. Pymoo: Multi-Objective Optimization in Python. *IEEE Access* **2020**, *8*, 89497–89509.
32. Klise, K.A.; Murray, R.; Haxton, T. An Overview of the Water Network Tool for Resilience (WNTR). **2018**.