



HAL
open science

Demand Estimation In Water Distribution Systems: Solving Underdetermined Problems Using Genetic Algorithms

Nhu Cuong Do, Angus Simpson, Jochen Deuerlein, Olivier Piller

► **To cite this version:**

Nhu Cuong Do, Angus Simpson, Jochen Deuerlein, Olivier Piller. Demand Estimation In Water Distribution Systems: Solving Underdetermined Problems Using Genetic Algorithms. *Procedia Engineering*, 2017, 186 (2017), pp.193-201. 10.1016/j.proeng.2017.03.227 . hal-01550013

HAL Id: hal-01550013

<https://hal.science/hal-01550013>

Submitted on 29 Jun 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



XVIII International Conference on Water Distribution Systems Analysis, WDSA2016

Demand Estimation In Water Distribution Systems: Solving Underdetermined Problems Using Genetic Algorithms

Nhu Do^{a*}, Angus Simpson^b, Jochen Deuerlein^c, Olivier Piller^d

^aPhD Candidate, The University of Adelaide, Adelaide SA 5000, Australia

^bProfessor, The University of Adelaide, Adelaide SA 5000, Australia

^cSenior Researcher, 3S Consult GmbH, Karlsruhe, Germany

^dSenior Researcher, Irstea UR ETBX, Dept. of Water, Cestas, France

Abstract

Modeling of water distribution systems is fundamental for the design, analysis and operation of any water network. As with all hydraulic models, water demands are one of the most important input components in the model. However, estimation of the demand parameters is usually complicated due to the stochastic behavior of the water consumptions. Several methods have been proposed for estimating water demands. Most of them have been developed based on given frameworks where the number of unknown parameters is assumed to be equal or less than the number of measurements. The outcomes, therefore, rely on this assumption, which can lead to significant approximation errors in real water distribution systems.

The approach proposed in this paper does not require the number of known inputs to be equal to the number of variables. In fact, nodes in the model could each have a different demand pattern. The genetic algorithm approach adopted here shows that the average results of multiple GA runs can estimate the demand patterns at each node. Moreover, the model can also be used to estimate the flow rates and nodal heads at non-measured locations of the water network, although the accuracy of the estimation depends on number, type and location of the measurements. Results are shown and discussed for a literature case study tested for a 24-hour time period.

© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Peer-review under responsibility of the organizing committee of the XVIII International Conference on Water Distribution Systems

* Corresponding author. Tel: +61 414 503 468; Fax: +61 8-8313-4359
Email: nhu.do@adelaide.edu.au

Keywords: Genetic algorithms; optimization; demand estimation; water distribution systems

1. Introduction

Water distribution system infrastructure has been constructed and developed for hundreds of years across the world. Together with the population growth and urbanization, water distribution systems (WDS) have expanded, and have become more complex and more difficult to operate. Modeling of WDS is, therefore, becoming increasingly important due to the need of understanding the behavior of these systems. Various simulation software solutions have been developed and broadly used for the design, analysis and operation of WDS, including EPANET, CWSNET or HydraulCAD. However, a new major challenge to deal with these models is the requirement for consistency between observed data of the real networks and simulated data from simulations models. Estimation of the model parameters, hence, is required so that the model is able to represent the real system.

In WDS, estimation is a process of fitting the outputs from the computer model, usually the pressures and flow rates at particular locations in the network, with the field measurements as well as calculating the parameters of interest [1]. Various model parameters, such as pipe diameters, roughness coefficients, valve resistances, or nodal demands, are required to be estimated. While most of these parameters are time invariant or vary slowly, nodal demands are the only parameter that can cause immediate changes in the model output. Estimation of model parameters during short periods of time using supervisory control and data acquisition (SCADA) systems, therefore, usually focuses on the demand parameter.

In the literature, the estimation of water demand has been studied by numerous researchers based on different methods, for example, extended Kalman filtering [2], tracking state estimation and Kalman filtering [1], Genetic Algorithms [3], and Particle Filtering [4]. However, these models have been developed based on given frameworks where the measurement locations were predetermined and the calibration parameters are grouped to be less than the number of measurements. The outcomes, therefore, rely on these additional assumptions, which can lead to large approximations in real water distribution systems. Only few papers have directly dealt with underdetermined systems such as a proportional demand method [5] and singular value decomposition (SVD) [6], [7].

This paper presents a study by [8] for the estimation of water demands in WDS over a period of 24 hours. A methodology is proposed to find the demand multiplier factors for an underdetermined system where the number of measurements is less than the number of demand parameter variables. The EPANET toolkit is used to solve the system of water network equations while Genetic Algorithms (GAs) are applied to find the best match between known measurements inputs and their estimated values. The mean values of multiple GA runs are suggested to be used as the best estimation of the flow rates and nodal heads as well as the estimation of the nodal demands in a system.

2. Methodology

The proposed model applies an implicit technique for the steady-state hydraulic simulation where the estimation process is formulated as an optimization problem. The objective function is a weighted least squares function in order to minimize the differences between simulated values from the model and their corresponding measured values.

The objective function at each time step is given by:

$$\text{Min}F = \sum_{i=1}^{N_H} w_i (H_i^{\text{Meas}} - H_i^{\text{Sim}})^2 + \sum_{j=1}^{N_Q} w_j (Q_j^{\text{Meas}} - Q_j^{\text{Sim}})^2 \quad (1)$$

where H_i^{Sim} , Q_j^{Sim} are the simulated nodal head and flow rate for the i^{th} node and j^{th} pipe, respectively; H_i^{Meas} , Q_j^{Meas} are the measured head and flow rate at the i^{th} node and j^{th} pipe (in this case these values are known exactly as they have been generated from a forward model run by a hydraulic simulator); N_H , N_Q are the number of head and flow measurement sites in the network and w_i , w_j are the weighting factors applied to different terms to ensure they have similar magnitude.

The decision variables for the optimization problem are the demand multiplier factors $f_{k,t}$, which are used to calculate the nodal demands at each time step:

$$D_{k,t} = D_{0,k} * f_{k,t} \tag{2}$$

where $D_{0,k}$ is base demand at the k^{th} node, which is calculated using quarter/annual billing information for water usage; and $f_{k,t}$ ($k = 1 \dots N_{DM}$) is the demand multiplier factor at the k^{th} node at time step t .

A bounded range of demand factors may apply as:

$$f_k^{min} \leq f_k \leq f_k^{max} \tag{3}$$

where (f_k^{min}, f_k^{max}) are the bounds of decision variables.

The genetic algorithm (GA) estimation model implemented for this research has been written in C-sharp language. The flowchart of the algorithm is shown in Figure 1. An initial population of chromosomes is randomly generated and decoded into corresponding demand multiplier factor values of each chromosome. For each node of the network exactly one of these DMFs is assigned and EPANET is subsequently called to simulate the steady state hydraulics of the system. Simulated flows and heads (Q^{Sim}, H^{Sim}) at the measurement locations are obtained and compared with their actual measured values via the calculation of the objective function of Eq.(1). The inverse of the objective function (in the expression of $1/(F+1)$ to avoid an indeterminate form when F equals to zero) is applied to define the fitness function for each member of the GA population. This is the measure for the quality of each member, and is used to determine its opportunity of survival.

By applying GA selection, crossover and mutation, new generations that inherit features of previous generation are created, and the estimation process is then repeated until the stopping criteria is met.

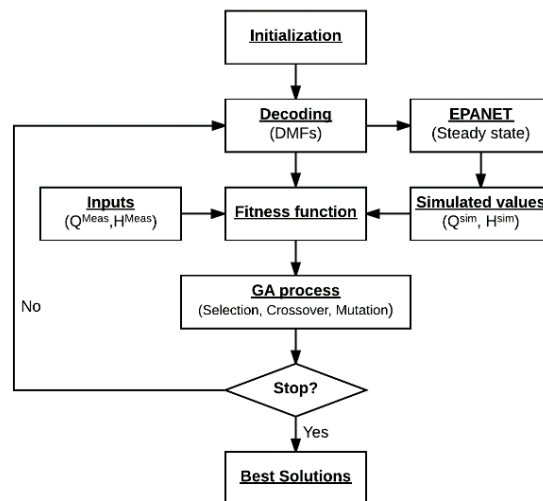


Fig. 1. Flowchart for GA estimation of demand multiplier factors

It should be noted that the proposed estimation model is to be applied to cases where the number of measurements is less than the number of unknown variables. In other words, the estimation of the demand in water distribution systems is mathematically a nonlinear underdetermined problem. A local solution of the problem can be found by local linearization methods such as QR decomposition, SVD or using the Moore-Penrose pseudoinverse matrix in Newton-Raphson method. However, due to the possibility of non-uniqueness of the solutions, the results from mathematical methods may be either far from the actual solution or result in negative demands at some nodes. A single run of the GA model, therefore, might converge to any of the non-unique solutions or be trapped in a local optimal solution where the simulated values cannot perfectly match the measured values at measurement locations. As a result, it appears that a good approximation of the demand multiplier factors estimation problem can only be obtained if multiple runs of the GA model are implemented and then averaged. The following section shows the results of the

multiple runs of the proposed GA model of a WDS that has been studied by the SVD model from [6] for the same estimation problem.

3. Case study

The case study used to evaluate the methodology is shown in Figure 2. The network has 9 nodes, 12 pipes, one tank, one pump and one reservoir. The network topology and all information such as the pipe parameters, length, roughness and pump characteristic can be found from EPANET example [9], namely the Net1 network.

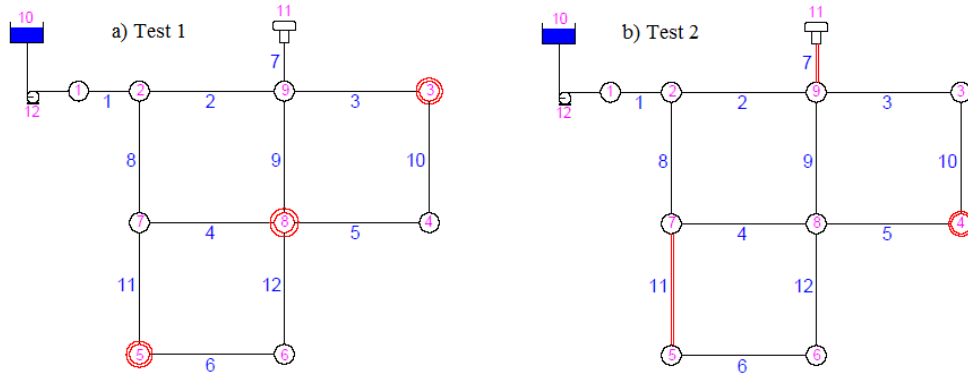


Fig. 2. Study network for the estimation problem.

The GA estimation model was tested for two scenarios of measurement sites according to [6]. In the first test, three pressure sensors are assumed to be located at nodes 3, 5 and 8. The second test assumes two flow meters are set at pipes 7 and 11, respectively and a pressure sensor is placed at node 4.

3.1. Input for GA estimation model

In practice, input data for the estimation process are collected from a supervisory control and data acquisition (SCADA) system. In this research, input data is generated using EPANET toolkit as follows: (1) known demand multiplier factors are assigned to nodal demands; (2) run EPANET to retrieve the corresponding “true” or “known” pipe flow rates and nodal heads; (3) select the flows and heads at the selected locations as input for the GA model. The output flows and heads for selected pipes and nodes based on the simulation of the “true” values are used as the measured values (H^{Meas} , Q^{Meas}) for the estimation process.

3.2. GA operators and model parameters

For the selection of GA operators, tournament selection was chosen because of its better convergence compared to proportionate selection or ranking selection; two-point crossover operator with the relatively high probability of $P_c = 0.8$ and the bitwise mutation with the probability of $P_m = 1/str \approx 0.13$ (str is the length of the string) are chosen, which are suggested by a study from [10].

The variation of the water demands at each node is presented by the range of the water demand multiplier factors. These values, after multiplying with their base demand, are expected to cover all the water usage throughout the day, including the lowest water use (possibly zero demand) and some extreme cases where the water demand is much larger than its average values. The range of decision variables (i.e. the demand multiplier factors), therefore, was selected to be from $f_k^{min} = 0.0$ to $f_k^{max} = 4.0$ which is selected based on typical values of the demand factors reported in [11].

One of the factors that may affect the accuracy of the GA estimation model is the increment ($\Delta\theta$) of the DMFs. The selection of a large increment for the decision variables leads to faster convergence of the GA model although it may result in a coarser approximation of the estimated demands. Alternatively, the GA model may give better results if selecting smaller increment steps. However, the model requires more computational effort to converge due to the larger search space size. In this case study, the increment step of the decision variables was selected as $\Delta\theta = 0.05$, corresponding to a search space size of $81^8 = 1.85 \cdot 10^{15}$ possible solutions for the GA model.

An integer-coding scheme was selected for the GA model. Each decision variable was coded by an integer number, ranging from 0 to 80 based on the choice table of the demand multiplier factors (corresponding to $\Delta\theta = 0.05$). By

using this coding information, the chromosome in the GA process was decoded into a set of demand multiplier factors that are multiplied by the base demand and can be used for the hydraulic simulation process. An example of GA model process for the estimation problem is shown in Figure 3.

Chromosome 1	10	12	34	1	71	5	42	16						
Chromosome 2	35	38	48	14	80	65	58	73						
Mutation	10	12	48	14	80	49	42	16						
	N	N	N	N	N	Y	N	N						
	2	38	34	1	71	65	58	73						
Y	N	N	N	N	N	N	N	N						

Crossover	10	12	48	14	80	5	42	16					
	35	38	34	1	71	65	58	73					
Child 1	10	12	48	14	80	49	42	16					
Decode	0.5	0.6	2.4	0.7	4	2.45	2.1	0.8					
Child 2	2	38	34	1	71	65	58	73					
Decode	0.1	1.9	1.7	0.05	3.55	3.25	2.9	3.65					

Fig. 3. Example of a GA process for the estimation problem

The population size for GA model was selected as $N=100$. Finally, due to the non-uniqueness and the stochastic behavior of the problem, for each GA application, multiple runs with different seeds were implemented. The stopping criteria for each GA run is defined by the number of generations, which equals to $N_{stop} = 1000$ generations.

4. Results and discussions

4.1. Comparison between two scenarios of measurement sites

A series of 100 runs (each initiated with a different random number seed) of the GA model was implemented for each scenario of measurement sites in a steady state simulation. The results are shown in Table 1 and Table 2.

Table 1. Average nodal demands and nodal pressures of the GA model for Test 1 and Test 2 (shown in Figure 2)

Node	Nodal demands (GPM)							Nodal pressure (psi)						
	Actual	Test 1 (100 runs averaged)	Test 2 (100 runs averaged)	$\Delta_{Test 1}$	$\Delta_{Test 2}$	%Error Test 1	%Error Test 2	Actual	Test 1 (100 runs averaged)	Test 2 (100 runs averaged)	$\Delta_{Test 1}$	$\Delta_{Test 2}$	%Error Test 1	%Error Test 2
2	150	260.83	146.18	110.83	3.82	73.89	2.55	119.26	119.66	119.55	0.40	0.29	0.34	0.24
3	100	114.05	135.6	14.05	35.60	14.05	35.60	118.67	<i>118.67</i>	118.08	<i>0.00</i>	0.59	<i>0.00</i>	0.50
4	150	116.93	110.03	33.07	39.97	22.05	26.65	120.74	123.81	<i>120.74</i>	3.07	<i>0.00</i>	2.54	<i>0.00</i>
5	100	83.9	91.35	16.10	8.65	16.10	8.65	115.86	<i>115.86</i>	116.51	<i>0.00</i>	0.65	<i>0.00</i>	0.56
6	100	88.15	112.65	11.85	12.65	11.85	12.65	110.79	113.43	109.39	2.64	1.40	2.38	1.26
7	150	237.3	135.23	87.30	14.77	58.20	9.85	117.66	119.39	118.3	1.73	0.64	1.47	0.54
8	200	130.9	175.1	69.10	24.90	34.55	12.45	118.76	<i>118.76</i>	118.83	<i>0.00</i>	0.07	<i>0.00</i>	0.06
9	150	241.7	162	91.70	12.00	61.13	8.00	117.02	120.06	117.02	3.04	0.00	2.60	0.00

italic - measured locations, Δ - absolute differences between actual values and estimated values

Table 1 presents the average values of nodal demands and nodal pressures from 100 runs of the GA model. For the estimation of the nodal demands, Test 2 (2 pipe flows and 1 nodal pressure are measured) provides better results than Test 1 (3 nodal pressures are measured). The maximum error was found at node 3 for Test 2, of 35.6%, while for Test 1 the maximum error occurred at node 2 with the corresponding percentage of 73.88%. Test 2 also shows the best estimate at node 2 with an estimated nodal demand of 146.18 GPM (the actual value is 150 GPM, corresponding with only 2.55% error). On the other hand, the minimum estimate error of Test 1 was found at node 6 with a proportion of 11.85%. For the estimation of the nodal pressures, the GA model achieved very good results, with the maximum error less than 2.6% for all nodes in the network.

Table 2. Average pipe flow rates of the GA model for Test 1 and Test 2 (shown in Figure 2)

Pipe	Flow rates (GPM)						
	Actual	Test 1 (100 runs averaged)	Test 2 (100 runs averaged)	$\Delta_{Test 1}$	$\Delta_{Test 2}$	% error Test 1	% error Test 2
1	1866.18	1863.60	1864.32	2.58	1.86	0.14	0.10
2	1234.21	1108.63	1244.62	125.58	10.41	10.17	0.84
3	129.34	128.40	153.39	0.94	24.05	0.73	18.59
4	191.16	103.23	202.67	87.93	11.51	46.00	6.02
5	120.66	102.58	128.36	18.08	7.70	14.98	6.38
6	40.81	29.71	55.47	11.10	14.66	27.20	35.92

7	766.18	509.85	766.19	256.33	0.01	33.46	0.00
8	481.97	454.14	473.52	27.83	8.45	5.77	1.75
9	188.70	188.69	166.30	0.01	22.40	0.01	11.87
10	29.34	14.35	51.23	14.99	21.89	51.11	74.61
11	140.81	113.61	140.80	27.20	0.01	19.32	0.00
12	59.19	58.44	63.44	0.75	4.25	1.26	7.19

**italic - measured locations, Δ - absolute differences between actual values and estimated values*

Table 2 presents the average values of pipe flow rates from 100 runs of the GA model. It can be seen from the table that Test 2 again shows better results than for Test 1. In Test 2, the estimate of the flows is relatively accurate for the pipe with average flows (>50GPM) and high flows (>400 GPM). Large estimate errors only occur at pipes with low flows (<50GPM) such as pipe 10 (74.61%) and pipe 6 (35.92%). For Test 1, relatively large estimate errors can be found in all ranges of flow magnitude, for example, pipe 7 (the actual value is 766.18 GPM, the estimate value is 509.85 GPM, corresponding to an error of 33.46%), pipe 4 (the actual value is 191.16 GPM, the estimate value is 103.23 GPM, estimated error of 46%) or pipe 10 (the actual value is 29.34 GPM, the estimate value is 14.35 GPM, estimated error of 51.23%). This result highlights the important role of the locations and types of the measurements in the demand estimation problem.

4.2. Results of GA model for extended period simulation

The GA model is tested in a 24-hour extended period simulation for Test 2 where the measurements are placed at pipes 7, 11 (flow measurements) and node 4 (a pressure measurement), respectively. In order to evaluate the effects of the number of GA runs on the model output, the results of an individual 20 GA runs and 100 GA runs are examined. The estimated flow rates at selected pipes and estimated demands at selected nodes are plotted in Figure 4 and Figure 6. By increasing the number of GA runs from 20 to 100, the model provides slightly better estimates of flows (e.g. flow at pipe 5 – Figure 6), while giving very good approximations of the water demands at all nodes.

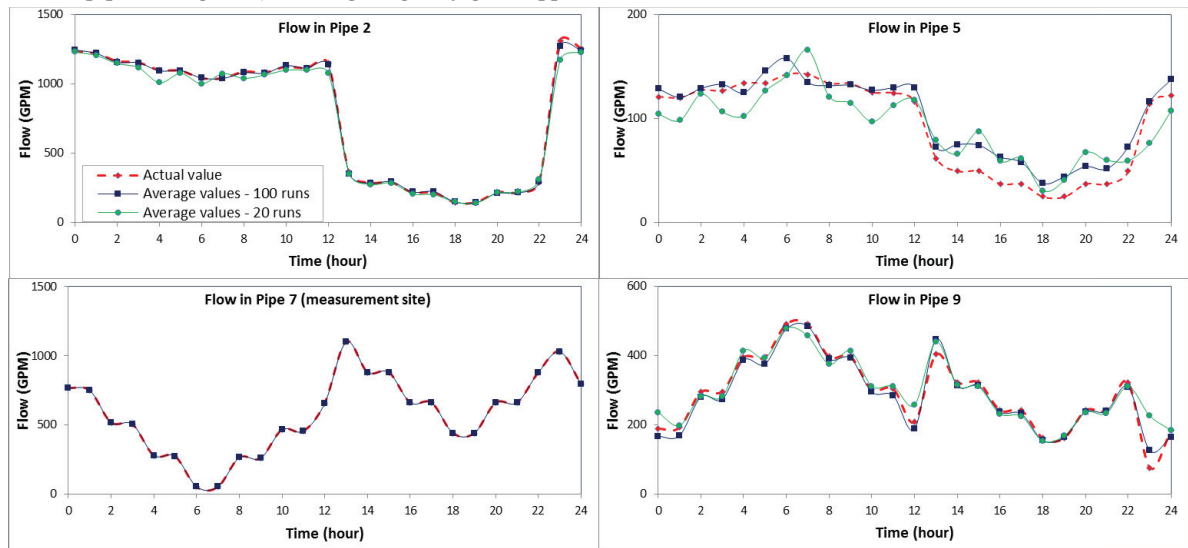


Figure 5 displays the scattergrams and correlation coefficients (R^2) of the average estimated flows of 100 GA runs versus actual flows at all pipes of the network. Similarly, Figure 7 plots the scattergrams and correlation coefficients of the average estimated demands of 100 GA runs versus actual demands at all nodes. It can be seen that the correlations are strong and for both flow estimates and demand estimates, except the flow in pipe 10 ($R = 0.715$) and the demand at node 7 ($R=0.677$). The strong correlation values indicate that the estimation is consistently good over time for the entire network.

Fig. 4. Estimation of flow rates at pipe 2, 5, 7 and 9.

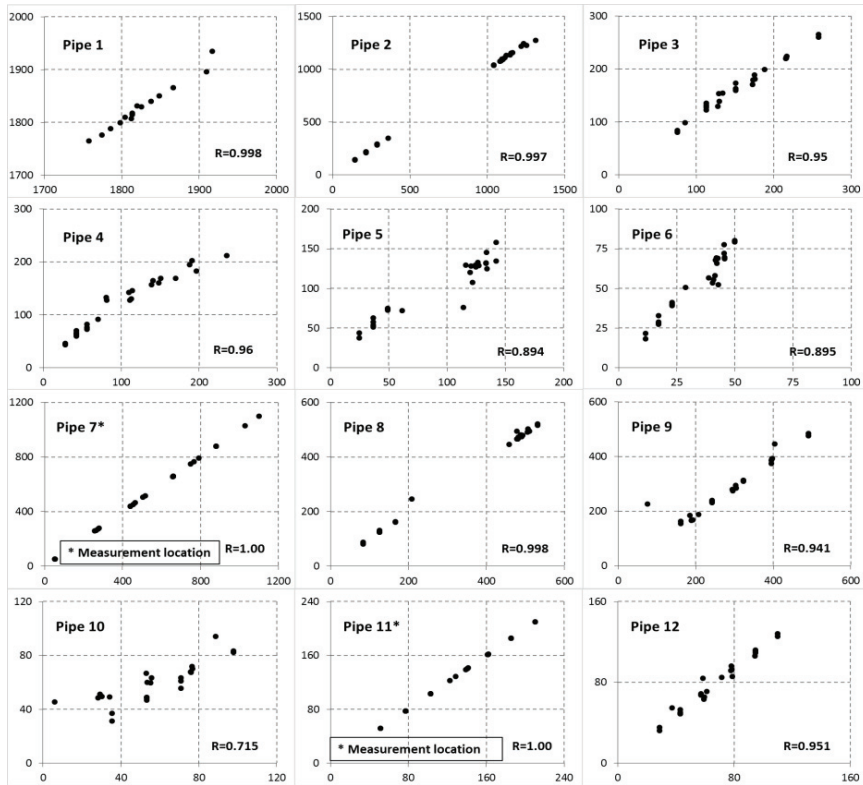


Fig. 5. Scattergrams and correlation coefficients between flow estimates (average of 100 GA runs) and actual flows (units for both axes are GPM)

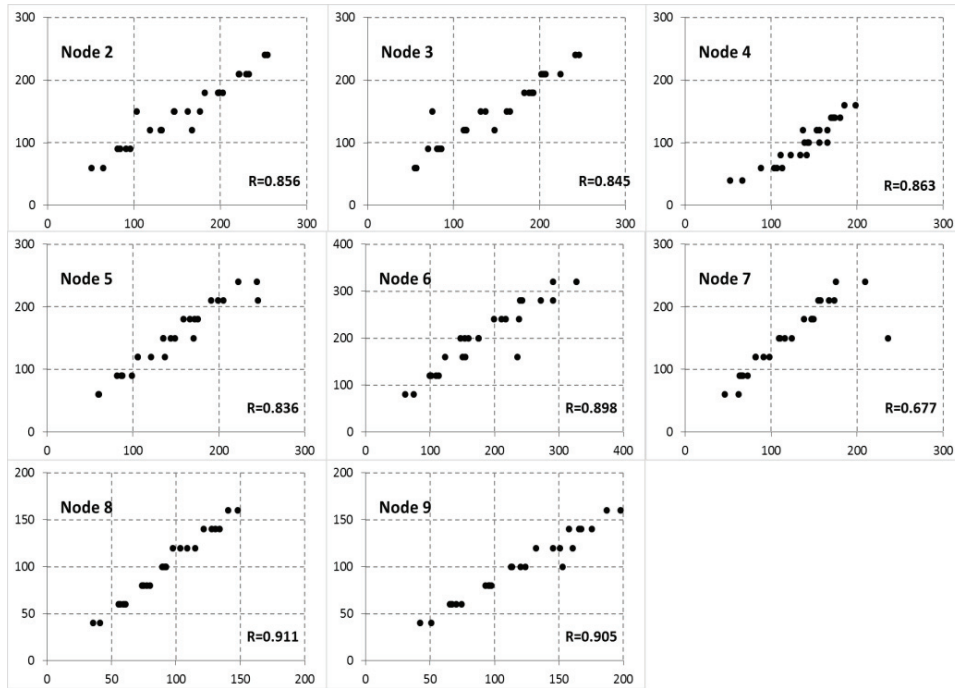
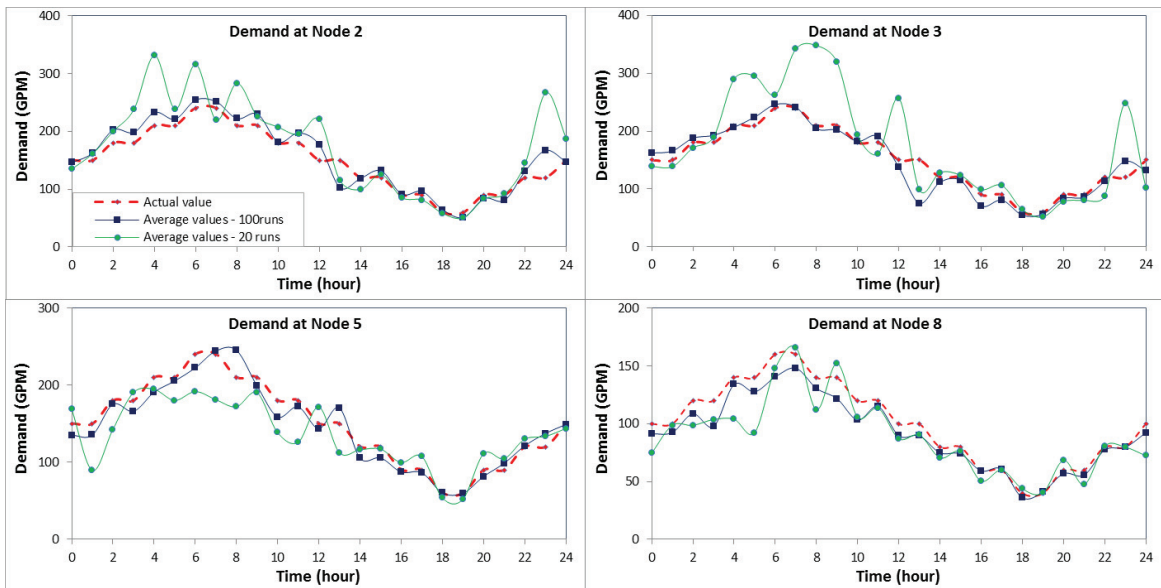


Fig. 6. Estimation of demands at node 2, 3, 5 and 8.

Fig. 7. Scattergrams and correlation coefficients between demand estimates (average of 100 GA runs) and actual demands (units for both axes are GPM)

5. Conclusions

Estimation of water demand in water distribution systems is problematical due to the limited number of



measurement sites. In this paper, a GA model based on integer coding has been introduced for the estimation of the water demand multipliers for underdetermined problems in water distribution systems where the number of measurement sites is less than the number of unknown parameters. The results provided from the case study show that

multiple runs of the GA model can produce relatively good approximation of the state in a water network. Future research efforts will involve finding advanced methods for the estimation of the demand to reduce the computational time. In addition, uncertainty of the model is another consideration given the presence of errors in measurement data. Finally, addressing the problem of leakage in the network is also important in achieving reliable results.

References

- [1] Kang, D. and Lansey, K. , Real-time demand estimation and confidence limit analysis for water distribution systems. *Journal of Hydraulic Engineering*, 2009. 135(10): p. 825-837.
- [2] Shang, F., et al. Real time water demand estimation in water distribution system. in 8th Annual Water Distribution Systems Analysis Symposium 2006, August 27, 2006 - August 30, 2006. 2006. Cincinnati, OH, United States: American Society of Civil Engineers.
- [3] Preis, M. Allen, and A.J. Whittle. On-line hydraulic modeling of a water distribution system in Singapore. 2012. Tucson, AZ.
- [4] Hutton, C.J., et al., Real-time demand estimation in water distribution systems under uncertainty in WDSA 2012: 14th Water Distribution Systems Analysis Conference. 2012.
- [5] Davidson, J. and F.-C. Bouchart, Adjusting nodal demands in SCADA constrained real-time water distribution network models. *Journal of Hydraulic Engineering*, 2006. 132(1): p. 102-110.
- [6] Cheng, W. and Z. He, Calibration of nodal demand in water distribution systems. *Journal of Water Resources Planning and Management*, 2010. 137(1): p. 31-40.
- [7] Kun, D., et al., Inversion Model of Water Distribution Systems for Nodal Demand Calibration. *Journal of Water Resources Planning and Management*, 2015.
- [8] Do, N.C., et al., Calibration of water demand multipliers in water distribution systems using Genetic Algorithms. *Journal of Water Resources Planning & Management (Accepted on April 2016)*, 2016.
- [9] Rossman, L.A., *EPANET 2: Users Manual*. 2000, US Environmental Protection Agency (EPA): USA.
- [10] Goldberg, D.E. and K. Deb, A comparative analysis of selection schemes used in genetic algorithms. *Foundations of Genetic Algorithms*, 1991. 1: p. 69-93.
- [11] Chin, D.A., A. Mazumdar, and P.K. Roy, *Water-resources engineering*. Vol. 12. 2000: Prentice Hall Englewood Cliffs.