Modeling highway runoff pollutant levels using a data driven model

T. Opher, A. Ostfeld and E. Friedler

ABSTRACT

Pollutants accumulated on road pavement during dry periods are washed off the surface with runoff water during rainfall events, presenting a potentially hazardous non-point source of pollution. Estimation of pollutant loads in these runoff waters is required for developing mitigation and management strategies, yet the numerous factors involved and their complex interconnected influences make straightforward assessment almost impossible. Data driven models (DDMs) have lately been used in water and environmental research and have shown very good prediction ability. The proposed methodology of a coupled MT-GA model provides an effective, accurate and easily calibrated predictive model for EMC of highway runoff pollutants. The models were trained and verified using a comprehensive data set of runoff events monitored in various highways in California, USA. EMCs of Cr, Pb, Zn, TOC and TSS were modeled, using different combinations of explanatory variables. The models' prediction ability in terms of correlation between predicted and actual values of both training and verification data was mostly higher than previously reported values. Pb_{Total} was modeled with an outcome of R^2 of 0.95 on training data and 0.43 on verification data. The developed model for TOC achieved R^2 values of 0.91 and 0.49 on training and verification data respectively.

Key words | data driven model (DDM), event mean concentration (EMC), genetic algorithm (GA), highway runoff, model tree (MT)

T. Opher
A. Ostfeld
E. Friedler (corresponding author)
Department of Environmental,
Water and Agricultural Engineering,
Faculty of Civil and Environmental
Engineering,
Technion—Israel Institute of Technology,
Haifa 32000,
Israel
E-mail: *tamaro@tx.technion.ac.il*;
ostfeld@tx.technion.ac.il;
eranf@tx.technion.ac.il

INTRODUCTION

Highway or road runoff under certain circumstances can be a significant non-point source of pollutants. Vehicles, road wear and road maintenance produce a range of toxic contaminants such as heavy metals and polycyclic aromatic hydrocarbons (PAHs). Under certain conditions, related to the nature and characteristics of the highway, the rainfall-runoff event and the receiving water body or ecosystem, pollutants in highway runoff may exert acute or chronic impact on the receiving environment. The ecological impact of polluted runoff water on soil- and water-based ecosystems and its threat to aquifers and surface water has been elucidated, however, the processes affecting the buildup, transformation and reduction of these pollutants on the road surface during dry periods and their washoff, transport and dispersion during stormwater runoff events is a much more complex phenomenon and not yet well understood. Physical, chemical and biological processes are involved throughout this sequence of events. Though it has been the subject of numerous research projects, there are still open questions regarding the identity and mutual influences of the many factors affecting pollutant concentrations in road runoff. The lack of detailed physical, chemical and hydrological understanding of all processes involved has lead us to believe that methodology of Data Driven Modeling (DDM) may be ideal for confronting the challenge of predicting runoff pollutant concentrations. As so called "Grey Box" models, modeling tools of this type require only partial denotation of the underlying processes, while taking

doi: 10.2166/wst.2009.289

advantage of past events and available computing resources to deduce the likely outcomes of future events. In this study an attempt was made at isolating the major factors involved in determining pollutant contents in highway runoff and using these as explanatory variables for developing a data driven model.

METHODS

The proposed approach combines two data-driven methodologies, model trees (MT) and a genetic algorithm (GA) in a coupled scheme of alternating execution. The GA searches for optimal model coefficients which are then incorporated by the MT into the tree-structured model.

Model tree (MT)

MTs are a generalization of Decision Trees (DT), which are widely used in solving classification problems and more specifically very common in data mining applications. Whereas DTs handle qualitative or discrete-value attributes only, MTs deal with continuous values. An MT is a data driven algorithm, built as a rule-based predictive structure using a top–down induction approach. The tree is fitted to a training data set by splitting the data into homogeneous subsets based on the data attributes. The tree is constructed so that the target variable of all training cases is predicted by the tree leaves. Each leaf is a linear regression model which incorporates the numerical decision attributes and predicts continuous values for the target variable. The tree is then pruned bottom–up and transformed into a set of *if–then* rules, a process which simplifies its structure and improves its ability to classify new instances (Quinlan 1992). The predictive ability of the MT is measured using a correlation coefficient for the training and validation data sets.

Genetic algorithm (GA)

GAs are heuristic search procedures based on the mechanisms of genetics and Darwin's natural selection principles, combining an artificial survival of the fittest with genetic operators abstracted from nature (Holland 1975). GAs differ from other search techniques in that they search among a population of points and use probabilistic rather than deterministic transition rules. As a result, GAs search more globally (Wang 1997; Haupt & Haupt 1998).

An initial random population of genomes within the search space is generated. Each genome represents a possible solution to the search/optimization problem and is represented by a string of values (genes), one *per* each search variable. Survival of the fittest is accomplished by evaluating each genome's fitness through an appropriate objective function and a biased random selection procedure of individuals for "reproduction", where higher rated genomes are more likely to be selected. Generation of a new population is achieved by means of crossover (partial exchange of information between pairs of strings) and mutation (random change in a random location within a string). The fittest individuals are transferred unchanged to the next generation, an approach known as 'elitism'. Every new generation of genomes is expected to be more closely concentrated in the vicinity of the optimal solution. The process is repeated until a convergence criterion is met or a pre-set maximum number of generations reached. GA input parameters include: population size, number of generations, range limits of each gene, crossover and mutation rates and a fitness function for genome evaluation.

Source of data

The models in this study were trained and verified using a comprehensive data set of 68 runoff events monitored in 92 highway sites in California, USA between 1998 and 2004. Data was obtained from the Caltrans stormwater quality database (Caltrans 2004).

THE PROPOSED MT-GA MODEL

Figure 1 presents the structure of the proposed model. The GA module uses the MT's correlation coefficient (R^2) as its objective function and so optimization is guided by the accuracy of prediction achieved by the MT model, using each specific set of coefficients. In every generation the GA module calls the MT module for each of the genomes in the current population. The MT module constructs a model

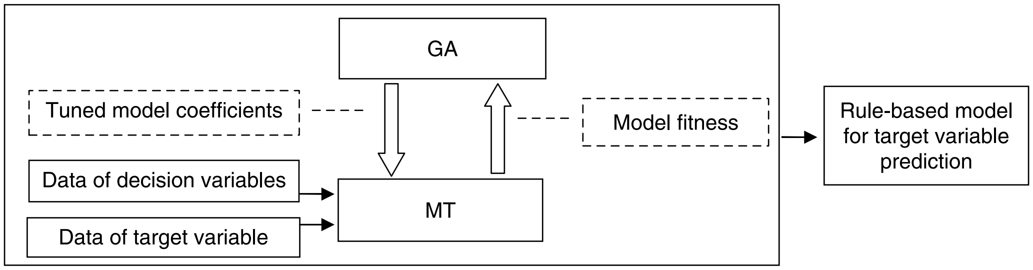
GA
Tuned model coefficients
Model fitness
Data of decision variables
MT
Data of target variable
Rule-based model for target variable prediction

Figure 1 | Schematic diagram of the proposed methodology.

using the coefficients coded by the genome and passes back to the GA module this model's R^2 for the training data. This value serves as the genome's fitness value. As the GA population advances towards the objective function optimum, the corresponding MT constructed with the tuned coefficients becomes more accurate in predicting the target value of the training data. As explained below, the MT decision variables are five site and storm characteristics and its target value is a given pollutant's event mean concentration (EMC). A similar approach has been proposed for flow and water quality predictions in watersheds and applied for daily loads of nutrients with very good results (Preis & Ostfeld 2008).

The coupled MT-GA model was coded in C#. The software incorporates the commercial Cubist M5 model tree protocol (Rulequest-Research 2007) as the core of the MT module. A graphical user interface (GUI), which enables the user to conveniently enter necessary modeling parameters, was also developed within the framework of this study.

Variables and coefficients

Target variables

EMC of five pollutants commonly found in highway stormwater runoff, representing three pollutant categories, were chosen as target variables for testing and demonstrating the proposed modeling approach: Pb_{Total}, Cr_{Total} and Zn_{Total} (total = particulate + dissolved fractions), TOC and TSS. TSS was selected for its significant positive correlation with many harmful pollutants found in highway runoff, making it an important goal for modeling, as it may serve as an indicator for other pollutants.

Decision variables

Selection of appropriate model inputs is extremely important in any prediction model. Often in DDM applications all input variables that might possibly have an influence on the model outputs are included and the DDM is left to determine which inputs are significant. However, presenting a large number of inputs and relying on the DDM to determine the critical model inputs, often results in the inclusion of insignificant model inputs (Solomatine 2003). In this study, based on available literature and on characterization and statistical investigation of the acquired data set, five variables were selected as potential inputs for the DDM, namely: annual average daily traffic (AADT) [10^3 vehicles/d], antecedent dry period (ADP) [d], event rainfall [mm], maximum 5-minute rain intensity [mm/h] and antecedent event rainfall [mm]. Of these five explanatory variables, all possible sub-group combinations were examined, to find the best minimal combination for each of the target variables.

Model coefficients

Unlike the standard use of DDMs, in the proposed modeling approach some of the available physical knowledge of the modeled phenomena was integrated into the model through mathematical expressions. Each decision variable was used in a specifically ascribed mathematical formula which is thought to roughly approximate its effect on pollutant EMCs (Table 1). In this study, unlike most applications of MT methodology, non-linear formulas were introduced to the modeling process which resulted in non-linear sub-models at the leaf nodes. The eight coefficients of these formulas were optimized by the GA in search of a set of values that will result in the best possible model.

Table 1 | MT-GA attributes and coefficients (genes)

Attribute	Equation	Description	Gene
AADT	$\alpha_1 \cdot \mathrm{AADT}$	Linear	α_1
ADP	$\alpha_2 \cdot [1 - \exp(-\alpha_3 \cdot \mathrm{ADP})]$	Saturation curve (Driscoll *et al.* 1990)	α_2
			α_3
Rainfall	$\alpha_4 \cdot \frac{\mathrm{Rainfall}}{\alpha_5} \cdot \exp\left(1 - \frac{\mathrm{Rainfall}}{\alpha_5}\right)$	Early maximum then fading	α_4
			α_5
Max. rain intensity	$\mathrm{Max\,Intensity}^{\alpha_6}$	Power (Yuan *et al.* 2001; Francey *et al.* 2004)	α_6
Ant. rainfall	$\alpha_7 \cdot \mathrm{Ant\,Rain}^{-\alpha_8}$	Power	α_7
			α_8

The linear equation for the effect of daily traffic (AADT) on runoff concentrations represents the accumulation of vehicle-originated substances on the highway surface. The impact of ADP is represented as a cumulative process with a saturation curve, emanating from the paved surface's carrying capacity, beyond which processes of removal (by air currents, chemical or biological decomposition, volatilization, etc.) restrain further accumulation of substances on the surface. The effect of rainfall is illustrated by a curve presenting an initial climb followed by a gradual descent. This function represents the increasing loads of pollutants washed from the road with the initially growing force of runoff flow, known as the First Flush phenomenon (Bertrand-Krajewski *et al.* 1998; Han *et al.* 2006) and then a decrease, as less matter is left of the road surface to be washed off, while growing quantities of stormwater have a diluting effect on the overall concentrations. The effect of maximum rainfall intensity is portrayed by a positive power function, as the increasing shear force produced by raindrops on the pavement may release substances with stronger adhesion, or those located deeper in the asphalt crevasses. Antecedent rainfall is thought to have an inverse proportion to the current storm's EMC, since a previous heavier rainfall leaves smaller loads of pollutants on the surface, to be washed off by the current storm. This effect is represented here by a negative power function.

Model training and verification

Every possible combination of 1 to 5 attributes was examined for predicting the EMC of each of the five pollutants. For each combination of variables a simple MT was first constructed. Every model which incorporates two or more model coefficients was then optimized, using the proposed MT-GA approach, resulting in 29 models per target variable. Model training was carried out in two phases: in the first phase the MT-GA application was run with a population of 10 genomes and 50 generations. In the second phase those models which showed a good potential for further improvement were run with 20 genomes per population and 500 generations, again starting from a random initial population.

A set of 850–1,100 data entries was available for each target variable, the statistics of which are presented in Table 2. Each data set was randomly divided into two subsets: 70% used for model training and 30% for verification. Evaluation of the MT-GA model is based on the fitness (R_T^2, training R^2) of the model. R_T^2 expresses the correlation between predicted and observed target variable values in the training data. Altogether 145 models were created; each one was tested using its relevant set of verification data. The correlation between predicted and actual verification data (R_V^2, verification R^2) was used as a measure of the model's predictive performance.

RESULTS AND DISCUSSION

Models

The five best models, one for each pollutant, vary in length, in the set of explanatory variables and in accuracy of prediction. Table 3 presents, as an example, the sequence of classification rules constituting the model for $\mathrm{Zn}_{\mathrm{Total}}$. This model is the most compact of the five; others consist

Table 2 | Summary statistics of all data sets used in the modeling process

	Data type	Count	Minimum	Maximum	Median	Mean	STD
Cr [μg/L]	Training	571	0.5	86	5.8	8.3	8.78
	Verification	244	0.5	98	5.7	8.1	10.30
Pb [μg/L]	Training	608	0.0	2,600	11.0	54	187
	Verification	261	0.0	1,400	11.3	57	154
Zn [μg/L]	Training	608	1.0	2,100	130	196	238
	Verification	261	2.5	1,665	120	188	217
TOC [mg/L]	Training	734	0.5	550	13.2	19.8	31.7
	Verification	315	2.0	151	15.0	20.9	21.1
TSS [mg/L]	Training	773	0.5	2,400	63	108	169
	Verification	332	0.5	2,988	63	110	198

Table 3 | Classification rules of the model for Zn_{Total}

Rule 1	If	AADT-K < = 50.81
	Then	$Zn_{Total} = 4.6881 + 2.92$ AADT-K + 171 Ant Event Rain + 88 ADP
Rule 2	If	AADT-K > 93.15 and Max Intensity > 2.8374
	Then	$Zn_{Total} = 420.89 - 50$ Max Intensity + 141 ADP
Rule 3	If	AADT-K > 50.81
	Then	$Zn_{Total} = 175.03 + 282$ ADP − 25 Max Intensity
Rule 4	If	AADT-K > 93.15 and Max Intensity < = 2.8374 and Ant Event Rain > 0.1092
	Then	$Zn_{Total} = -365.48 + 4.1$ AADT-K + 1166 ADP − 155 Max Intensity
Rule 5	If	AADT-K > 93.15 and Max Intensity < = 2.8374 and Ant Event Rain < = 0.1092
	Then	$Zn_{Total} = 92.60 + 5865$ Ant Event Rain + 95 ADP

of 6–15 rules. Since the MT-GA output is designated for MT-GA automated predictions, variable names in the models' rules represent the mathematical expressions used to encode them (Table 1). Whereas the expressions for computing pollutant EMCs in the model rules appear to be linear, they are actually often non-linear, once variable names are substituted with their corresponding mathematical expressions.

The resulting MT-GA models are unlike those previously reported, such as MLR or process-based equations, in that they ascribe a few different equations to each pollutant target. And yet, similarly to previously reported models, for pollutant loading (Irish *et al.* 1998; Kim *et al.* 2005*a*) or EMCs (Kayhanian *et al.* 2007), each sub-model in an MT-GA leaf node takes the general form of a multi-term summation equation, consisting of the significant factors affecting the target variable. For example, rule number 1 in the model for Zn_{Total} (Table 3), after substitution, takes the form:

$$\text{If AADT}-\text{K} \leq 50.8/\alpha_1 \text{ Then}$$

$$Zn_{Total} = 4.7 + 2.9\cdot\alpha_1\cdot\text{AADT}-\text{K} + 171\cdot\alpha_7\cdot\text{Ant Event Rain}^{-\alpha 8}$$
$$+ 88\cdot\alpha_2\cdot[1 - \exp(1 - \alpha_3\cdot\text{ADP})]$$

Model attributes

Each of the five target variables was best modeled by a different set of explanatory variables. Table 4 presents a summary of model attributes and performance of the developed models. For each category two models are presented: the one with best performance on the training data and the one with best performance on the verification data. The number of attributes used in these models varies between 2 and 5, proving that it is not always advisable to

Table 4 | Model attributes and performance for the best models in each target category

	R^2	Type of R^{2*}	AADT	ADP	Rainfall	Max. rain intensity	Antecedent rainfall	No. of variables
Cr_{Total}	0.77	T	■	■	■	■	■	5
	0.56	V	■		■			2
Pb_{Total}	0.95	T	■	■	■	□†	□†	4
	0.43	V	■	■	■		■	4
Zn_{Total}	0.84	T	■	■	■		■	4
	0.49	V	■	■		■	■	4
TOC	0.93	T	■	■	■	■	■	5
	0.49	V	■	■	■			3
TSS	0.82	T	■	■	■		■	4
	0.32	V	■		■	■		3

*T, training; V, verification.
†Either maximum rain intensity or antecedent rainfall as fourth variable gives the same R^2_T.

have the DDM work with all optional input variables. Though $R^2_T$of the full 5-variable models was always among the highest for each target variable, R^2_v was often considerably lower than that of certain other models using fewer variables. This implies possible over-fitting when using redundant attributes. For example, the 5-attribute model for TOC displays R^2_T of 0.93, which is the highest among all TOC models (closely followed by 0.92 for a 4-attribute model), but its R^2_v is a mere 0.22, which is much lower than the 0.49 maximum, achieved by another 4-attribute model.

In the case of Pb_{Total} the 5-attribute model achieved identical prediction accuracy for both training and verification data (0.95 and 0.43 respectively) as the 4-attribute one in which the attribute maximum rain intensity was left out. The five-variable model consists of 12 rules, while the four-variable model contains 15 rules. Maximum rain intensity is used in the five-variable model as a variable in the equations at the leaf nodes, in only 6% of the cases. This minor contribution of additional data obviously makes it possible to make predictions of the same quality using a more compact tree structure.

Annual average daily traffic is clearly indicated as the most influencing factor on the EMCs of the pollutants modeled. Not a single model in Table 4 disregards it. Moreover, looking through the full result tables of the training process (not presented in the scope of this paper for lack of space), it becomes obvious that whenever AADT is left out of a model its performance is significantly compromised. The second most common attribute within the best models is event rainfall, participating in nine of the ten models.

Table 5 compares attribute combinations of the models in this study with those used in models from three previously reported studies, all using multiple linear regression analysis (comparable data for TOC was unavailable). Many common traits are apparent for selected explanatory variables in all four studies. Generally, the sets of variables chosen in this study are closer to those presented by Kayhanian *et al.* (2003, 2007) than to those presented by Irish *et al.* (1995). Though this difference may be coincidental, it should be noted that Kayhanian *et al.* used data collected from the same geographical area as that used in the current study (California), while Irish *et al.* modeled highway runoff data collected in a different geographic and climatic location (Austin, Texas). Traffic-related variables were identified as the most significant factors in modeling highway runoff quality according to the current and additional three studies. These, in different forms, were found to be significant influencing factors in all but one model (that for TSS by Irish *et al.* 1995). It is noteworthy that some reports have concluded that there is no definitive relationship between AADT and pollutant concentrations (Driscoll *et al.* 1990) and others suggested that such a relationship exists only for certain contaminants and in high AADT sites (Kayhanian *et al.* 2003).

Attributes relating to rainfall volume are the second most commonly used type of attributes, participating in 11 of the total 14 models in Table 5. Disagreement between

Table 5 | Comparison of participating attributes in various models

Variable	Model	Traffic	ADP	Rainfall	Intensity	Previous storm	Others
Cr_{Total}	Current study	AADT		Rainfall			
	Kayhanian *et al.* (2003)	AADT	ADP	Rainfall			SCR
Pb_{Total}	Current study	AADT	ADP	Rainfall		PRAINFALL	
	Kayhanian *et al.* (2007)	AADT					SCR
	Kayhanian *et al.* (2003)	AADT	ADP				SCR DA
	Irish *et al.* (1998)	VDS			Intensity	PINT	Flow
Zn_{Total}	Current study	AADT	ADP		Max. Intensity	PRAINFALL	
	Kayhanian *et al.* (2007)	AADT	ADP	Rainfall			SCR
	Kayhanian *et al.* (2003)	AADT	ADP	Rainfall	Max. Intensity		SCR DA
	Irish *et al.* (1998)	ATC				PFLOW PDUR PINT	DUR Flow
TSS	Current study	AADT		Rainfall	Max. Intensity		
	Kayhanian *et al.* (2007)	AADT	ADP	Rainfall			SCR
	Kayhanian *et al.* (2003)	AADT	ADP	Rainfall	Max. Intensity		SCR DA
	Irish *et al.* (1998)		ADP	Flow	Intensity	PINT	

DA, drainage area; DUR, storm duration; Flow, total volume of runoff per unit area of watershed; Intensity, flow divided by duration; PDUR, duration of previous storm event; PFLOW, total volume of runoff per unit area of watershed during the previous storm event; PINT, PFLOW divided by PDUR (L/m^2/min); PRAINFALL, Previous storm rainfall; SCR, Seasonal cumulative rainfall; VDS, Single-lane vehicle count during storm event.

the current study and all other three studies exists for rainfall volume in the Zn_{Total} model (present in all but the currently presented model) and for ADP in the TSS model (used in all other three models, but left out of the one presented here).

Model training

The process of training was very instructive, as it gradually revealed unexpected general trends regarding the ability of the various combinations of attributes to explain the variability of the different pollutant concentrations. Different training runs displayed distinctive courses of progress. Figure 2 shows a random selection of thirteen first phase training runs. The best fitness (i.e. the MT's R_T^2) in each generation is plotted along the course of the 50 generations. Some graphs are continuously climbing, such as model number 13223, representing a model for Zn_{Total} with attributes ADP and event rainfall. Some start out with a good rate of improvement but converge to their maximum quite quickly, such as 14225 (TOC explained by ADP and antecedent rainfall). Other models displayed no improvement in fitness whatsoever, as did number 12215 (Cr_{Total} explained by AADT and antecedent rainfall), which remained steady at a fitness of 0.46 throughout the training process.

Models showing a potential of further improvement (i.e. those which displayed any increase in fitness within the 50 generations of the first phase runs) were trained again in phase 2, with a larger genome population and a much longer GA evolution of 500 generations. In phase 2 the majority of models converged to their optimum solution within the first 300 generations. A few continued improving as far as generation number 485, but these were all models

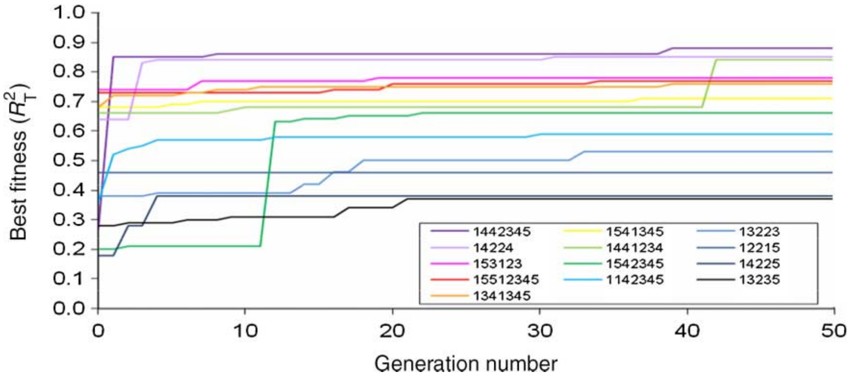

Figure 2 | Fitness improvement curves for a sample of first phase model training runs.

with a fitness value considerably inferior to the best model in their pollutant category and were therefore not pursued further.

Highest R^2_T values of the chosen models for all target variables are satisfactory, ranging from 0.77 to 0.95. These values were found to be significantly higher than the R^2_T of most other models reported in the literature for Cr, Pb and Zn and among the highest for TOC and TSS (Table 6).

Model evaluation

Each model, once trained and optimized by MT-GA, was evaluated using a set of verification data (not used for model training). Verification data sets consisted 242–330 cases (Table 2). Coefficient of correlation between predicted and actual values of the verification data (R^2_v) was used for evaluating each model's predictions of previously unseen cases. It should be noted that no comparable values were found in the literature, as most studies report only correlation of predictions on data used for model calibration (R^2_T presented in Table 6).

Table 6 | R^2_T of models developed in this study and other models from the literature

	Current study	**Kayhanian *et al.* (2007)**	**Kim *et al.* (2005*b*)**	**Kayhanian *et al.* (2003)**	**Irish *et al.* (1995)**
Cr_{Total}	0.77			0.21	
Pb_{Total}	0.95	0.36		0.35	0.68
Zn_{Total}	0.84	0.51		0.45	0.92
TOC	0.93	0.14	0.98		
TSS	0.82	0.25	0.84	0.19	0.93

Adjusted R^2_v values range between 0.32 (for TSS) and 0.56 (for Cr_{Total}), generally displaying some underestimation of extreme high EMCs. The results show that there is no consistent correlation between training and verification R^2 values, i.e. a model's high R^2_T does not necessarily indicate a high R^2_v. This is demonstrated in Figure 3, which shows correlation coefficients between modeled and actual values for two selected models per target pollutant, one for highest R^2_T and another for highest R^2_v. This lack of consistency between a model's relative performance on training and test data sets is disappointing, since the assumption that a model's accuracy of predictions on its training data indicates its future performance on unseen cases is at the basis of this methodology and of the concept of data driven modeling at large. R^2_T should be a good indication of R^2_v, which represents the use of the model as a prediction tool for future events. Better correlation between R^2_T and R^2_v may be achieved by applying a different method of partitioning of the data into training and verification sets, which would result in different model rules. In this study data partitioning was performed arbitrarily, yet dividing the available data by seasons or by monitoring sites may be more informative for the models' training process. Amplifying the relative weight of extreme high cases may reduce the difference between training and verification scores. It is also reasonable to assume that a larger set of training data would be more representative of the general regularities characteristic of the modeled phenomena and result in a more indicative evaluation of future predictions. Another possible reason for this poor correlation could be existence of other explanatory variables affecting the EMC not considered by the model.

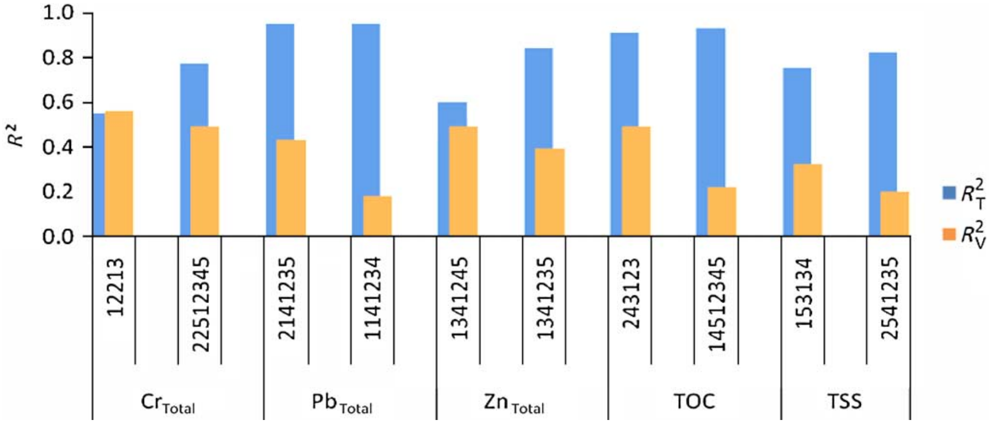

Figure 3 | R_T^2 and R_V^2 of selected models for each target pollutant.

CONCLUSIONS

The innovative approach of a coupled MT-GA modeling technique was implemented. Unlike most previous studies, in the current study the MT included some non-linear equations. Models for five highway runoff pollutants (Cr_{Total}, Pb_{Total}, Zn_{Total}, TOC and TSS) were trained and tested using an extensive data set from the Caltrans stormwater monitoring database. The coupled model was found to be a convenient and effective methodology for highway runoff quality predictions.

Five key factors known to affect runoff pollutant concentrations were selected as optional modeling attributes. All combinations of 1 to 5 explanatory variables of the five variables selected were tested for modeling each constituent's EMC. Each constituent was found to be best modeled by a different set of attributes. Of the five candidate variables, the most frequently used attribute is AADT, implying that this is the most influencing factor on runoff EMCs. The second most common variable in the developed models is event rainfall, left out of only one of the ten best models.

Correlations between predicted and actual EMCs for the models' training data were very good, ranging from 0.77 to 0.95, and better in most cases than those achieved by multiple linear regression models reported in the literature. Correlation coefficients for predicted and actual EMCs of the verification data set were significantly lower than those of the training data, ranging between 0.32 and 0.56. This suggests that there may be other explanatory variables affecting the EMC not considered by the model. Another possibility is that dividing the available data non-randomly (by seasons or by monitoring sites) could result in a better correlation between training and test performance. Comparison of our models' accuracy of prediction on unseen input data was prevented since comparable model verification data is unavailable in the literature.

ACKNOWLEDGEMENTS

The research was financed by The Grand Water Research Institute, Technion. The authors wish to thank Mr. T. Granicher of CSUS Office of Water Programs, Sacramento CA, Caltrans Stormwater Monitoring Database administrator, for providing the data.

REFERENCES

Bertrand-Krajewski, J.-L., Ghassan, C. & Saget, A. 1998 Distribution of pollutant mass *vs* volume in stormwater discharges and the first flush phenomenon. *Water Res.* **32**(8), 2341–2356.

Caltrans 2004 Caltrans Stormwater Quality Data Base. Sacramento, CA, California Department of Transportation (Caltrans).

Driscoll, E. D., Shelley, P. E., Strecker, E. W., FHWA & US Department of Transportation 1990 Pollutant loadings and impacts from highway stormwater runoff Volume 3: Analytical investigation and research report (FHWA-RD-88-008).

Francey, M., Duncan, H. P., Deletic, A. & Fletcher, T. D. 2004 *An advance in modelling pollutant loads in urban runoff*. International Conference on Urban Drainage Modelling, Dresden.

Han, Y., Lau, S.-L., Kayhanian, M. & Stenstorm, M. K. 2006 Characteristics of highway stormwater runoff. *Water Environ. Res.* **78**(12), 2377–2388.

Haupt, R. L. & Haupt, S. E. 1998 *Practical Genetic Algorithms*. John Wiley & Sons, Inc., New York.

Holland, J. H. 1975 *Adaptations in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor.

Irish, L. B. J., Barrett, M. E., Malina, J. F. J. & Charbeneau, R. J. 1998 Use of regression models for analyzing highway stormwater loads. *J. Environ. Eng.* **124**, 987–993.

Irish, L. B. J., Lesso, W. G., Barrett, M. E., Malina, J. F. J., Charbeneau R. J. & Center for Research in Water Resources 1995 An evaluation of the factors affecting the quality of highway runoff in the Austin, Texas area (Technical Report CRWR 264). pp. 248.

Kayhanian, M., Singh, A., Suverkropp, C. & Borroum, S. 2003 Impact of annual average daily traffic on highway runoff pollutant concentrations. *J. Environ. Eng.* **129**(11), 975–990.

Kayhanian, M., Suverkropp, C., Ruby, A. & Tsay, K. 2007 Characterization and prediction of highway runoff constituent event mean concentration. *J. Environ. Manage.* **85**, 279–295.

Kim, L. H., Kayhanian, M., Lau, S. L. & Stenstorm, M. K. 2005*a* A new modeling approach for estimating first flush metal mass loading. *Water Sci. Technol.* **51**(3–4), 159–167.

Kim, L. H., Kayhanian, M., Zoh, K. D. & Stenstorm, M. K. 2005*b* Modeling of highway stormwater runoff. *Sci. Total Environ.* **348**, 1–18.

Preis, A. & Ostfeld, A. 2008 A coupled model tree–genetic algorithm scheme for flow and water quality predictions in watersheds. *J. Hydrol.* **349**, 364–375.

Quinlan, J. R. 1992 *Learning with Continuous Classes*. Fifth Australian Joint Conference on Artificial Intelligence, World Scientific, Singapore.

Rulequest-Research 2007 Data Mining with Cubist. http://www.rulequest.com/cubist-info.html, visited 27 February, 2006.

Solomatine, D. P. 2003 Model trees as an alternative to neural networks in rainfall–runoff modelling. *Hydrol. Sci.* **48**(3).

Wang, Q. J. 1997 Using genetic algorithms to optimise model parameters. *Environ. Model. Softw.* **12**(1), 27–34.

Yuan, Y., Hall, K. & Oldham, C. 2001 A preliminary model for predicting heavy metal contaminant loading from an urban catchment. *Sci. Total Environ.* **266**(1–3), 299–307.