Probabilistic building block identification for the optimal design and rehabilitation of water distribution systems

Ralph J. Olsson, Zoran Kapelan and Dragan A. Savic

ABSTRACT

The multi-objective design and rehabilitation of water distribution systems (WDS) is defined as the search for the set of system designs which offers the best trade-off between competing design objectives. Typically these objectives will consist of the cost of implementing a system design and a measure of the performance of that system. These measures are often in competition since improvements in the performance of a system generally come at a cost. Here three genetic algorithms which use probabilistic methods to identify building blocks—the Univariate Marginal Distribution Algorithm (UMDA) (Mühlenbein 1997), the hierarchical Bayesian Optimisation Algorithm (hBOA) (Pelikan 2002) and the Chi-Square Matrix methodology (Aporntewan & Chongstitvatana 2004)—are compared to the well-known multi-objective evolutionary algorithm NSGAII (Deb *et al.* 2002) for the multi-objective design and rehabilitation of water distribution systems. For single-objective problems the identification of building blocks has been seen to make evolutionary algorithms more scalable to large problems than simple genetic algorithms. In this paper these algorithms are shown to offer significantly better solutions than NSGA-II for the case of large systems. However, this improvement comes at the expense of diversity of solutions in the fronts identified.

Key words | building block identification, genetic algorithms, multi-objective optimisation, water distribution system design and rehabilitation

Ralph J. Olsson (corresponding author)
Zoran Kapelan
Dragan A. Savic
Centre for Water Systems,
University of Exeter,
Harrison Building,
North Park Road,
Exeter EX4 4QF,
UK
Tel.: +44 1392 264075
E-mail: *ralpholsson@btinternet.com*

NOTATION

N	the number of decision variables
m	the number of objectives in the optimisation problem
n	the size of the set of promising solutions
$\bar{X} = (X_0, \ldots, X_{N-1})$	a vector of (typically binary) decision variables
$\bar{x} = (x_0, \ldots, x_{N-1})$	a specific instance of the vector $\bar{X}$
Π_i	the set of parents of the decision variable X_i
π_i	a specific instance of the set Π_i
M	an N by N matrix in which the (i,j)th entry is the chi-square goodness of fit statistic for the variables X_i and X_j
$\bar{y} = (y_0, \ldots, y_{m-1})$	the vector of objective values for a given solution
$F = \{\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_f\}$	a front of size f of mutually non-dominated solutions
$S(F)$	the size of space covered metric value for the front F
$C(F_1, F_2)$	the coverage of two sets metric for the fronts F_1 and F_2

doi: 10.2166/hydro.2009.047

INTRODUCTION

The problem of the optimal design and rehabilitation of Water Distribution Systems (WDS) has received much attention in the literature. Early methodologies employed

local search algorithms to find solutions (Schaake & Lai 1969; Alperovits & Shamir 1977; Morgan & Goulter 1985; Loganathan *et al.* 1995). However, the complexity, non-linearity and discrete nature of WDS optimisation problems seriously limits these methodologies as they can be trapped in a local optima in the solution space (the set of all possible solutions).

Evolutionary algorithms (EAs) have proved to be a valuable tool in the search for good solutions to WDS optimisation problems. EAs were first employed to solve WDS design and rehabilitation problems in which only a single objective was optimised (Simpson *et al.* 1994; Dandy *et al.* 1996; Savic & Walters 1997; Vairavamoorthy & Ali 2000; Babayan *et al.* 2005). Any further competing objectives and/or constraints (e.g. minimum pressure constraints) were incorporated into the fitness evaluation of a solution as a penalty function added to the solution's cost.

More recently multi-objective evolutionary algorithms such as NSGA-II (Deb *et al.* 2002) have been applied to the design and rehabilitation of water distribution systems in which more than one objective compete in the solution space (Halhal *et al.* 1997; Walters *et al.* 1999; Farmani *et al.* 2003, 2005*a*; Prasad & Park 2004; Kapelan *et al.* 2005; Babayan *et al.* 2007). NSGA-II is a genetic algorithm which searches for a Pareto optimal front which is the best possible set of mutually non-dominated solutions. The ability of NSGA-II to identify the best trade-off between a set of competing objectives makes it a useful tool throughout the water industry with recent applications including the calibration of storm water runoff models (Hejazi *et al.* 2008) and the calculation of reservoir operating rules (Kim *et al.* 2008).

Simple EAs, including NSGA-II, perform operations such as crossover and mutation on a pair of parent chromosomes to produce a pair of child chromosomes. These operations make no allowance for any dependence between decision variables and so the optimisation process can be deceived into ignoring important solutions. Consider a simple WDS with a reservoir at node A and a demand at node D. If the design options are to install pipes in series between nodes A and B, nodes B and C, and nodes C and D then, due to the cost of including a pipe, a simple genetic algorithm which considers each pipe separately will prefer not to put a pipe at any position unless there are already pipes in the other two positions, since to do so would increase the cost with no improvement in pressure at node D. Suppose this design problem is formulated as the vector (i,j,k) where i, j and k are 1 if pipes are to be laid between nodes A and B, B and C, and C and D, respectively, and zero otherwise. The evolutionary techniques of crossover and mutation applied to single bits will tend to lead toward the local optimum (0,0,0) rather than the global optimum (1,1,1). This type of function is known in the literature as the trap function (see, for example, Deb & Goldberg 1991).

To solve such problems an EA capable of learning important interactions between decisions (i.e. capable of 'linkage learning') is desirable. Initial tests (Olsson *et al.* 2007*a*,*b*) indicate that such an EA has significant potential to reduce the number of evaluations required to identify the optimal rehabilitation solution. To investigate this further three well-known Probabilistic Model Building Genetic Algorithms (PMBGAs) (see Pelikan *et al.* 1999) are compared to the previously used NSGA-II algorithm in the multi-objective WDS optimisation context. PMBGAs perform linkage learning by using probabilistic methods to identify short, highly dependent groups of decision variables known as building blocks (BBs). The three PMBGAs analysed here are as follows: the Univariate Marginal Distribution Algorithm (Mühlenbein 1997), the hierarchical Bayesian Optimisation Algorithm (Pelikan 2002) and the Chi-Square Matrix methodology (Aporntewan & Chongstitvatana 2004).

WDS DESIGN AND REHABILITATION PROBLEM

The problem of the optimal design or rehabilitation of WDS is to find the most effective way of investing money into a new (design) or existing (rehabilitation) distribution network. Investment in network infrastructure can take many forms including laying new pipes, installing new tanks and pumps, and cleaning existing pipes. Effectiveness will typically be measured as a benefit relative to some worst case design (for example, Walters *et al.* (1999), who consider benefit relative to the system performance before rehabilitation) or as a shortfall relative to a set of design ideals (for example, Simpson *et al.* (1994), who consider nodal pressure shortfalls for a system).

For each of the problems considered in this paper network layouts are predefined based on the locations and elevations of water sources and demands, and pipelines are predetermined based on either engineering considerations (such as following the paths of roads) in the case of new designs or on the positions of existing pipes in the case of rehabilitations. Each system has a set of requirements such as the need to supply each node with a given demand at a given pressure. The potential investments for the system will be encoded as a set of design variables. These will include the diameters of new pipes used, the positions and sizes of new tanks, pump operation schedules, and the decision whether to clean old pipes or not.

For a given set of design variables the system will be analysed and the measure of effectiveness will be calculated. In this paper the hydraulic solver Epanet2 (Rossman 2000) is used to analyse the system. Epanet2 uses a model in which a set of nonlinear equations is formed representing energy and mass balances within the system. Epanet2 solves these equations to find the pressures and flows within the system. The measure of effectiveness is calculated based on these pressures and flows and the total cost of the investment represented by the design variables is calculated.

The aim of the multi-objective optimisation presented here is to find the sets of design variables (solutions) which offer the best trade-off between maximising the effectiveness (minimising the shortfall) and minimising the investment cost.

OPTIMISATION ALGORITHMS

The four algorithms considered are described below. Each algorithm is designed to search for the Pareto optimal front of a problem. The Pareto optimal front is the set of all non-dominated solutions in a solution space, that is, the solutions in the Pareto optimal front cannot be out-performed in all objectives by any other possible solution. The goal of Pareto optimisation is to find a front which is both as close as possible to the Pareto optimal front, and has as much diversity as possible among its members.

In the following section solution A is said to be dominated by solution B ($A \prec B$) if and only if B performs better than A in at least one objective, and no worse than A in all other objectives. If $A \not\prec B$ and $B \not\prec A$, A and B are said to be mutually non-dominated.

THE NON-DOMINATED SORTING GENETIC ALGORITHM-II

The Non-dominated Sorting Genetic Algorithm-II (NSGA-II) (Deb *et al.* 2002) is a multi-objective evolutionary algorithm which employs the traditional techniques of crossover and mutation in the search for the Pareto optimal front. The algorithm works as follows. In the first step, the initial population is created and sorted into fronts whereby the first front contains those solutions which are not dominated by any other solutions in the population, the second front contains those solutions which are only dominated by members of the first front, and so on. Next the solutions within each front are assigned a crowding distance which gives a measure of how dense the front is in the vicinity of that solution (see Deb *et al.* 2002).

In the second algorithm step, an offspring population is created by selecting members of the current population and performing the operations of crossover and mutation to produce new solutions. When selecting solutions (using, for example, tournament or roulette wheel selection) candidates are compared by their front number with lower numbered fronts given preference. In the event of two solutions coming from the same front the solution with the greater crowding distance is chosen.

In the third algorithm step, a new population is created by sorting solutions from both the current population and the offspring population into fronts. Once the assignment of a new front would raise the population size above the required number, solutions are added from the next front starting with the solution with the greatest crowding distance until the population is full.

The search process then involves repeating steps two and three until some termination criteria is met.

In this paper selection is achieved using tournament selection, in which a set of t individuals is chosen at random from the population, and the best solution in the set is selected. The termination criteria is that a maximum number of generations has been reached.

Each of the following algorithms operates identically to NSGA-II except in the creation of the offspring population (i.e. the second step).

THE UNIVARIATE MARGINAL DISTRIBUTION ALGORITHM

The Univariate Marginal Distribution Algorithm (Mühlenbein 1997) creates an offspring population by building and sampling a probabilistic model describing a set of promising solutions. This probabilistic model approximates the distribution of solutions in this set as the product of the univariate marginal distributions for each gene.

First the set of promising solutions, the parent population, is selected from the current population. The size of this set is given by a parameter of the algorithm as a percentage of the population size. The solutions in this set can be thought of as instances of the random vector $\bar{X} = (X_0, \ldots, X_{N-1})$, where N is the number of decision variables. In UMDA, the probability that a solution will have the vector $\bar{x} = (x_0, \ldots, x_{N-1})$ is approximated as the product of the univariate marginal probabilities that the variables X_i will take the values x_i:

$$P(\bar{X} = \bar{x}) \approx \prod_{i=0}^{N-1} P(X_i = x_i) \tag{1}$$

These univariate marginal distributions are calculated from the parent population as follows:

$$P(X_i = x_i) = \frac{\text{number of solutions in population with } X_i = x_i}{\text{number of solutions in population}} \tag{2}$$

Once the $P(X_i = x_i)$ have been calculated for all instances of all decision variables, the model is sampled to create the offspring population. To create a new solution each decision variable is randomly generated based on the $P(X_i = x_i)$. Once the required number of new solutions have been assigned to the offspring population, the next generation is created as in the third step of NSGA-II described above.

THE HIERARCHICAL BAYESIAN OPTIMISATION ALGORITHM

The hierarchical Bayesian Optimisation Algorithm (Pelikan 2002) differs from UMDA in the way the probabilistic model is built and sampled. hBOA employs a Bayesian network to learn an approximation of the probability $P(\bar{X}_i = \bar{x}_i)$ in which important interactions between decisions are encoded as dependences between variables.

A Bayesian network is an acyclic directed graph which approximates the probability distribution as follows:

$$P(\bar{X} = \bar{x}) \approx \prod_{i=0}^{N-1} P(X_i = x_i | \Pi_i = \pi_i) \tag{3}$$

where Π_i, with instance π_i, is the set of parents of X_i. The network consists of a set of nodes representing the variables X_i and a set of directed edges representing the dependences such that an edge from X_j to X_i denotes that X_j is a parent of X_i. In hBOA the network is built by using a hill climbing algorithm to build a set of decision graphs, one for each variable, in such a way as to maximise a scoring metric. Any dependence identified in a decision graph corresponds to the addition of an edge in the Bayesian network. Any dependence which would create a cycle in the network is discarded. For details of the building of the Bayesian network see Pelikan (2002).

Once a Bayesian network describing the set of promising solutions has been built, it can be sampled to create the offspring population. Variables are sampled using the corresponding decision graph, working in such an order that all of the parents of a variable are sampled before that variable (this will always be possible since the Bayesian network is acyclic).

hBOA has been seen to solve a number of challenging problems in low order polynomial time (Pelikan 2002). For the case of the least cost design and rehabilitation of water distribution systems, hBOA has been seen to offer significant improvements over simple GAs, particularly for large systems (Olsson *et al.* 2007*a*). However, for particularly large systems hBOA's computational complexity has proven to limit its applicability (Olsson *et al.* 2007*b*).

THE CHI-SQUARE MATRIX FOR BUILDING BLOCK IDENTIFICATION

While the two PMBGAs above build and sample a probabilistic model describing the joint distribution $P(\bar{X}_i = \bar{x}_i)$, the Chi-Square Matrix approach (Aporntewan & Chongstitvatana 2004) instead identifies BBs by using probabilistic analysis to partition the decision vector. This method assumes the decision vector has binary entries.

Given a set of promising solutions, a chi-square matrix is built. This matrix, M, is an N by N symmetric matrix in which the element $m_{i,j}$ is the chi-square statistic for the variables X_i and X_j:

$$m_{i,j} = \begin{cases} \text{Chi}-\text{Square}(X_i; X_j); & i \neq j; \\ 0; & \text{otherwise} \end{cases} \tag{4}$$

The chi-square statistic is defined as follows:

$$\text{Chi}-\text{Square}(X_i, X_j) = \sum_{(x_i,x_j)} \frac{(X^{x_i,x_j}(X_i, X_j) - n/4)^2}{n/4}, \quad (x_i, x_j) \in \{0, 1\} \tag{5}$$

where $C^{x_i,x_j}(X_i, X_j)$ is the number of individuals in the set of promising solutions with $X_i = x_i$ and $X_j = x_j$, and n is the total number of promising solutions.

If there is no dependence between X_i and X_j then the expected value of $C^{x_i,x_j}(X_i, X_j)$ will be $n/4$ for each (x_i,x_j). Thus the chi-square statistic will be close to zero. Once the chi-square matrix as been computed, a partitioning algorithm is employed to sort the variables X_i into sets such that the chi-square values of variables in the same partition subset are high, and the chi-square values of variables in different subsets are low. For details of this partitioning algorithm see Aporntewan & Chongstitvatana (2004). Each partition subset represents a building block.

Once the variables have been partitioned, parents are paired up and a crossover operator is applied in such a way that all variables in the same BB are taken from the same parent. Once the offspring population has been generated the algorithm continues as with NSGA-II.

For large systems, the chi-square method is considerably less computationally demanding than hBOA.

CASE STUDIES

The three probabilistic evolutionary algorithms outlined above are compared to NSGA-II for the multi-objective optimisation of three WDS design and rehabilitation problems, the New York tunnels problem, the Anytown network and a real-world WDS. In each case the following two performance metrics of Zitzler & Thiele (1999) are used to compare the fronts found.

For an m objective optimisation, let $\bar{y} = (y_0, \ldots, y_{m-1})$ be the vector of objective values for a solution, and let $F = \{\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_f\}$ be a front of size f. The *size of space covered* metric, $S(F)$, gives the union of the spaces covered by each of the $\bar{y}_i$, where the space covered is the volume enclosed by the $\bar{y}_i$ and the axes. In the case of a two-objective optimisation this is simply the area of the rectangle formed by the two axes and the point (y_0,y_1). In order for this metric to be meaningful it is assumed that all objective values are non-negative and that either all objectives are being maximised or all are being minimised. This metric considered alone can be deceptive as it can favour fronts with poor coverage over fronts with good coverage. For example, consider the two fronts $F_1 = \{(2,2)\}$ and $F_2 = \{(3,1), (2,2),(1,3)\}$, in which two objectives are being minimised. The first front has the metric value $S(F_1) = 4$ while the second has a metric value of $S(F_2) = 6$. Since the objectives are being minimised a better front should cover a smaller area, yet the second front is clearly better than the first.

For this reason a coverage metric is also used. For two fronts, F_1 and F_2, the *coverage of two sets* metric $C(F_1,F_2)$ gives the proportion of points in F_2 which are covered by (dominated by or equal to) points in F_1. This gives a comparative measure of two fronts, but does not provide any information about how good either front is in its own right. For the case of the two fronts considered above, all of the points in F_1 are covered by F_1 (since the point (2,2) in F_1 covers the same point in F_1) so $C(F_2,F_1) = 1$: however, only one of the three points in F_1 is covered by F_1 so $C(F_1,F_2) = 1/3$. Thus F_2 can be seen to give better coverage than F_1.

The New York tunnels problem

The New York tunnels system, as shown in Figure 1, consists of a single reservoir and a network of 21 pipes

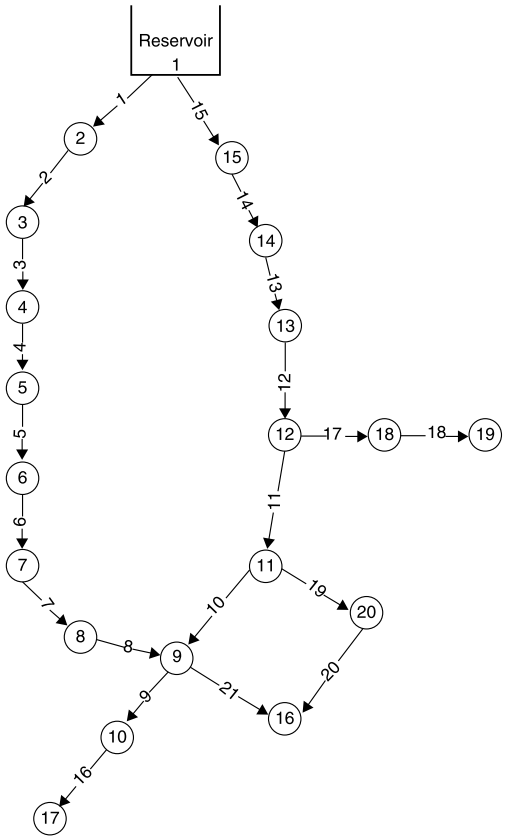

Figure 1 | The New York tunnels network.

(see Schaake & Lai 1969). There is a minimum head requirement of 255 ft (77.72 m) at all demand nodes except 16 and 17, which have head requirements of 260 ft (79.25 m) and 272.8 ft (83.15 m), respectively. The reservoir at node 1 has a fixed head of 300 ft (91.44 m). To satisfy these requirements any of the 21 existing pipes may be duplicated by laying a new pipe alongside. New pipes can be chosen from 15 available diameters and so there are 16 options for each pipe including 'do nothing'. Each of these decisions is encoded using a 4-bit binary string and so each candidate solution is represented by a vector with 84 bits.

The cost in US dollars ($) of a candidate solution is given by

$$C = \sum_i 1.1 D_i^{1.24} L_i \tag{6}$$

where D_i and L_i are the diameter (in inches) and length (in feet) of pipe i, and the sum is taken over all pipes. The two objectives considered in this paper are to minimise the rehabilitation cost C and to minimise the total head deficit H. H is the sum of heads required minus heads available over all nodes for which the head available is less than the head required, where all heads are measured in feet.

For each algorithm a sensitivity analysis was performed to find the values of the key parameters which gave the best fronts after 50,000 evaluations. These parameters are the population size, the number of generations, the number of offspring produced (as a percentage of the population size) and the number of promising solutions on which the BB identification is based (as a percentage of the population size). Due to the time required to run each optimisation this sensitivity analysis is necessarily limited: however, the parameters used (see Table 1) are considered to give a good indication of the relative convergence speeds of the four algorithms.

Each algorithm was run 10 times with the identified parameters. In order to give a comparison of the number of fitness evaluations required for convergence, each algorithm was stopped after 200,000 evaluations, and the results after this many evaluations compared. Table 2 gives the minimum, mean and maximum S metric values for each of the algorithms. Based on this metric UMDA and CSM both appear to perform better than NSGA-II while hBOA performs worse. However, NSGA-II offers far better consistency in terms of the variation in areas enclosed by its fronts. Table 3 gives the minimum, mean and maximum C metric values for each pair of algorithms. Considering the diagonal entries first it is clear that UMDA and NSGA-II offer a much more consistent set of fronts than hBOA and CSM. The minimum values of zero for the latter two show that a front given using one random seed can completely dominate a front given by the same algorithm using a different random seed. The non-diagonal entries show a broadly similar performance between NSGA-II and UMDA. hBOA is seen to perform poorly against NSGA-II and UMDA, and not much better against CSM. The high maximum scores for the CSM demonstrate that, with the right random seed, this algorithm can provide a good front, though the low minimum and mean scores suggest that the fronts are more often poor.

Table 1 | Algorithm parameters for the New York tunnels problem

Optimisation algorithm	Population size	Number of generations	Offspring (%)	Parents (%)
NSGA-II	200	1,000	100	–
UMDA	1,000	200	100	100
hBOA	2,000	100	100	100
CSM	1,000	200	100	100

The best front in each case (as determined using the C metric values) is presented in Figure 2 and the worst front is presented in Figure 3. In the best case the algorithms UMDA and CSM agree well with NSGA-II while hBOA is seen to perform as well at specific points along the front but poorly elsewhere. In the worst case it is clear that all three building block identification methods fail to give good coverage of the front. The poor performance of these algorithms for some random seeds is thought to be caused by their concentrating on BBs which are important for solutions in some parts of the front but not for others.

For the case of hBOA, Khan *et al.* (1999) address this problem by introducing clustering. At each generation the parent population is split into k clusters (different regions in the solution space) and a probabilistic model is built for each cluster. The new population is produced by sampling an equal number of new solutions from each of the k models. This approach is able to force such algorithms to concentrate on a number of points along the front simultaneously: however, it also has the drawback of increasing the computational complexity of the algorithm and the population size required in order to identify BBs.

The average run time was 26 min for NSGA-II, 54 min for UMDA, 59 min for CSM and 223 min for hBOA. It is therefore clear that, as well as offering less consistent results, the three BB identification algorithms are more time-consuming.

Table 4 shows the costs and total pressure deficits for a selection of designs given by each algorithm. In each case the solution with the lowest cost, the solution with the median cost and the solution with the greatest cost are selected from the best front. Only NSGA-II and UMDA identify the minimal cost solution (no pipe duplication at all). While this solution is trivial, the failure of hBOA and CSM to find it demonstrates the poor front coverage offered by these methods. Although all four algorithms find solutions with zero pressure deficits only NSGA-II identifies the best known zero deficit solution with cost \$38.64 million. UMDA, hBOA and CSM identify progressively more expensive solutions, with CSM giving a cost nearly \$8 million greater than NSGA-II.

Table 2 | S metric values (in millions) for the New York tunnels problem

Optimisation algorithm	Minimum	Mean	Maximum
NSGA-II	3,087	3,119	3,141
UMDA	1,559	2,580	3,122
hBOA	3,301	4,265	5,388
CSM	1,885	3,036	3,721

Table 3 | C metric values for the New York tunnels problem (minimum, mean and maximum)

		F_2			
		NSGA-II	UMDA	hBOA	CSM
F_1	NSGA-II	0.495	0.346	0.958	0.404
		0.764	0.613	0.998	0.849
		1	1	1	1
	UMDA	0.394	0.500	0.718	0.598
		0.672	0.793	0.955	0.866
		0.902	1	1	1
	hBOA	0	0	0	0
		0.002	0.003	0.528	0.085
		0.030	0.064	1	0.550
	CSM	0	0	0.225	0
		0.310	0.359	0.896	0.588
		0.772	1	1	1

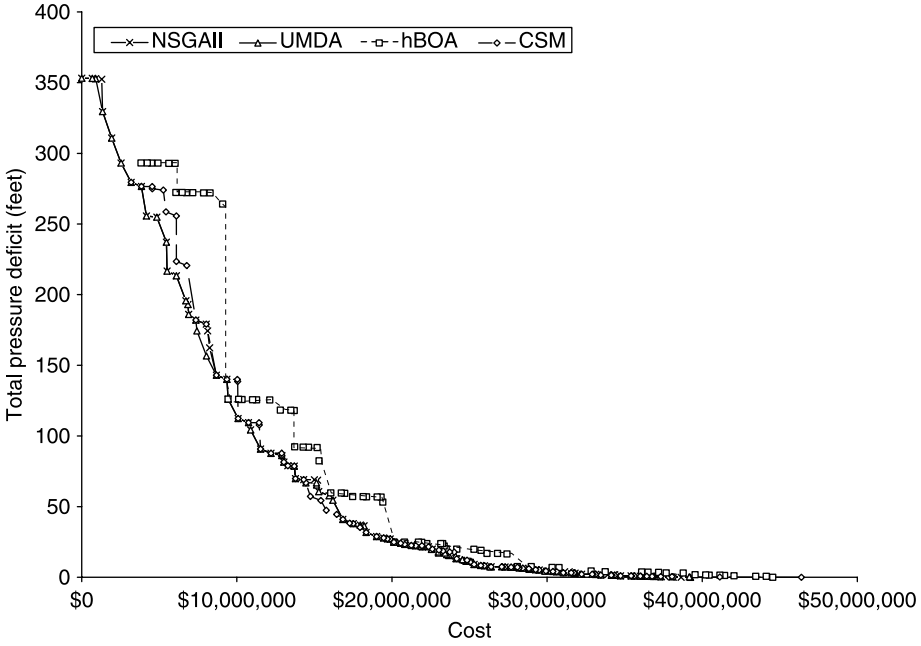

Figure 2 | Best fronts for the New York tunnels problem.

The Anytown network

The Anytown water distribution system (see Figure 4) was introduced by Walski *et al.* (1987) as an example of a more challenging WDS rehabilitation problem. The network consists of a city region in which the existing pipes (thick solid lines) are difficult to access so that cleaning or duplication work is more expensive, and a residential region in which pipes (thin solid lines) are easier to access, and therefore cheaper to duplicate or clean. As part of the

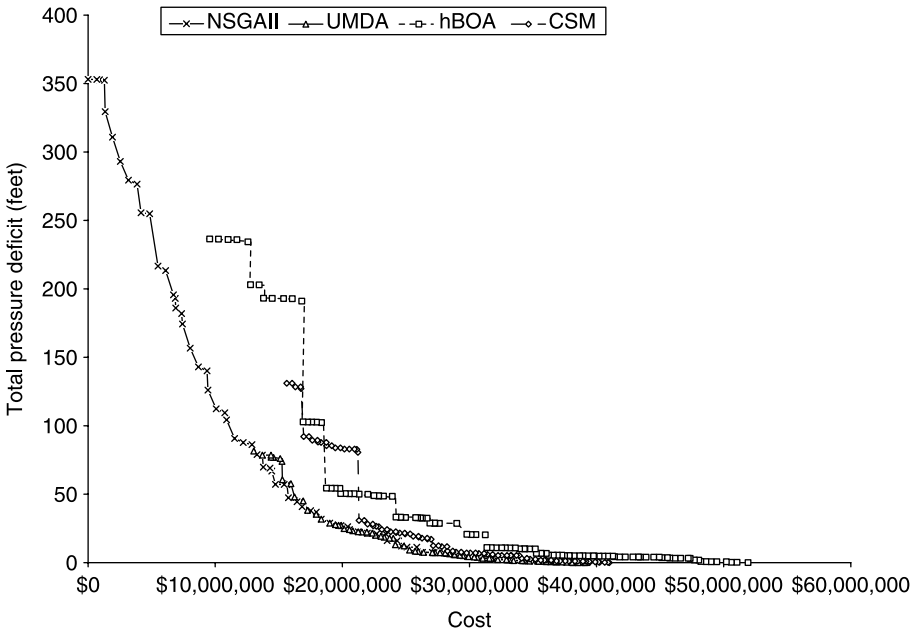

Figure 3 | Worst fronts for the New York tunnels problem.

Table 4 | Costs (thousands) and total pressure deficits of selected designs for the New York tunnels problem

	Minimum cost		Median cost		Maximum cost	
	Cost	Pres. deficit	Cost	Pres. deficit	Cost	Pres. deficit
NSGA-II	\$0	353.14 ft (107.64 m)	\$20 129	24.87 ft (7.58 m)	\$38 644	0 ft (0 m)
UMDA	\$0	353.14 ft (107.64 m)	\$22 541	19.80 ft (6.04 m)	\$39 246	0 ft (0 m)
hBOA	\$3822	293.10 ft (89.34 m)	\$19 415	53.18 ft (16.21 m)	\$44 555	0 ft (0 m)
CSM	\$3183	279.47 ft (85.18 m)	\$22 599	19.97 ft (6.09 m)	\$46 401	0 ft (0 m)

rehabilitation a new extension is planned to the north of the city (thin dashed lines). The network has two existing tanks and water is pumped into the system from a nearby treatment works by three parallel pumps. Each existing pipe in the network may be left as it is, cleaned and lined, or duplicated. New pipes can be chosen from 10 possible diameters. Any node which is not the site of an existing tank is considered as a potential site for a new tank. Each tank has an emergency volume and a normal operating volume. The cost of a new tank is a function of its volume. Up to two new tanks may be added.

In addition to the rehabilitation of the network the number of pumps operating during each hour of a typical day is to be decided. The total design cost is the combined cost of network rehabilitation and pump operation over 20 years at 12% interest. The system should be capable of satisfactorily operating under 5 different load conditions; average day flow, instantaneous peak flow and three

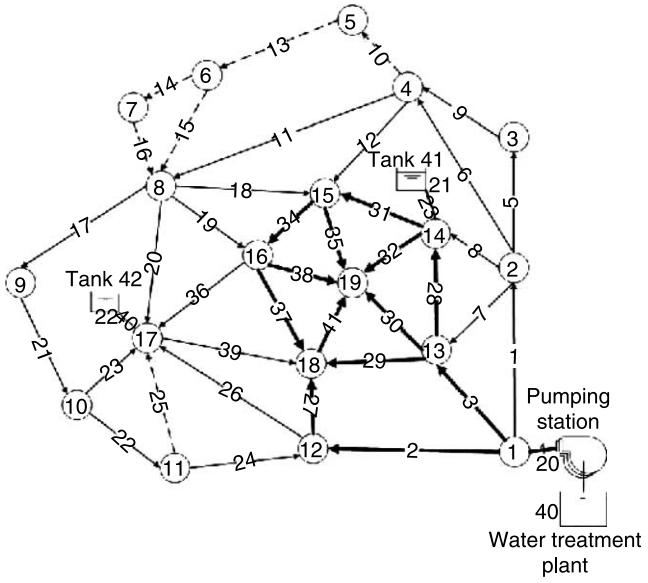

Figure 4 | The anytown network.

fire flows. Under each condition all tanks start at their minimum daily operating levels. Under average daily flow each tank should fill and empty over its normal operating range without going into the emergency volume and a minimum pressure of at least 40 psi (28 m) should be provided at all nodes. The system must provide, instantaneously, a minimum pressure of 40 psi (28 m) at each node when tanks start at their minimum working level and demands are 1.8 times the average daily demand. Under each of the 3 fire conditions the system must provide minimum pressures of 20 psi (14 m) for 2 h. Only the fire flows need be provided at those nodes under fire conditions and demands of 1.3 times the average daily demand must be satisfied at all other nodes. The emergency tank volumes may be used under each of the fire conditions.

The cost (in dollars) of a design for the Anytown network is a combination of pipe laying and cleaning costs, tank installation costs and the net present value of pumping costs over a 20 year operating period (assuming 12% interest). For full details of this cost function see Walski *et al.* (1987). The need to optimise pumping schedules, to design new tanks that fill and empty over average daily flows and to allow for emergency flows combine to make it difficult to choose between one solution and another. For example, one solution may satisfy all pressure requirements under average daily flows, but the end-of-day tank levels may differ from the start-of-day levels. Another solution may have tanks which end the day with the same water levels as they start the day, but may fail to meet pressure requirements under instantaneous peak flows. The main challenge of the Anytown network is to derive suitable measures for the quality of solutions. Since this paper concentrates on comparing the evolutionary algorithms used to optimise the system rather than the measure of quality itself a simple shortfall function is used as the

second objective:

$$S = w\text{NPS} + (1 - w)\text{TLD} \quad (7)$$

where w is a weighting, NPS is the total nodal pressure shortfall across all nodes under all conditions and TLD is the total tank level discrepancy across all tanks during average daily flows. This is calculated as the sum of the difference between intended and observed minimum levels, the difference between intended and observed maximum levels, and the difference between start-of-day and end-of-day levels.

While this shortfall function has the benefit of simplicity it is worth noting that the weighting is arbitrary and so the optimisation can be biased toward solutions that perform well in terms of NPS and poorly in terms of TLD, or vice versa. For this reason a number of more refined measures of quality have been developed including the benefit function of Walters *et al.* (1999) and the reliability (resilience) index of Farmani *et al.* (2005*b*). More recently a three-objective cost vs. reliability vs. water quality optimisation has been proposed by Farmani *et al.* (2006).

Solutions are encoded as 276-bit binary strings. A weight of $w = 0.25$ is used in the shortfall function (thus the optimisation is encouraged to look more favourably on solutions with low TLD than those with low NPS). As for the New York tunnels example a limited sensitivity analysis was performed to find the set of parameters which offered the best convergence after 50,000 evaluations over a selection of runs. The parameters used are given in Table 5. Each algorithm was run five times with the identified parameters. In order to keep CPU times to a manageable level the algorithms were stopped after 200,000 fitness evaluations (where each evaluation involves 5 hydraulic simulations) and the convergence up to that point is compared here.

Table 5 | Algorithm parameters for the Anytown network

Optimisation algorithm	Population size	Number of generations	Offspring (%)	Parents (%)
NSGA-II	200	1,000	100	–
UMDA	1,000	200	100	100
hBOA	2,000	100	100	100
CSM	1,000	200	100	100

Table 6 | *S* metric values (in millions) for the Anytown network

Optimisation algorithm	Minimum	Mean	Maximum
NSGA-II	38,691	42,605	45,643
UMDA	11,198	24,927	38,243
hBOA	12,068	22,561	34,086
CSM	5,936	26,627	47,374

Table 6 gives the minimum, mean and maximum S metric values for each algorithm. The difference between minimum and maximum values is significantly smaller for NSGA-II than for the three BB identification algorithms, suggesting that this algorithm is less sensitive to changes in the initial population. The S metric values for the 3 BB identification algorithms are generally smaller, though this alone does not indicate whether these fronts generally dominate those of NSGA-II or whether they cover a smaller range of solutions. Table 7 shows the minimum, mean and maximum C metric values for each pair of algorithms. The diagonal entries suggest that NSGA-II and hBOA offer greater consistency of coverage than UMDA and CMS. The first row shows that NSGA-II rarely produces solutions which dominate those of the BB identification methodologies. While the fronts given by CSM cover a significant portion of the solutions found by NSGA-II, they cover a fairly small percentage of solutions found by UMDA and hBOA.

Table 7 | *C* metric values for the Anytown network (minimum, mean and maximum)

		F_2			
		NSGA-II	UMDA	hBOA	CSM
F_1	NSGA-II	0.212	0	0	0
		0.572	0.049	0.046	0.121
		1	0.140	0.180	0.313
	UMDA	0.348	0.022	0.388	0.450
		0.482	0.583	0.487	0.656
		0.622	1	0.622	0.837
	hBOA	0.485	0.148	0.266	0.433
		0.697	0.207	0.592	0.801
		0.927	0.238	1	0.997
	CSM	0.409	0	0	0.003
		0.589	0.063	0.082	0.531
		0.829	0.208	0.328	1

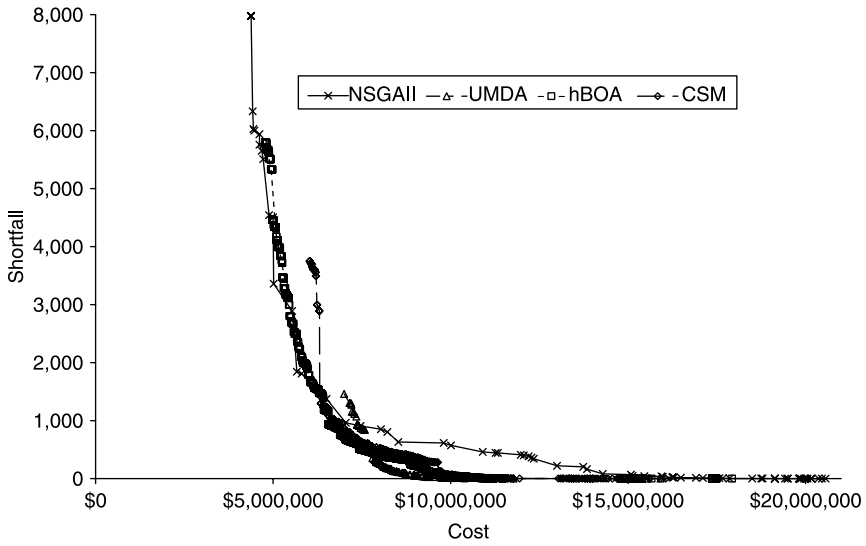

Figure 5 | Best fronts for the anytown network.

The best front for each algorithm is presented in Figure 5. As before, it can be seen that NSGA-II offers the best range of solutions between high shortfall–low cost designs and high cost–low shortfall designs. The three BB identification methods concentrate on the low shortfall end of this range. However, in this area, the BB identification methods offer better results. Figure 6 shows the worst front for each algorithm. Again, it is clear that the probabilistic methods are converging better for different parts of the front.

Table 8 gives three solutions from the best fronts produced by each of the algorithms (the last row will be discussed in the next paragraph). The lowest cost solutions vary considerably both in terms of cost, ranging from just over $4 million in the case of NSGA-II to $7 million in the case of UMDA, and in terms of shortfall, ranging from 7,980 for NSGA-II to only 1,460 for UMDA. These shortfalls are a weighting of the NPS (measured in psi in the original formulation of the problem) and the TLD (measured in feet in the original formulation). These measures are also

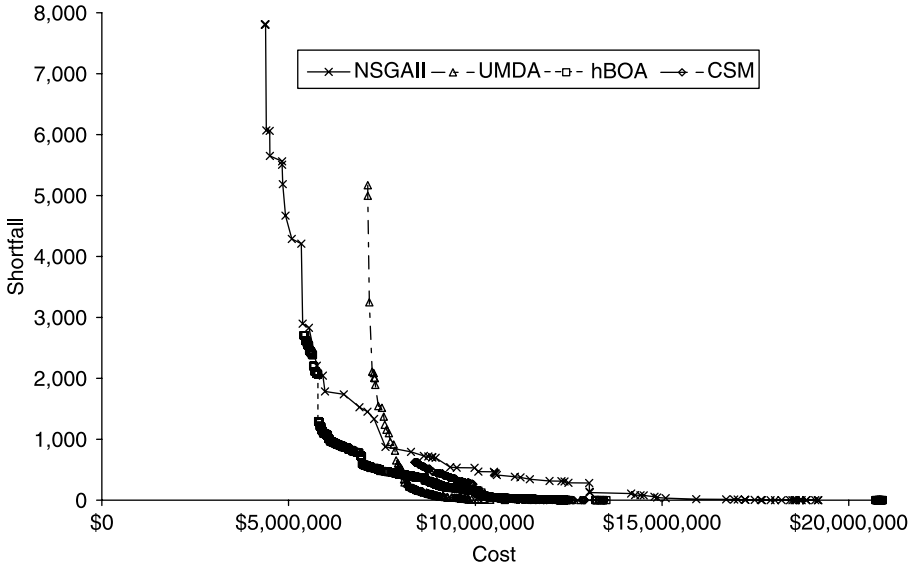

Figure 6 | Worst fronts for the anytown network.

Table 8 | Costs for selected solutions to the Anytown problem

		Cost	Shortfall	NPS (psi)	TLD (ft)
NSGA-II	Minimum cost	$4 374 658.14	7980.35	31651.38	90
	Median cost	$14 284 317.21	79.79	53.49	88.56
	Maximum cost	$20 557 354.48	<0.01	0	<0.01
UMDA	Minimum cost	$7 001 005.02	1459.95	5434.81	135.00
	Median cost	$8 372 609.86	139.51	204.02	118.00
	Maximum cost	$15 911 801.04	2.04	2.90	1.75
hBOA	Minimum cost	$4 775 135.40	5798.46	22 923.84	90
	Median cost	$7 912 085.89	425.67	1582.68	40.00
	Maximum cost	$17 925 418.71	0.01	0	0.02
CSM	Minimum cost	$6 029 639.10	3752.36	14 739.40	90.01
	Median cost	$8 710 608.05	421.80	468.17	73.00
	Maximum cost	$16 054 845.15	<0.01	0	<0.01
NSGA-II (long)	Minimum cost	$4 364 427.14	7808.21	30 962.84	90.01
	Median cost	$11 420 176.80	370.11	1333.23	49.07
	Maximum cost	$19 370 332.00	<0.01	0	<0.01

provided in Table 8. As can be seen the lowest cost solutions all perform badly in both of these measures. The maximum cost (least shortfall) solution provided by UMDA has a NPS of 2.9 psi (2 m) and a TLD of 1.75 ft (0.5 m). The rest of the algorithms produce near-zero shortfalls. In particular CSM gives a comparable shortfall to NSGA-II but at a cost saving of $4.5 million.

For this larger system a difference is apparent in the time each algorithm takes to run. NSGA-II took on average 459 min, UMDA took 515 min, CSM took 848 min and hBOA took 2319 min. These increased run times for the BB identification methods are due to the time spent modelling the interaction between variables. A single run of NSGA-II was made with the same parameters as before except that the number of generations was doubled to 2,000 to give roughly the same run time as CSM (893 min). The minimum, median and maximum cost solutions found by this run are presented in the last row of Table 8. While the least shortfall solution of NSGA-II is improved with the increased run length, it is still more expensive than those given by the BB identification methods.

A real-world system

The network shown in Figure 7 is a real-world WDS consisting of 535 demand nodes fed by a single reservoir. The system is connected by 632 pipes. The cost in Euros (€) of a design is the sum over all pipes of the pipe diameter multiplied by its length. Olsson *et al.* (2007*b*) perform a least cost optimisation of this system using hBOA, where the maximum head deficit across all demand nodes was multiplied by a penalty and added to the cost. With the 20 potential pipe diameters available the solution space for this system contains approximately 1.78×10^{822} solutions. With this many potential solutions, hBOA was found to perform poorly because the computational complexity of the algorithm prevented a suitably large population size

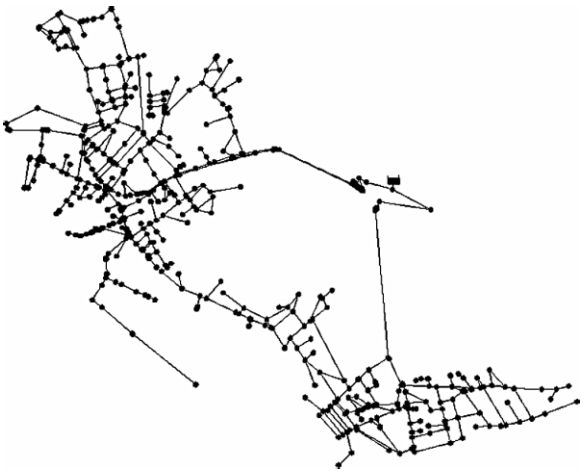

Figure 7 | A real-world system.

Table 9 | Algorithm parameters for the real-world system

Optimisation algorithm	Population size	Number of generations	Offspring (%)	Parents (%)
NSGA-II	200	1,250	100	–
UMDA	1,000	250	100	100
hBOA	100	1,000	100	100
CSM	2,500	100	100	100

from being used. By restricting the number of pipe diameter options to 8 the solution space was reduced to approximately 5.7×10^{570}. This allowed larger populations to be used within hBOA, which in turn allowed the algorithm to more accurately estimate the structure of the problem. In this case hBOA was found to offer noticeable improvement over solutions from the literature (Keedwell & Khu 2006).

Here the restricted set of pipe diameters is used and a multi-objective optimisation is performed to minimise both the cost and the maximum nodal head deficit. Each solution is coded as an 1896-bit binary string. Any infeasible solutions (solutions incapable of satisfying demand with any positive pressure) are penalised with a cost of €10 million. A limited sensitivity analysis is performed and the parameters shown in Table 9 selected. It is noted that, despite the use of a reduced set of pipe diameters, the population size used with hBOA is restricted by the memory requirements of building Bayesian networks for large datasets covering a large number of variables.

Each algorithm was run five times and stopped after 250,000 fitness evaluations with the exception of hBOA, which was stopped after only 100,000 evaluations due to the long computation times taken by this algorithm. Table 10 shows the *S* metric values for the algorithms. NSGA-II, UMDA and CSM all give relatively consistent *S* metric values compared to hBOA, which produces fronts enclosing areas ranging from 384 million up to 19,831 million. Over the five runs hBOA produces fronts ranging in size from 74 solutions down to only 2. This lack of consistency may in part be due to the reduced number of fitness evaluations performed by hBOA: however, it is thought that the main cause is the limited population size leading hBOA to evolve toward a specific part of the Pareto optimal front with a high dependence on the initial population (i.e. random seed).

Table 11 gives the *C* metric values for the four algorithms. All of the diagonal entries show a minimum metric value of 0, meaning that for some random seeds the algorithms produce fronts which are entirely covered by the same algorithm using a different random seed. The non-diagonal entries show that hBOA does not produce any solutions which cover solutions provided by the other algorithms. All of the solutions offered by NSGA-II cover solutions given by hBOA but do not cover solutions given by UMDA or CSM. UMDA and CSM fail to cover some

Table 10 | *S* metric values (in millions) for the real-world system

Optimisation algorithm	Minimum	Mean	Maximum
NSGA-II	230	275	341
UMDA	10	38	85
hBOA	384	4,303	19,831
CSM	36	81	121

Table 11 | *C* metric values for the real-world system (minimum, mean and maximum)

		F_2			
		NSGA-II	UMDA	hBOA	CSM
F_1	NSGA-II	0	0	1	0
		0.592	0	1	0
		1	0	1	0
	UMDA	0.758	0	1	0.316
		0.941	0.565	1	0.921
		1	1	1	1
	hBOA	0	0	0	0
		0	0	0.533	0
		0	0	1	0
	CSM	0.763	0	1	0
		0.970	0.011	1	0.594
		1	0.218	1	1

of the solutions given by NSAG-II. In combination with the fact that NSGA-II does not cover any solutions from these algorithms this suggests that UMDA and CSM offer worse coverage of the front while giving better solutions for the parts of the front they do cover.

The best front from each algorithm is presented in Figure 8. hBOA is seen to give only two solutions, both with significantly worse objective values than solutions offered by the other algorithms. The restriction on population size due to computer memory constraints limits the amount of data hBOA can use to build Bayesian networks, and thus the ability to use BBs to drive the evolutionary process. For the remaining algorithms any increase in the quality of solutions is seen to be matched with a decrease in the range of solutions covered. Figure 9 shows the worst fronts for each algorithm. Here it is seen that, while hBOA offers a more diverse set of solutions, the quality of those solutions is decreased. For the other algorithms the same pattern of quality solutions being found at the expense of diversity among solutions is seen.

Table 12 shows the lowest cost, median cost and highest cost solutions for each of the algorithms taken from the fronts identified as best based on the *C* metric values. The NSGA-II, UMDA and CSM algorithms find solutions which fully satisfy the nodal pressure requirements of the system. NSGA-II does this with a cost of €8.8 million. CSM reduces this cost by a quarter while UMDA halves the cost. It is noticeable that even the minimum cost solution offered by NSGA-II is more expensive than the maximum cost solutions offered by UMDA and CSM.

For this large system NSGA-II took, on average, 437 min to run, UMDA took 695 min, CSM took 5,087 min and hBOA took 6,039 min (for less than half the fitness evaluations). An additional run of NSGA-II is performed with the same population size for 15,000 generations, giving 12 times the number of fitness evaluations. This run took 4,874 min. The front for this run is included in Figure 8. It can be seen that, although the additional generations offer noticeable improvement in solutions, NSGA-II still does not find solutions as good as

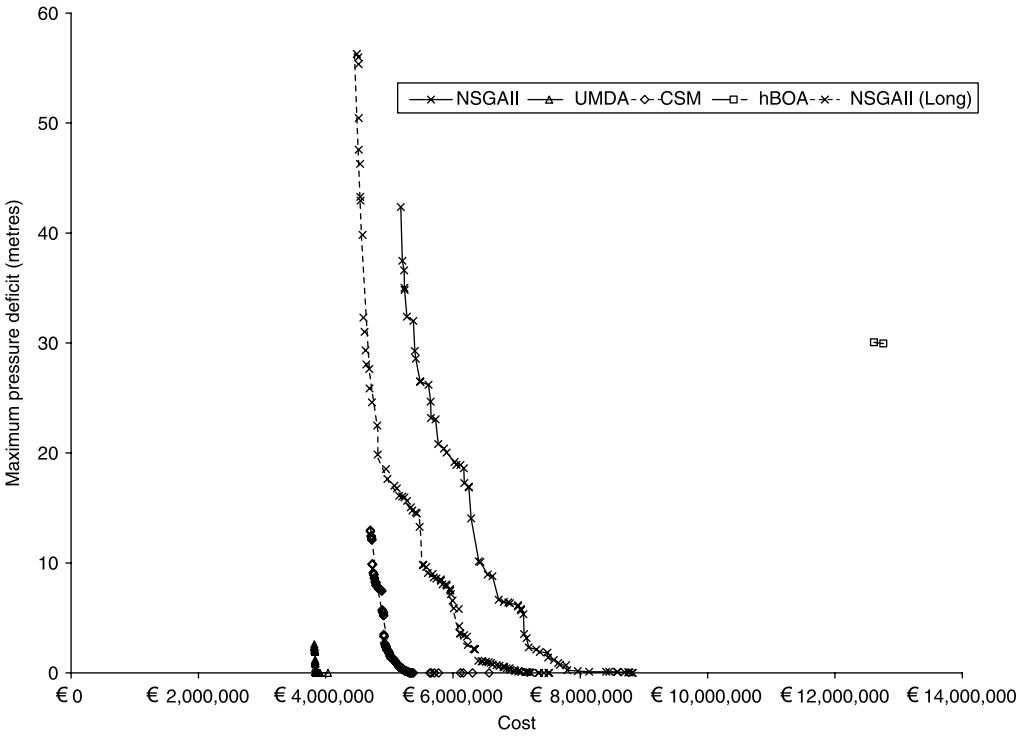

Figure 8 | Best fronts for the real-world system.

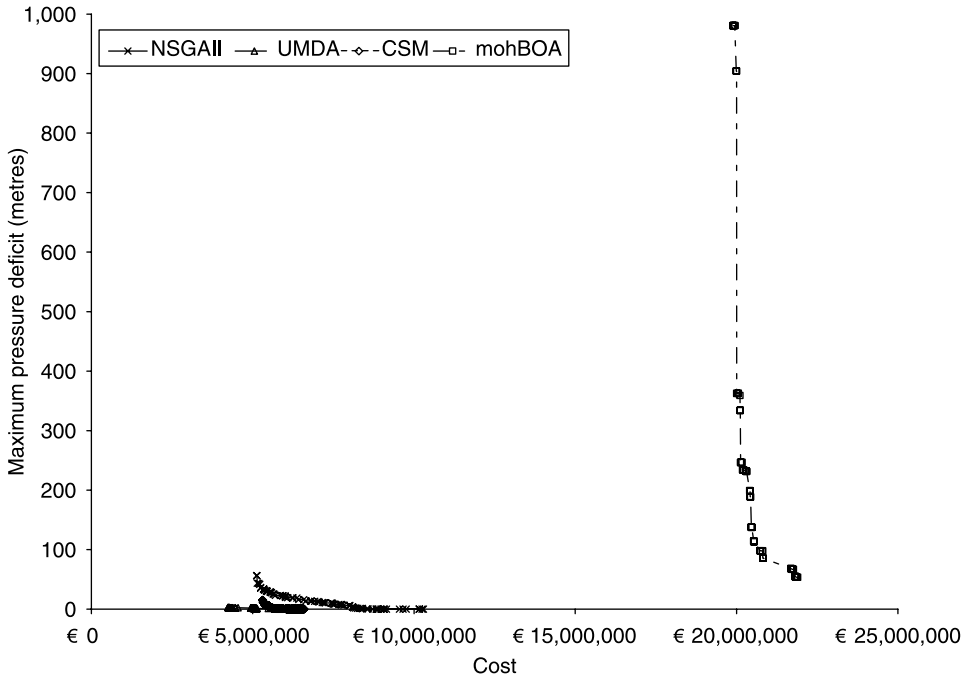

Figure 9 | Worst fronts for the real-world system.

Table 12 | Costs (thousands) and maximum nodal pressure deficits of selected designs for the real-world system

	Minimum cost		Median cost		Maximum cost	
	Cost	Pres. deficit	Cost	Pres. deficit	Cost	Pres. deficit
NSGA-II	€5180	42.37 m	€6799	6.46 m	€8822	0 m
UMDA	€3821	2.6 m	€3848	0.06 m	€4037	0 m
hBOA	€12 612	30.06 m	–	–	€12 759	29.95 m
CSM	€4702	12.98 m	€5018	1.4 m	€6569	0 m

those found by UMDA or CSM. This supports the conclusion that the use of building blocks to guide the search for optimal solutions can significantly improve the scalability of evolutionary algorithms to large problems.

CONCLUSIONS

In this paper three optimisation algorithms which identify and preserve building blocks by employing probabilistic analysis to link decision variables were compared to NSGA-II for the multi-objective optimisation of water distribution systems. For a number of mathematical test functions, BB identification has been seen to significantly enhance the efficiency of genetic algorithms (Pelikan 2002; Aporntewan & Chongstitvatana 2004) and this improvement has recently been shown to extend to single-objective WDS design and rehabilitation (Olsson *et al.* 2007*a*,*b*).

For the case of multi-objective WDS optimisation it is clear that the complexity of these problems has a strong effect on the ability of the probabilistic methods to provide good Pareto fronts. This conclusion is supported by Sastry *et al.* (2005), who investigate the scalability of multi-objective PMBGAs and conclude that a large number of Pareto optimal solutions can overwhelm the niching methods (the methods which preserve alternative solutions) and cause significant problems in maintaining a good coverage of the Pareto optimal front. By allowing BBs to

become disrupted and employing mutation NSGA-II is able to find a good spread of solutions. By modelling the distribution of variables in a population CSM and hBOA are able to guide the evolutionary process toward better solutions, but do so at the expense of good coverage. By ignoring interactions in this model UMDA guides the process toward a specific part of the front, and can therefore give better solutions than CSM and hBOA. However, this improvement comes at the expense of further loss of front coverage. Methods such as clustering (Khan *et al.* 1999) may be able to improve the coverage of probabilistic BB identification methodologies, though when Pareto optimal fronts contain a large number of solutions it is likely that a large number of clusters would be needed and that computation time would therefore be increased significantly.

Although the loss of coverage may make BB identification algorithms unsuitable for small WDS design and rehabilitation problems such as the New York tunnels problem, for larger, real-life WDS problems it is clear that the scalability gained through the identification of BBs can outweigh problems of coverage and allow multi-objective optimisations which offer serious advantages over NSGA-II.

ACKNOWLEDGEMENTS

This research work has been carried out under the EPSRC Platform Grant GR/T26054/01. The authors are also grateful to Prof. David Goldberg and Dr Martin Pelikan for providing the single-objective hBOA software code and permission for its use.

REFERENCES

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

Aporntewan, C. & Chongstitvatana, P. 2004 Chi-square matrix: an approach for building-block identification. In *Proceedings of 9th Asian Computing Science Conference*, Springer, Berlin, Heidelberg, pp. 63–77.

Babayan, A., Kapelan, Z., Savic, D. & Walters, G. 2005 Least-cost design of water distribution networks under demand uncertainty. *J. Water Res. Plan. Manage.* **131** (5), 375–382.

Babayan, A. V., Savic, D. A., Walters, G. A. & Kapelan, Z. 2007 Robust least-cost design of water distribution networks using redundancy and integration based methodologies. *J. Water Res. Plan. Manage.* **133** (1), 67–77.

Dandy, G., Simpson, A. & Murphy, L. 1996 An improved genetic algorithm for pipe network optimization. *Water Resour. Res.* **32** (2), 449–458.

Deb, K. & Goldberg, D. 1991 *Analyzing Deception in Trap Functions*. Technical Report 91009, IlliGAL.

Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. 2002 A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans Evolut. Comput.* **6** (4), 182–197.

Farmani, R., Savic, D. & Walters, G. 2003 Multi-objective optimization of water systems: a comparative study. In *Pumps, Electromechanical Devices and Systems Applied to Urban Water Management* (ed. in E. Cabrera & E. Cabrera, Jr), Taylor & Francis, London, pp. 247–256.

Farmani, R., Savic, D. & Walters, G. 2005*a* Evolutionary multi-objective optimization in water distribution network design. *Eng. Optimiz.* **37**, 167–183.

Farmani, R., Walters, G. & Savic, D. 2005*b* Trade-off between total cost and reliability for anytown water distribution network. *J. Water Res. Plan. Manage.* **131** (3), 161–171.

Farmani, R., Walters, G. & Savic, D. 2006 Evolutionary multi-objective optimization of the design and operation of water distribution network: total cost vs. reliability vs. water quality. *J. Hydroinf.* **8** (3), 165–179.

Halhal, D., Walters, G., Ouazar, D. & Savic, D. 1997 Water network rehabilitation with a structured messy genetic algorithm. *J. Water Res. Plan. Manage.* **123** (3), 137–146.

Hejazi, M. I., Cai, X. & Borah, D. K. 2008 Calibrating a watershed simulation model involving human interference: an application of multi-objective genetic algorithms. *J. Hydroinf.* **10** (1), 97–111.

Kapelan, Z., Savic, D. & Walters, G. 2005 Multiobjective design of water distribution systems under uncertainty. *Wat. Resour. Res.* **41**, W11407.

Keedwell, E. & Khu, S.-T. 2006 Novel cellular automata approach to optimal water distribution network design. *J. Comput. Civil Eng.* **20** (1), 49–56.

Khan, N., Goldberg, D. E. & Pelikan, M. 1999 *Multi-objective Bayesian Optimization Algorithm*. Technical Report 2002009, IlliGAL.

Kim, T., Heo, J.-H., Bae, D.-H. & Kim, J.-H. 2008 Single-reservoir operating rules for a year using multiobjective genetic algorithm. *J. Hydroinf.* **10** (2), 163–179.

Loganathan, G., Greene, J. & Ahn, T. 1995 Design heuristic for globally minimum cost water-distribution systems. *J. Water Res. Plan. Manage.* **121** (2), 182–192.

Morgan, D. & Goulter, I. 1985 Optimal urban water distribution design. *Water Resour. Res.* **21** (5), 642–652.

Mühlenbein, H. 1997 The equation for response to selection and its use for prediction. *Evolut. Comput.* **5** (3), 303–346.

Olsson, R., Kapelan, Z. & Savic, D. 2007*a* The design and rehabilitation of water distribution systems using the hierarchical bayesian optimisation algorithm. In *Proceedings of*

the World Environmental and Water Resources Congress, Tampa, FL, ASCE, p. 497.

Olsson, R., Kapelan, Z. & Savic, D. 2007*b* The hBOA based methodology for the least cost design of large water distribution systems. In *Proceedings of the 9th International Conference on Computing and Control for the Water Industry (CCWI), Leicester, UK*. Taylor and Francis, London, pp. 455–458.

Pelikan, M. 2002 *Bayesian Optimization Algorithm: From Single Level to Hierarchy*. PhD Thesis, University of Illinois at Urbana-Champaign (IlliGal report no. 2002023).

Pelikan, M., Goldberg, D. & Lobo, F. 1999 *A Survey of Optimization by Building and Using Probabilistic Model*. Technical Report 99018, IlliGAL.

Prasad, T. D. & Park, N.-S. 2004 Multiobjective genetic algorithms for design of water distribution networks. *J. Water Res. Plan. Manage.* **130** (1), 73–82.

Rossman, L. A. 2000 *EPANET 2 Users Manual*. US Environmental Protection Agency, Washington, DC.

Sastry, K., Pelikan, M. & Goldberg, D. 2005 Limits of scalability of multiobjective estimation of distribution algorithms. In: *Proceedings of the Congress on Evolutionary Computation*. Vol. 3, IEEE, pp. 2217–2224.

Savic, D. A. & Walters, G. A. 1997 Genetic algorithms for least-cost design of water distribution networks. *J. Water Res. Plan. Manage.* **123** (2), 67–77.

Schaake, J. & Lai, D. 1969 *Linear Programming and Dynamic Programming Applications to Water Distribution Network Design*. Report 116. Hydrodynamics Laboratory, Department of Civil Engineering, MIT, Cambridge, MA.

Simpson, A., Dandy, G. & Murphy, L. 1994 Genetic algorithms compared to other techniques for pipe optimization. *J. Water Res. Plan. Manage.* **120** (4), 423–443.

Vairavamoorthy, K. & Ali, M. 2000 Optimal design of water distribution systems using genetic algorithms. *Comput.-Aid. Civil Infrastruct. Eng.* **15**, 374–382.

Walski, T., Brill, E., Gessler, J., Goulter, I., Jeppson, R., Lansey, K., Lee, H., Liebman, J., Mays, L., Morgan, D. & Ormsbee, L. 1987 Battle of the network models: epilogue. *J. Water Res. Plan. Manage.* **113** (2), 191–203.

Walters, G., Halhal, D., Savic, D. & Ouazar, D. 1999 Improved design of "anytown" distribution network using structured messy genetic algorithms. *Urban Water* **1** (1), 23–38.

Zitzler, E. & Thiele, L. 1999 Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE Trans. Evolut. Comput.* **3** (4), 257–271.

First received 7 May 2008; accepted in revised form 27 September 2008.