Redundant binary codes in genetic algorithms: multi-objective design optimization of water distribution networks

Tiku T. Tanyimboh

ABSTRACT

Genetic algorithms have been shown to be highly effective for optimization problems in various disciplines, and binary coding is generally adopted as it is straightforward to implement and lends itself to problems with discrete-valued decision variables. However, a difficulty associated with binary coding is the existence of redundant codes that do not correspond to any element in the finite discrete set that the encoded parameter belongs to. A common technique used to address redundant binary codes is to discard the chromosomes in which they occur. Effective alternatives to the outright removal of redundant codes are lacking in the literature. This article presents illustrative examples based on the problem of optimizing the design of water distribution networks. Two benchmark networks in the literature and two different multi-objective design optimization models were considered. Different fixed mapping schemes gave significantly different solutions in the search space. The main inference from the results is that mapping schemes that improved diversity in the population of solutions achieved better results, which may pave the way for the development of practical and effective mapping schemes.

Key words | genetic algorithm, infrastructure resilience, redundant binary codes, statistical entropy, uncertainty-based design, water distribution network

Tiku T. Tanyimboh
School of Civil and Environmental Engineering,
University of the Witwatersrand,
Private Bag 3, Johannesburg, WITS 2050,
South Africa
E-mail: *tiku.tanyimboh@wits.ac.za*

HIGHLIGHTS

- Redundant binary codes occur routinely in GAs.
- Redundant binary codes occur if the number of options available is not a power of 2.
- Even mapping of redundant codes promotes search diversity and efficiency.
- Utility of technique for handling redundant codes is demonstrated.

INTRODUCTION

Genetic algorithms (GAs) are used widely to solve optimization problems in various disciplines including civil engineering (Yazdi *et al.* 2017; Abdy Sayyed *et al.* 2019; Fadaee *et al.* 2020; Feng *et al.* 2020; Lotfian *et al.* 2020; Naseri *et al.* 2020; Rathi *et al.* 2020; Xu *et al.* 2020). Powerful and robust, GAs are highly effective for optimization problems with discontinuous, complex and poorly understood solution spaces. In addition to the space of the objective functions, GAs may operate on both the phenotype and genotype spaces, i.e. the solution and coding spaces, respectively. Coding and decoding translate solutions between the genotype and phenotype spaces. Binary coding is used frequently as it is straightforward to implement and lends itself to problems with discrete-valued decision

doi: 10.2166/ws.2020.329

variables. It is used widely due to its relative simplicity. It also lends itself to theoretical analyses.

While binary coding is common, a challenge associated with it is the existence of redundant codes that arise frequently. Redundant codes do not correspond to any element in the finite discrete set to which the encoded parameter belongs. They arise if the encoded parameter belongs to a finite discrete set whose cardinal number is not a power of 2. Approaches for handling redundant codes (Saleh & Tanyimboh 2014) include discarding or assigning relatively low fitness values to the chromosomes in which the redundant codes occur. This inevitably entails the premature loss of valuable information and useful genes (Saleh & Tanyimboh 2014) and, consequently, a reduction in the effectiveness of the GA.

Alternatively, the redundant codes may be mapped to valid codes, i.e. real elements in the phenotype space, in a manner that is fixed, random or probabilistic. Fixed mapping essentially assigns the redundant codes to selected phenotype elements *a priori*. Self-evidently, random mapping assigns the redundant codes to the valid phenotype elements randomly. Random mapping is bias free but the quality of the information that is transmitted from one generation to the next deteriorates, which affects the algorithm's performance adversely.

Also, if the number of decision variables is large, real coding (Pattanaik *et al.* 2020) may be used instead as binary coding results in long chromosomes. In real coding, genes are represented by real numbers (Gen *et al.* 2008). Real coding is suitable for problems with continuous and complex solution spaces. Comparisons between real and binary coded GAs indicate that real coding is more powerful on constrained optimization problems with respect to computational efficiency (Eshelman & Schaffer 1993). In reality, the choice between real and binary coding depends on a number of factors as both perform well in different computational environments. For example, Phan *et al.* (2013) observed that real-coded GAs (RCGAs) required large population sizes to achieve good results consistently. Also, RCGAs require recombination and mutation operators that yield genes belonging to the intervals in which the variables they represent lie (Deb 2001: 106–120).

Traditionally most optimization models were binary coded and continuous decision variables were discretised frequently (Vamvakeridou-Lyroudia *et al.* 2005; Siew *et al.* 2016). The challenges associated with the handling of redundant binary codes have been neglected in the literature and methods that are computationally efficient and practical are lacking (Czajkowska 2016). This article investigates redundant binary codes based on the optimal design of water distribution networks (Haghighi *et al.* 2011; Creaco *et al.* 2015; Mohammadi-Aghdam *et al.* 2015; Jabbary *et al.* 2016; Avila-Melgar *et al.* 2017; Ciaponi *et al.* 2017; Surco *et al.* 2018; Shende & Chau 2019; Poojitha *et al.* 2020). Results from two different constrained multi-objective optimization models and two benchmark networks in the literature are included and assessed.

DESIGN OPTIMIZATION MODELS

Two different efficient design optimization models that were developed recently were considered. The first model addresses the resilience of a water distribution network through statistical entropy, an information-theoretic measure of uncertainty (Shannon 1948). The second is a least-cost design problem based on the construction cost.

Case A: uncertainty based optimization of initial construction cost

Design optimization problem formulation

The objectives considered in the design optimization model comprised the minimization of the initial construction cost and maximization of the statistical entropy (Jaynes 1957) of the pipe flow rates in the network. The decision variables of the design optimization problem were the pipe diameters. The discrete pipe diameter options considered were the commercially available pipe sizes.

The constraints comprised the minimum residual pressure constraints at the demand nodes and conservation of mass and energy. The nodal mass balance and energy conservation equations were satisfied by hydraulic simulation using EPANET 2 (Rossman 2000). The minimum nodal pressure constraints were incorporated as a third objective. In other words, the pressure deficit at the critical demand node was minimised by being reduced to zero. The critical demand node is the demand node that has the largest shortfall in the required head. As the nodal demands vary spatially and temporally, the location of the critical node is dynamic. In general, it cannot be predicted accurately before completing the design, followed by extensive simulation and verification, as its location depends additionally on the pipe diameters.

The statistical flow entropy (Tanyimboh 2017) is an extension of Shannon's informational entropy measure (Shannon 1948) and is essentially a function of the pipe flow rates whose formulation was inspired by the conditional entropy formula (Khinchin 1953, 1957). Its inclusion in the design optimization model addresses the uncertainty that is inherent in the water system's design and operation.

Given a set of nodal demands, and assuming steady-state conditions, the statistical flow entropy function (Tanyimboh 2017) is

$$S = S_0 + \sum_{n=1}^{NN} p_n S_n \tag{1}$$

where S is the flow entropy of the network; S_0 is the entropy due to any differences between the flows from the supply nodes; S_n is the entropy of the flows at node n; $p_n = T_n/T$ is the fraction of the total flow through the network that reaches node n; T_n is the total flow that reaches node n; T is the sum of the demands; NN is the number of nodes.

The differences in the flows from the supply nodes are accounted for as

$$S_0 = -\sum_{n \in I} \frac{Qn_{0n}}{T} \ln\left(\frac{Qn_{0n}}{T}\right) \tag{2}$$

where Qn_{0n} is the flow from supply node n; and I represents the set of supply nodes. Similarly, the entropy of the flows at node n is

$$S_n = -\frac{Qn_{n0}}{T_n} \ln\left(\frac{Qn_{n0}}{T_n}\right) - \sum_{ij \in ND_n} \frac{Qp_{ij}}{T_n} \ln\left(\frac{Qp_{ij}}{T_n}\right); \quad n = 1, \ldots, NN \tag{3}$$

where Qn_{n0} is the demand at node n; Qp_{ij} is the flow rate in pipe ij with nodes i and j as the upstream and downstream nodes, respectively; the set ND_n represents the pipe flows from node n; and T_n is the total flow that reaches node n.

The optimization problem may be summarized briefly as in Equations (4a) to (8a):

$$\text{Minimize the initial construction cost: } f_1 = \sum_{ij \in IJ} g_{ij}(d_{ij}, L_{ij}) \tag{4a}$$

$$\text{Maximize the statistical flow entropy: } f_2 = S \tag{5}$$

$$\text{Minimize the residual head deficits: } f_3 = \text{Max}\langle \max\,[0, (H_n^{req} - H_n)];\ \forall n \rangle \tag{6}$$

$$\text{Subject to: } d_{ij} \in D;\ \forall ij \in IJ \tag{7a}$$

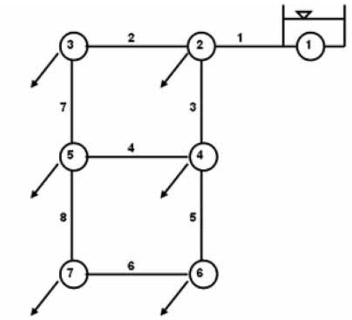
$$\boldsymbol{h}(\boldsymbol{Q_p}, \boldsymbol{H_n}) = \boldsymbol{0} \tag{8a}$$

$g_{ij}(d_{ij}, L_{ij})$ is the cost of pipe ij with diameter d_{ij} and length L_{ij}, and IJ represents the set of pipes in the network. The set D comprises the available discrete pipe diameters. S is the flow entropy. H_n and H_n^{req} are the available and required heads at node n, respectively. The required head is the head above which the demand would be satisfied in full. Simulations in EPANET 2 provided the nodal heads and pipe flow rates, i.e. the solution of Equation (8a) that represents the system of nonlinear equations for the conservation of mass and energy, where the vectors $\boldsymbol{h}$, $\boldsymbol{Q_p}$ *and* $\boldsymbol{H_n}$ represent the nonlinear equations, pipe flow rates and nodal heads, respectively.

Design specifications for Network 1

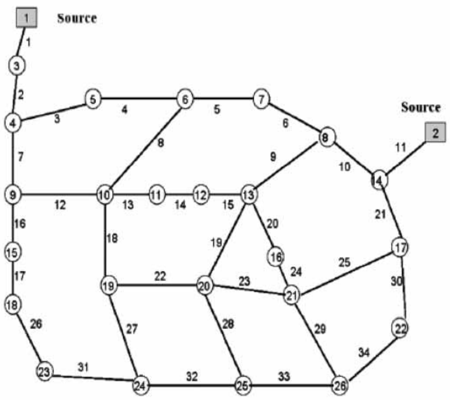
The benchmark network in Figure 1(a) (Alperovits & Shamir 1977) (Network 1) has pipes that are 1,000 m long. The minimum pressure required at the demand nodes to satisfy the demands in full was 30 m. A Hazen-Williams roughness coefficient value of 130 was assumed for new pipes. The other properties of the nodes and discrete pipe diameter options are summarized in Table 1. The design problem of Network 1 has been tackled by numerous researchers, e.g. Sivakumar *et al.* (2016).

(a)



Network 1. The source node (node 1) has a constant water level of 210 m.

(b)



Network 2. The sources have constant water levels of 100 and 95 m respectively.

Figure 1 | Network topologies with node and pipe identifiers. Nodal demand and required head values are in Tables 1 and 4. (a) Network 1. The source node (node 1) has a constant water level of 210 m. (b) Network 2. The sources have constant water levels of 100 and 95 m respectively.

Computational solution method

The elitist non-dominated sorting genetic algorithm NSGA II (Deb *et al.* 2002) was used for the computational solution of the optimization problem. It is practical, robust, efficient and used widely in various disciplines. It maintains diversity by seeking an even distribution of the non-dominated solutions in the objective space using the crowding distance concept (Deb *et al.* 2002). The crowding distance is a measure of the spatial density of the solutions in the objective space and is based on the average distance between a solution and its nearest neighbours.

The problem formulation and solution approach adopted here is penalty-free. The residual head deficit function f_3 in Equation (6) allows infeasible solutions to participate in the evolutionary process, without being impeded by extraneous constraint violation tournaments or penalties. Penalty-free multi-objective evolutionary optimization models have been shown to be highly competitive, having achieved numerous new best-known solutions in the literature on several challenging benchmark problems (Saleh & Tanyimboh 2014; Siew *et al.* 2014).

Experiments were carried out (Czajkowska 2016) to determine suitable values of the GA parameters. Two offspring were produced from two parents using a single-point crossover operator. The probabilities of crossover and random bit-wise mutation were 1.0 and 0.03125, respectively. The population sizes ranged from 100 to 500, and 1,000 generations were allowed. A 4-bit binary coding scheme was used, which gave $2^4 = 16$ binary codes, i.e. two more than the 14 pipe diameter options in Table 1. The average central processing unit (CPU) time for a single optimization run comprising 200,000 function evaluations was about 10 minutes on a PC (2.4 GHz CPU; and 3 GB RAM).

The two redundant binary codes mentioned above were mapped between the phenotype and genotype spaces using two different schemes. In the first mapping scheme, the two redundant codes were distributed evenly in an ordinal

Table 1 | Node data and pipe diameter options for Network 1

Node	Elevation (m)	Demand (L/s)	Required head (m)	Diameter (mm)	Unit cost ($/m)	Diameter (mm)	Unit cost ($/m)
1	[a]210	–	–	25.4	2	304.8	50
2	150	27.78	30	50.8	5	355.6	60
3	160	27.78	30	76.2	8	406.4	90
4	155	33.33	30	101.6	11	457.2	130
5	150	75.00	30	152.4	16	508.0	170
6	165	91.66	30	203.2	23	558.8	300
7	160	55.56	30	254.0	32	609.6	350

[a]210 m is the total head at the supply node (reservoir).

sense among the candidate pipe diameters. More specifically, the pipe diameters of 152.4 and 406.4 mm were doubled in the genotype space. In the second mapping scheme, the two redundant codes were mapped to the smallest pipe diameter, i.e. the 25.4 mm diameter was tripled in the genotype space.

RESULTS AND DISCUSSION

The Pareto sets achieved with a population of 200 were the best. The assessment of the quality of the Pareto sets was based on the variations in the spatial distribution and density of the solutions. Figure 2 (Czajkowska 2016) shows

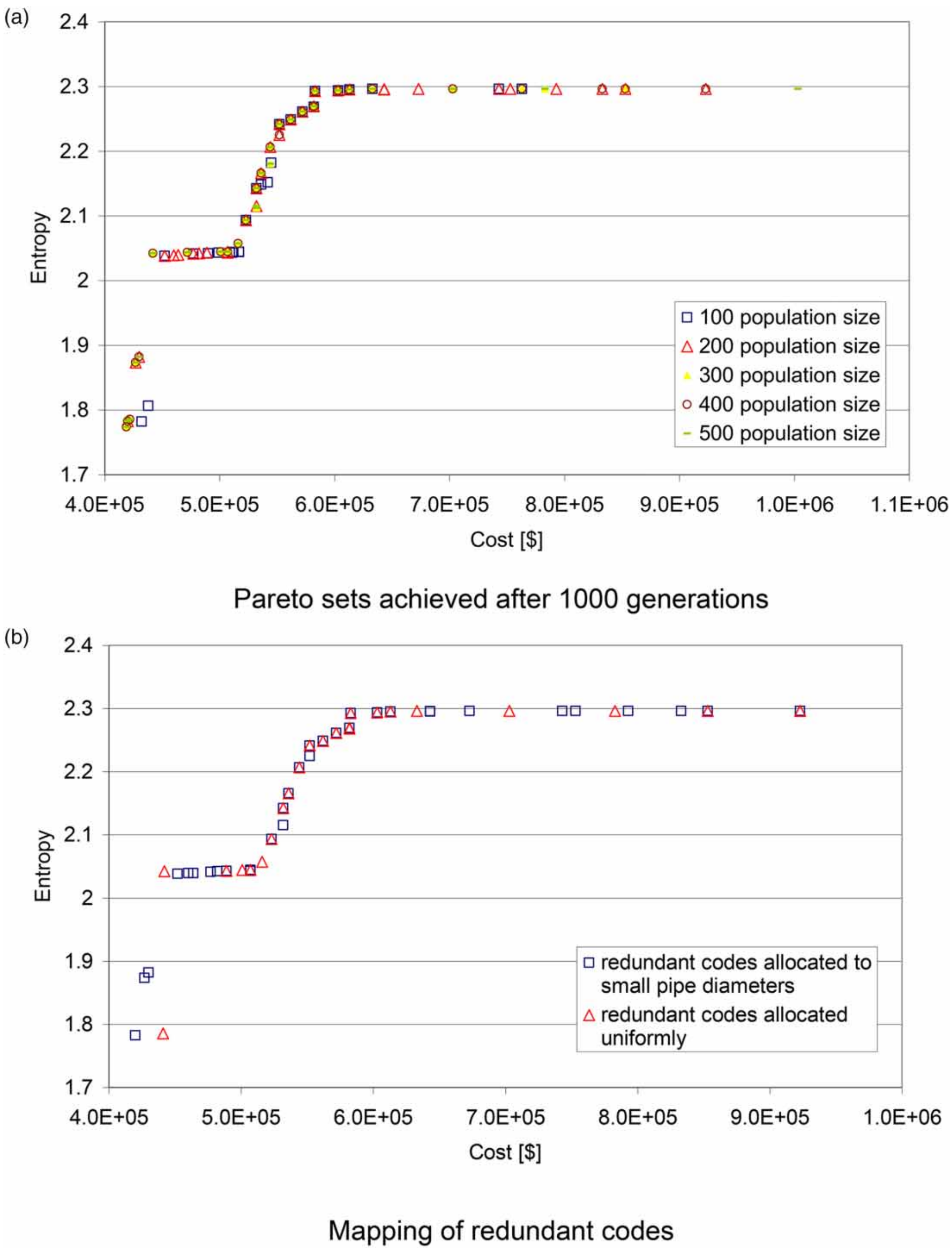


Figure 2 | Effects of population size and mapping of redundant codes on Network 1. (a) Pareto sets achieved after 1,000 generations. (b) Mapping of redundant codes.

that the Pareto set for a population size of 200 had more non-dominated solutions than any other.

The Pareto sets achieved had more infeasible solutions than feasible solutions. The reason is that, effectively, there were three objectives for the infeasible solutions, i.e. cost minimization, entropy maximization and minimization of the residual head deficits. Accordingly, significantly more infeasible solutions were non-dominated compared to feasible solutions (Saxena *et al.* 2013). Conversely there were effectively only two objectives for the feasible solutions, i.e. cost minimization and entropy maximization. The third objective of minimizing the residual head deficits does not apply to feasible solutions, as their deficit value is zero.

With 14 pipe diameter options and two redundant substrings (codes) ($2^4 = 16$), two alternative mappings of the redundant substrings (codes) were investigated. The first (case i) involved a balanced and even distribution in the phenotype space, i.e. the pipe diameters of 152.4 mm and 406.4 mm were doubled in the genotype space. In the second (case ii), the two redundant codes were mapped to the smallest pipe diameter, i.e. the pipe diameter of 25.4 mm was tripled in the genotype space.

Scheme (i) aims to minimize rather than eliminate any representational bias by adopting a relatively even distribution of the redundant codes in the phenotype space. Scheme (ii) clearly favours the smallest pipe diameter and is obviously skewed and maximally biased. There are other mapping schemes that were not considered. As a first step, the objectives were merely to demonstrate a practical alternative to the arbitrary elimination of redundant codes and to compare a symmetrical and even allocation scheme to one that is maximally skewed.

The respective Pareto sets produced by the two mapping schemes used were mutually consistent and mostly overlapped; virtually all of the solutions belong to the same front. However, the mapping in scheme (ii) in which the smallest pipe diameter was over-represented provided more low-cost, low-entropy solutions as shown in Figure 2(b). The mapping scheme with a balanced and even distribution of the redundant codes in scheme (i) gave a better, i.e. more uniform, spatial distribution of the solutions in the Pareto set achieved, with less clustering in the objective space.

Case B: deterministic optimization of initial construction cost

Design optimization problem formulation

Two objectives were selected, namely the initial construction cost that was minimized and the demand satisfaction ratio (Siew & Tanyimboh 2012a) that was maximized. The demand satisfaction ratio is a measure of the hydraulic performance of a water distribution network and is the ratio of the flow delivered to the flow required. In other words, it is the fraction of the demand that the network satisfies at adequate pressure. For networks with satisfactory pressure its value is 1.0; it is less than unity otherwise.

The pressure in a water distribution network may be insufficient due to various reasons e.g. excessively large fire-fighting demands or insufficient flow-carrying capacity in the pipes collectively. Unfeasible candidate solutions inherently do not meet the minimum nodal pressure requirements; consequently, their demand satisfaction ratios are less than unity. Infeasible solutions are produced frequently by recombination and mutation operators in GAs.

The decision variables were the pipe diameters, selected from a set of discrete commercial pipe sizes. The constraints were the equations for the conservation of mass and energy, and the minimum residual pressure requirements at the demand nodes. The equations for the conservation of mass and energy were satisfied using a pressure-driven hydraulic simulation model (EPANET-PDX) (pressure-dependent extension) (Siew & Tanyimboh 2012a). Recent results by Abdy Sayyed *et al.* (2019), who compared pressure-driven and demand-driven hydraulic simulation models in evolutionary algorithms, revealed that the former was superior.

The minimum nodal residual pressure constraints were addressed using the second objective function that is based on the demand satisfaction ratio, i.e. Equation (10). The demand satisfaction ratio was obtained by pressure-driven hydraulic simulation using EPANET-PDX. The robustness and computational efficiency of the EPANET-PDX hydraulic simulation model and accuracy of the results have previously been verified extensively (e.g. Seyoum & Tanyimboh 2017).

The design optimization model is summarized briefly as follows:

$$\text{Minimize } f_1 = \chi^2; \quad 0 < \chi \leq 1.0 \tag{9}$$

$$\text{Maximize } f_2 = \pi^4; \quad 0 \leq \pi \leq 1.0 \tag{10}$$

$$\text{Subject to: } d_{ij} \in D; \quad \forall ij \in IJ \tag{7b}$$

$$\boldsymbol{h}(\boldsymbol{Q_p}, \boldsymbol{H_n}) = \boldsymbol{0} \tag{8b}$$

IJ represents the set of pipes in the network. χ *and* π are normalized network cost and performance functions, respectively. π is the demand satisfaction ratio. Set *D* comprises the discrete pipe diameter options. For solution *s*, χ was defined as

$$\chi_s = \frac{C_s}{\text{Max}(C_s; \ \forall s)}; \quad s = 1, \ldots, PS \tag{11}$$

$$C_s = \sum_{ij \in IJ} g_{ij}(d_{ij}, L_{ij}); \quad s = 1, \ldots, PS \tag{4b}$$

where C_s is the initial construction cost; $g_{ij}(d_{ij}, L_{ij})$ is the cost of pipe *ij* with diameter d_{ij} and length L_{ij}; *PS* is the population size. Equation (8b) represents the constitutive equations.

Design specifications for Network 2

The benchmark network shown in Figure 1(b) was put forward by Kadu *et al.* (2008). It has 34 pipes, 24 demand nodes and two reservoirs (source nodes) with constant water levels of 100.0 and 95.0 m, respectively. The decision variables were the pipe diameters to be selected from a set of discrete commercial pipe sizes. The objective functions were the initial construction cost that was minimized and the demand satisfaction ratio that was maximized. The design optimization problem of Figure 1(b) has been investigated previously by many researchers including Haghighi *et al.* (2011), Mohammadi-Aghdam *et al.* (2015) and Jabbary *et al.* (2016).

The lengths of the pipes are shown in Table 2. The Hazen–Williams formula for the head loss due to friction in a pipe was used, with parameters in SI units as

Table 2 | Pipe connections and unit costs for Network 2

Pipe	Start node	End node	Length (m)	Pipe	Start node	End node	Length (m)	Diameter (mm)	Unit cost (rupees/m)
1	1	3	25	18	10	19	54	150	339.8354
2	3	4	68	19	13	20	63	200	487.6562
3	4	5	78	20	13	16	92	250	656.5072
4	5	6	61	21	14	17	55	300	847.3027
5	6	7	135	22	19	20	98	350	1059.1283
6	7	8	50	23	20	21	82	400	1296.8607
7	4	9	67	24	16	21	56	450	1576.3487
8	6	10	117	25	17	21	90	500	1856.7510
9	8	13	98	26	18	23	63	600	2495.885
10	8	14	63	27	19	24	75	700	3252.0573
11	2	14	18	28	20	25	54	750	3619.0186
12	9	10	58	29	21	26	128	800	4041.7556
13	10	11	26	30	17	22	61	900	4922.5846
14	11	12	42	31	23	24	98	1,000	5911.3075
15	12	13	163	32	24	25	138		
16	9	15	75	33	25	26	110		
17	15	18	71	34	22	26	271		

follows: $\alpha = 1.85$, $\beta = 4.87$ and $\omega = 10.68$. Thus $h_{ij} = \omega L_{ij}(Qp_{ij}/C_{ij})^{\alpha}/d_{ij}^{\beta}$ where h_{ij} is the head loss; Qp_{ij} is the volume flow rate; d_{ij} is the diameter; L_{ij} is the length; and C_{ij} is the roughness coefficient that was taken as 130 for all the pipes.

The nodal demands and minimum nodal heads required, these being the heads above which the demands would be satisfied in full, are shown subsequently along with the results (in Table 4). The 14 discrete pipe diameter options and their associated costs are shown in Table 2.

Computational solution method

The optimization algorithm developed by Siew & Tanyimboh (2012b) was used to solve the problem. It was selected because of its proven computational efficiency. It has outperformed other state-of-the-art algorithms in the literature and achieved the best results on key benchmark optimization problems including the Anytown network that has multiple pumps and service reservoirs (Siew *et al.* 2014).

A distinctive feature of the algorithm is that it allows infeasible solutions to take part in the evolutionary optimization processes along with the feasible solutions without being impeded artificially by any constraint violation penalties or tournaments. Thus, the Pareto-based fitness assessment of the solutions considers only the objective functions f_1 and f_2 (Equations (9) and (10)).

Two offspring were created from two parents using a single-point crossover operator; selection for crossover utilized a binary tournament; and random bit-wise mutation was applied. The initial populations were generated randomly. Also, the two solutions $\mathbf{d}^{\mathrm{U}}$ and $\mathbf{d}^{\mathrm{L}}$ that respectively maximize and minimize the 2-norm of the solution vector were included automatically by default, this being an inherent property of the solution methodology. Equation (12) defines the vectors $\mathbf{d}^{\mathrm{U}}$ and $\mathbf{d}^{\mathrm{L}}$. Equation (13) defines the 2-norm:

$$\boldsymbol{d}^{\mathrm{U}} = \boldsymbol{d}\colon d_{ij} = \max(D), \quad \forall ij \in IJ \\ \boldsymbol{d}^{\mathrm{L}} = \boldsymbol{d}\colon d_{ij} = \min(D), \quad \forall ij \in IJ \tag{12}$$

$$\|\boldsymbol{d}\|_2 = \left(\sum_{ij \in IJ} d_{ij}^2\right)^{1/2} \tag{13}$$

The 14 available discrete pipe diameters and their associated costs are shown in Table 2. With a four-bit substring ($2^4 = 16$), the two redundant codes were allocated to the smallest pipe diameter of 150 mm which, therefore, occurred three times in the genotype space. Ten independent optimization runs were executed. One thousand generations; a population size of 500; mutation rate of 0.05; and crossover probability of 1.0 were adopted in conformity with Siew *et al.* (2014). Thus, the number of hydraulic simulations allowed was 500,000 in each optimization run. The optimization was carried out on a desktop PC (2.5 GHz CPU; and 1.95 GB RAM).

RESULTS AND DISCUSSION

Figure 3(b) shows the progress of the optimization, i.e. the least-cost feasible solution. The initial improvement was very rapid as can be seen, followed by a steady and gradual progress that converged at 127.362 million rupees and 339,000 function evaluations (Run 5). The best solution in the literature costs 125.461 million rupees (Siew *et al.* 2014). Figure 3(a) shows the Pareto sets achieved. They are virtually identical, thus demonstrating the reliability and consistency of the computational solution approach. Similarly, the trajectories of the progress graphs in Figure 3(b) are highly consistent. These observations are supported further by the standard deviation and coefficient of variation of the cost of the least expensive feasible solution from the 10 optimization runs, i.e. 1.129 million rupees and 0.0087, respectively.

The pipe diameters and nodal heads achieved are shown in Tables 3 and 4, respectively, along with previous results from the literature. The nodal head results in Table 4 were obtained using EPANET 2 to allow direct comparisons with results elsewhere in the literature, and they verify further that the solution is feasible.

The results achieved in the 10 optimization runs may be summarized briefly as follows. The minimum, mean and maximum values of the cost of the least expensive feasible solution were: 127.362, 129.114 and 131.287 million rupees, respectively. The corresponding standard deviation and coefficient of variation of the cost of the least expensive feasible solution were 1.129 million rupees and 0.0087,

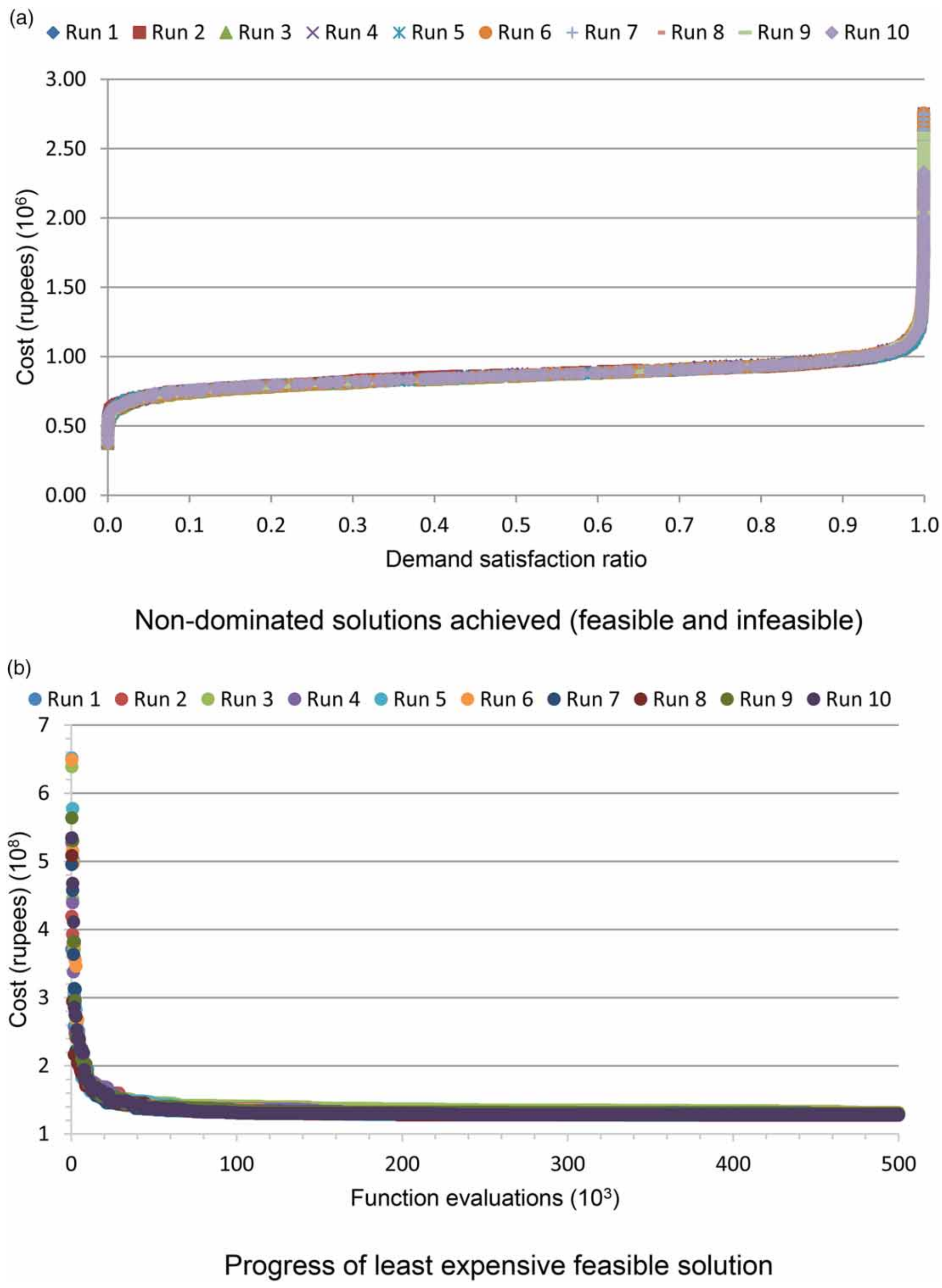


Figure 3 | Computational characteristics of genetic algorithm on Network 2 (a) Non-dominated solutions achieved (feasible and infeasible) (b) Progress of least expensive feasible solution.

respectively. It is thus inferred that consistently good results were obtained relative to the best solution in the literature (125.461 million rupees). Abdy Sayyed *et al.* (2019) have reported a competitive solution recently, with a cost of 125.754 million rupees, after refining the critical path-based reduced solution space in Kadu *et al.* (2008).

Siew *et al.* (2014) achieved the best-known solution, 125.461 million rupees, i.e. 1.5% less expensive than the solution here. Also, 38.2% of the diameters are different from Siew *et al.* (2014), which is somewhat surprising. The respective standard deviations of 1.129 and 1.506 million rupees demonstrate that each study achieved similar results

Table 3 | Pipe diameters from alternative solutions of Network 2

Pipes	Diameters (mm)					
	Kadu *et al.* (2008)	Haghighi *et al.* (2011)	Siew *et al.* (2014)		Abdy Sayyed *et al.* (2019)	Present study
			FSS	**RSS**		
1	1,000	1,000	1,000	900	900	900
2	900	900	900	900	900	900
3	400	400	350	400	350	400
4	350	350	300	250	300	250
5	150	150	150	150	150	150
6	250	250	250	200	250	200
7	800	800	800	800	800	800
8	150	150	150	150	150	150
9	400	400	450	600	450	400
10	500	500	500	600	500	600
11	1,000	1,000	900	900	800	800
12	700	700	700	700	750	750
13	800	800	500	500	500	600
14	400	400	500	500	450	450
15	150	150	150	150	150	150
16	500	500	500	500	500	500
17	350	350	350	350	350	350
18	350	350	400	350	350	450
19	150	150	150	450	150	150
20	200	150	150	150	150	150
21	700	700	700	600	700	700
22	150	150	150	150	150	150
23	400	450	450	150	450	450
24	400	400	350	350	350	350
25	700	700	700	600	700	700
26	250	250	250	250	250	250
27	250	250	250	300	250	250
28	200	200	300	300	300	250
29	300	300	200	200	200	250
30	300	300	250	300	300	250
31	200	200	150	150	200	150
32	150	150	150	150	150	150
33	250	200	150	150	150	150
34	150	150	150	150	150	150
Cost ($R10^6$)	131.679	131.313	125.461	125.826	125.754	127.362

FSS denotes the full solution space; and RSS denotes reduced solution space based on the critical-path solution space reduction method (Kadu *et al.* 2008).

Table 4 | Nodal heads from alternative solutions of Network 2

Nodes	Demands (L/s)	Required heads (m)	Available heads (m) Kadu *et al.* (2008)	Haghighi *et al.* (2011)	Siew *et al.* (2014) FSS	Siew *et al.* (2014) RSS	Abdy Sayyed	Present study
1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
2	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00
3	307.0	85.00	98.95	98.96	98.28	98.26	98.29	98.25
4	75.00	85.00	95.65	95.66	95.04	94.98	95.06	94.92
5	108.0	85.00	90.85	90.85	87.47	90.68	87.51	90.74
6	70.00	85.00	89.40	89.41	85.63	85.07	85.64	85.50
7	52.00	82.00	87.75	87.73	85.83	82.95	85.62	84.48
8	103.0	82.00	89.99	89.96	88.82	89.35	88.31	91.49
9	142.0	85.00	91.77	91.79	91.12	91.05	91.17	90.93
10	192.0	85.00	89.05	89.08	88.30	88.22	89.19	88.85
11	137.0	85.00	88.85	88.88	86.42	86.38	87.33	88.06
12	227.0	85.00	84.98	85.01	85.13	85.12	85.19	85.87
13	247.0	82.00	82.02	81.88	83.25	84.85	82.85	82.38
14	177.0	82.00	94.49	94.49	94.14	94.15	93.49	93.51
15	175.0	85.00	88.44	88.46	87.97	87.92	87.93	87.79
16	150.0	82.00	84.53	84.81	83.11	83.04	82.67	82.37
17	113.0	82.00	90.88	90.88	90.69	90.04	90.09	90.07
18	57.00	85.00	85.46	85.47	85.39	85.39	85.11	85.23
19	77.00	82.00	85.11	85.24	86.14	83.82	85.39	87.48
20	177.0	82.00	82.10	83.78	83.15	82.07	82.77	82.91
21	210.0	82.00	87.39	87.38	87.37	87.00	86.86	86.71
22	90.00	80.00	86.45	86.55	80.69	85.50	85.62	80.92
23	33.00	82.00	82.09	82.07	82.96	83.05	82.06	82.84
24	75.00	80.00	79.94	79.89	80.28	80.86	80.44	80.37
25	58.00	80.00	79.96	79.77	81.10	80.54	80.95	80.04
26	37.00	80.00	82.87	84.04	80.04	80.39	80.67	81.30
Total deficit (m)			0.12	0.46	0.00	0.00	0.00	0.00
Cost ($R10^6$)			131.679	131.313	125.461	125.826	125.754	127.362
Function evaluations			120,000	4,440	436,000	82,400	7,600	339,000

consistently. Furthermore, the maximum difference in the least cost in Siew *et al.* (2014) for various population sizes and mutation rates was 1.02%. Siew *et al.* (2014) mapped the two redundant codes to the largest pipe diameter.

The mapping scheme used here achieved a more expensive best solution than Siew *et al.* (2014). The difference in cost may be attributed to the effects of the mapping schemes, besides the number of optimization runs. However, the consistency of the results from the algorithm has been demonstrated previously to be very high; the coefficient of variation of 0.0087 achieved here reinforces this view. The mapping scheme used here with both redundant codes, corresponding to an over-representation of 12.5%, mapped to the smallest pipe diameter is highly skewed and maximally biased, with the potential to misdirect the search to prioritize the infeasible region of the solution space. Conversely, the largest of the least-cost values in Siew *et al.* (2014) that mapped the two redundant codes to the largest pipe diameter, an over-representation of 12.5%, was 132.986 million rupees compared to 131.287 million rupees in the present study. The difference is 1.699 million rupees (1.3%). Although the two mapping schemes seem to be mere opposites, the essential difference is that over-representation of the largest pipe diameter prioritizers the feasible region [with better results overall] compared to over-representation of the smallest pipe diameter that prioritizes the infeasible region.

The objective of minimizing the cost inherently favours the small pipe diameters as they are less expensive. However, it seems that over-representation of the largest pipe diameter increases diversity in the population as it increases the probability of selecting the largest diameter, which in turn helps to avoid premature convergence and consequently leads to better solutions. In this regard, it is interesting to note that the standard deviation of 1.506 million rupees in Siew *et al.* (2014) exceeds the corresponding value of 1.129 million rupees in the present study by 33%. This supports the idea that over-representation of the largest diameter helps to improve diversity in the population of candidate solutions.

Based on these observations, it can be stated that over-representation of the largest pipe diameter outperformed over-representation of the smallest pipe diameter. Over-representation of the largest pipe diameter yielded: (a) smaller values of the minimum and average cost of the best solution, i.e. the solutions were more cost effective overall; and (b) larger values of the standard deviation and maximum cost of the best solution, i.e. both the diversity and range – the difference between the largest and smallest values of the cost of the best solution – were superior. Conversely over-representation of the smallest pipe diameter yielded: (a) larger values of the minimum and average cost of the best solution, i.e. the solutions were more expensive overall; and (b) smaller values of the standard deviation and maximum cost of the best solution, i.e. both the diversity and range were inferior. Thus, over-representation of the largest pipe diameter outperformed over-representation of the smallest pipe diameter based on cost effectiveness and diversity.

Additional studies to investigate these observations further are worth considering for the future. More fundamentally, given the importance of diversity in the gene pool, it is possible that even better results in terms of solution quality and computational efficiency could be achieved with the most even and balanced mapping schemes. This view is consistent with the results and observations from Network 1, in which over-representation of the smallest pipe diameter achieved suboptimal results in terms of cost and diversity compared to the symmetrical and even mapping.

A wealth of evidence in the literature shows that better results are achieved in evolutionary algorithms if infeasible solutions are allowed to compete with feasible solutions on an equal basis, without being impeded by extraneous constraint violation penalties e.g. by being discarded summarily, with no further consideration whatsoever. The premature loss of infeasible solutions depletes the gene pool of valuable genetic materials which, therefore, results in a suboptimal search (Woldesenbet *et al.* 2009; Eskandar *et al.* 2012; Saleh & Tanyimboh 2014; Siew *et al.* 2014). Similarly, the chromosomes with redundant binary codes carry valuable information and, accordingly, more practical and effective mapping routines for redundant codes would improve the efficiency of GAs and the solutions achieved. Finally, the computational efficiency results in Abdy Sayyed *et al.* (2019) based on the function evaluations (Table 4) demonstrate the effectiveness of reducing the solution space *a priori*. More research on seamless solution space reduction procedures is thus indicated also.

CONCLUSIONS

Powerful and robust, GAs are highly effective for optimization problems with discontinuous, complex and poorly understood solution spaces. Coding and decoding translate solutions between the genotype and phenotype spaces, and binary coding is often adopted as it is straightforward to implement and lends itself to problems with decision variables that have discrete values. It is therefore used widely and underpins most existing GA formulations.

Redundant binary codes occur frequently. They arise if an encoded parameter belongs to a finite discrete set whose cardinal number is not a power of 2. A common technique used to address redundant codes is to discard any chromosomes in which they occur. This approach has the shortcoming that it inevitably leads to the premature loss of valuable information and useful genetic materials and a reduction in the effectiveness of the GA.

Redundant binary codes were investigated in this article using two benchmark networks in the literature in conjunction with two different GAs. Both GA formulations do not include constraint violation penalties. They allow *non-dominated infeasible* solutions to compete for selection and crossover without being impeded arbitrarily; the evidence in the literature shows that the subpopulation of competitive infeasible solutions improves the algorithm's effectiveness.

The results indicate that different mapping schemes with varying levels of representational bias between the genotype and phenotype spaces can lead to significantly different solutions in the phenotype space and suboptimal solutions in the objective space, depending on the available computational budget. Also, it was observed that a balanced, relatively even and unbiased allocation of the redundant binary codes with respect to the phenotype space as a whole, achieved good results in terms of the quality and spatial distribution of the solutions in the final Pareto set. Practical guidance on the handling of redundant binary codes is lacking. The main inference from the results is that mapping schemes that improved diversity in the population of candidate solutions achieved better results, which may pave the way for the development of practical and effective mapping schemes.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Abdy Sayyed, M. A. H., Gupta, R. & Tanyimboh, T. T. 2019 Combined flow and pressure deficit-based penalty in GA for optimal design of water distribution network. *ISH Journal of Hydraulic Engineering* doi:10.1080/09715010.2019.1604180.

Alperovits, E. & Shamir, U. 1977 Design of optimal water distribution systems. *Water Resources Research* **13** (6), 885–900.

Avila-Melgar, E. Y., Cruz-Chávez, M. A. & Martinez-Bahena, B. 2017 General methodology for using EPANET as an optimization element in evolutionary algorithms in a grid computing environment for water distribution network design. *Water Supply* **17** (1), 39–51. doi:10.2166/ws.2016.101.

Ciaponi, C., Creaco, E., Franchioli, L. & Papiri, S. 2017 The importance of the minimum path criterion in the design of water distribution networks. *Water Supply* **17** (6), 1558–1567. doi:10.2166/ws.2017.061.

Creaco, E., Fortunato, A., Franchini, M. & Mazzola, M. R. 2015 Water distribution network robust design based on energy surplus index maximization. *Water Supply* **15** (6), 1253–1258. doi:10.2166/ws.2015.091.

Czajkowska, A. M. 2016 *Maximum Entropy Based Evolutionary Optimization of Water Distribution Networks Under Multiple Operating Conditions and Self-Adaptive Search Space Reduction Method*. PhD thesis. University of Strathclyde, Glasgow, UK.

Deb, K. 2001 *Multi-objective Optimization Using Evolutionary Algorithms*. Wiley, Chichester, UK, pp. 106–120.

Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. 2002 A fast and elitist multiobjective genetic algorithm: NSGA II. *IEEE Transactions on Evolutionary Computation* **6**, 182–197.

Eshelman, L. J. & Schaffer, J. D. 1993 Real-coded GAs and interval-schemata. *Foundations of Genetic Algorithms* **2**, 187–202.

Eskandar, H., Sadollah, A., Bahreininejad, A. & Hamdi, M. 2012 Water cycle algorithm – a novel metaheuristic optimization method for solving constrained engineering optimization problems. *Computers and Structures* **110–111**, 151–166. doi:10.1016/j.compstruc.2012.07.010.

Fadaee, M., Mahdavi-Meymand, A. & Zounemat-Kermani, M. 2020 Suspended sediment prediction using integrative soft computing models: on the analogy between the butterfly optimization and GAs. *Geocarto International* doi:10.1080/10106049.2020.1753821.

Feng, Y., Wang, C., Briseghella, B., Fenu, L. & Zordan, T. 2020 Structural optimization of a steel arch bridge with genetic algorithm. *Structural Engineering International* doi:10.1080/10168664.2020.1773373.

Gen, M., Cheng, R. & Lin, L. 2008 *Network Models and Optimization: Multiobjective Genetic Algorithm Approach*. Springer-Verlag London Limited.

Haghighi, A., Samani, H. M. V. & Samani, Z. M. V. 2011 GA-ILP method for optimisation of water distribution networks. *Water Resources Management* **25** (7), 1791–1808.

Jabbary, A., Podeh, H. T., Younesi, H. & Haghiabi, A. H. 2016 Development of central force optimization for pipe-sizing of water distribution networks. *Water Science & Technology Water Supply* **16** (5), 1–13. doi:10.2166/ws.2016.051.

Jaynes, E. T. 1957 Information theory and statistical mechanics. *Physical Review* **106**, 620–630 and **108,** 171–190.

Kadu, M. S., Gupta, R. & Bhave, P. R. 2008 Optimal design of water networks using a modified genetic algorithm with reduction in search space. *Journal of Water Resources Planning-ASCE* **134** (2), 147–160.

Khinchin, A. I. 1953 The entropy concept in probability theory. *Uspekhi Matematicheskikh Nauk* **8** (3), 3–20.

Khinchin, A. I. 1957 *Mathematical Foundations of Information Theory*. Dover, New York, pp. 1–28.

Lotfian, R., Gholamnejad, J. & Lardkeyvan, Y. M. 2020 Effective solution of the long-term open pit production planning problem using block clustering. *Engineering Optimization* doi:10.1080/0305215X.2020.1771703.

Mohammadi-Aghdam, K., Mirzaei, I., Pourmahmood, N. & Pourmahmood-Aghababa, M. 2015 Application of dynamic mutated particle swarm optimization algorithm to design water distribution networks. *Journal of Water & Wastewater* **26** (4), 88–99.

Naseri, H., Fani, A. & Golroo, A. 2020 Toward equity in large-scale network-level pavement maintenance and rehabilitation scheduling using water cycle and GAs. *International Journal of Pavement Engineering* doi:10.1080/10298436.2020.1790558.

Pattanaik, J. K., Basu, M. & Dash, D. P. 2020 Improved real-coded genetic algorithm for fixed head hydrothermal power system. *IETE Journal of Research* doi:10.1080/03772063.2020.1785339.

Phan, D. T., Lim, J. B. P., Tanyimboh, T. T., Lawson, R. M., Xu, Y., Martin, S. & Sha, W. 2013 Effect of serviceability limits on optimal design of steel portal frames. *Journal of Constructional Steel Research* **86**, 74–84.

Poojitha, S. N., Singh, G. & Jothiprakash, V. 2020 Improving the optimal solution of GoYang network – using genetic algorithm and differential evolution. *Water Supply* **20** (1), 95–102. https://doi.org/10.2166/ws.2019.139.

Rathi, S., Gupta, R., Palod, N. & Ormsbee, L. 2020 Minimizing the variance of flow series using a genetic algorithm for reliability-based design of looped water distribution networks. *Engineering Optimization* **52** (4), 637–651.

Rossman, L. A. 2000 *EPANET 2 User's Manual*. Water Supply and Water Resources Division, National Risk Management Research Laboratory, US EPA, Cincinnati, OH.

Saleh, S. H. A. & Tanyimboh, T. T. 2014 Optimal design of water distribution systems based on entropy and topology. *Water Resources Management* **28** (11), 3555–3575. doi:10.1007/s11269-014-0687-y.

Saxena, K. S., Duro, J. A., Tiwari, A. & Deb, K. 2013 Objective reduction in many-objective optimization: linear and non-linear algorithms. *Evolutionary Computation* **17** (1), 77–99.

Seyoum, A. G. & Tanyimboh, T. T. 2017 Integration of hydraulic and water quality modelling in distribution networks: EPANET-PMX. *Water Resources Management* **31** (14), 4485–4503. doi:10.1007/s11269-017-1760-0.

Shannon, C. 1948 A mathematical theory of communication. *AT&T Technical Journal* **27**, 379–428.

Shende, S. & Chau, K. W. 2019 Design of water distribution systems using an intelligent simple benchmarking algorithm with respect to cost optimization and computational efficiency. *Water Supply* **19** (7), 1892–1898. doi:10.2166/ws.2019.065.

Siew, C. & Tanyimboh, T. T. 2012a Pressure-dependent EPANET extension. *Water Resources Management* **26** (6), 1477–1498. doi:10.1007/s11269-011-9968-x.

Siew, C. & Tanyimboh, T. T. 2012b Penalty-free feasibility boundary-convergent multi-objective evolutionary algorithm for the optimization of water distribution systems. *Water Resources Management* **26** (15), 4485–4507. doi:10.1007/s11269-012-0158-2.

Siew, C., Tanyimboh, T. T. & Seyoum, A. G. 2014 Assessment of penalty-free multi-objective evolutionary optimization approach for the design and rehabilitation of water distribution systems. *Water Resources Management* **28** (2), 373–389. doi:10.1007/s11269-013-0488-8.

Siew, C., Tanyimboh, T. T. & Seyoum, A. G. 2016 Penalty-free multi-objective evolutionary approach to optimization of Anytown water distribution network. *Water Resources Management* **30** (11), 3671–3688. doi:10.1007/s11269-016-1371-1.

Sivakumar, P., Prasad, R. K. & Chandramouli, S. 2016 Uncertainty analysis of looped water distribution networks using linked EPANET-GA method. *Water Resources Management* **30** (1), 331–358. doi:10.1007/s11269-015-1165-x.

Surco, D. F., Vecchi, T. P. B. & Ravagnani, M. A. S. S. 2018 Optimization of water distribution networks using a modified particle swarm optimization algorithm. *Water Supply* **18** (2), 660–678. doi:10.2166/ws.2017.148.

Tanyimboh, T. T. 2017 Informational entropy: a failure tolerance and reliability surrogate for water distribution networks. *Water Resources Management* **31**, 3189–3204.

Vamvakeridou-Lyroudia, L. S., Walters, G. A. & Savic, D. A. 2005 Fuzzy multiobjective optimization of water distribution networks. *Journal of Water Resources Planning-ASCE* **131** (6), 467–476.

Woldesenbet, Y. G., Yen, G. G. & Tessema, B. G. 2009 Constraint handling in multiobjective evolutionary optimization. *IEEE Transactions on Evolutionary Computation* **13** (3), 514–525.

Xu, Y., Liu, X., Hu, X., Huang, G. & Meng, N. 2020 A genetic-algorithm-aided fuzzy chance-constrained programming model for municipal solid waste management. *Engineering Optimization* **52** (4), 652–668.

Yazdi, J., Yoo, D. G. & Kim, J. H. 2017 Comparative study of multi-objective evolutionary algorithms for hydraulic rehabilitation of urban drainage networks. *Urban Water Journal* **14** (5), 483–492. doi:10.1080/1573062X.2016.1223319.

First received 11 May 2020; accepted in revised form 3 November 2020. Available online 19 November 2020