A preliminary coupled MT–GA model for the prediction of highway runoff quality

Tamar Opher, Eran Friedler *

Department of Environmental, Water and Agricultural Engineering, Faculty of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Technion City, Haifa 32000, Israel

A R T I C L E  I N F O

Article history:
Received 21 December 2008
Received in revised form 31 March 2009
Accepted 19 April 2009
Available online 15 May 2009

Keywords:
Data-driven model (DDM)
Event mean concentration (EMC)
Genetic algorithm (GA)
Highway runoff
Model tree (MT)

A B S T R A C T

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 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 tree–genetic algorithm) model provides an effective, accurate and easily calibrated predictive model for EMC (event mean concentration) 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. Sensitivity analysis was performed to examine the relative significance of each explanatory variable and the models' response to changes in input values.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Highway runoff may 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 an 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, transport and removal of these pollutants on and from the road surface involve a multitude of complex phenomena which have not yet been thoroughly clarified (Barrett et al., 1995; Irish et al., 1995). Physical, chemical and biological processes are involved in this cluster 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 pollutants concentrations in road runoff. The lack of detailed physical, chemical and hydrological understanding of all processes involved has lead us to an attempt at using the methodology of data-driven modeling (DDM) 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 advantage of past events and available computing resources to deduce the likely outcomes of future events (Solomatine, 2002; Solomatine and Ostfeld, 2008). In this study the major factors involved in determining pollutant contents in highway runoff were determined and used as explanatory variables in a coupled model tree (MT) and genetic algorithm (GA) optimized model.

2. Methods

2.1. Model variables

EMCs (event mean concentrations) of five pollutants commonly found in highway stormwater runoff, representing three pollutant categories (heavy metals, organics and suspended matter), were chosen as target variables for testing and demonstrating the proposed modeling approach: PbTotal, CrTotal and ZnTotal (total = particulate + dissolved fractions), TOC and TSS. TSS was selected for its significant positive correlation with many harmful pollutants found in highway runoff. These correlations make TSS an important goal for modeling, as it may serve as an indicator for other pollutants.

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). The inclusion of redundant attributes in a DDM may lead to a decrease in the model’s generalizability, which is critical for its performance on unseen (i.e. non-training) data. In this study, based on available literature (e.g. (Driscoll et al., 1990; Barrett et al., 1995; Irish et al., 1995; Legret and Pagotto, 1999)) 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.

2.2. Modeling approach

The proposed approach combines two data-driven methodologies, model trees (MTs) 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. Additionally, approximate procedural mathematical expressions are used to incorporate some of the accumulated physical knowledge of the involved processes.

2.3. Model tree (MT)

MTs are a generalization of Decision Trees (DTs), 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 correctly predicted in 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 strength of the MT is evaluated by the correlation coefficient for the predicted and actual values of its target variable.

2.4. 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 and 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 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 (a random change in a random location within the 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 is 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.

2.5. MT-GA coupled model

Unlike the standard use of DDMs, the modeling approach proposed here enables integration of available physical knowledge of the modeled phenomena into the model through mathematical expressions. Each explaining variable was placed in a specifically ascribed mathematical formula which is thought to roughly approximate its effect on pollutant EMCs (Table 1). In this study, as a result of the MT-GA coupling, 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 (Fig. 1). The five coefficients of these formulas were optimized by the GA in search of the set of values that will result in the best possible model.

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 ADT 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

![Table 1](image)
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.

The model tree and genetic algorithm were combined as outlined in Fig. 2. The GA module uses the correlation coefficient (R) between EMC values predicted by the MT and the observed EMC values 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 MT module is called by the GA module for each of the genomes in the current population. A model tree is constructed, using the coefficients coded by the genome and the model’s $R^2$ for the training data is passed back to the GA. 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 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 and Ostfeld, 2008).

The coupled MT–GA model was coded in C#, incorporating the commercial Cubist M5 model tree protocol (Rulequest-Research, 2007) as the core of the MT module.

2.6. 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, in which attribute values are taken as-is into the learning algorithm, was first constructed. Every model which incorporates two or more model coefficients was then optimized, using the proposed MT–GA approach, resulting in 26 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 20 genomes and 50 generations. In the second phase those models which showed a good potential for further improvement were run for 500 generations, again starting from a random initial population of 20 genomes. Following guidelines suggested by Jian-Xia et al. (2005) and Liu et al. (2007) and some preliminary exploratory testing, the GA algorithm was run with 80% crossover, 30% mutation rate and elitism of the top 10% of the population.

Evaluation of the MT–GA models in the training process is based on their fitness, represented by $R^2_T$ (training $R^2$), which expresses the correlation between predicted and observed target variable values in the training data. Each optimized model was then tested using its complete data set of training + verification data. The resulting correlation ($R^2_{All}$) was used as a measure of the model’s predictive performance (Preis and Ostfeld, 2008).

3. Data

The models 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 (unpublished). A total of 1105 data entries, each representing a single sample of runoff, were available (Table 2). A set of 850 to 1100 data entries was extracted for each target variable, the statistics of which are presented in Table 3. Each data set was randomly divided into two subsets: 70% used for model training and 30% left aside for verification. Data items reported as ‘not

Table 2
Summary statistics of explanatory variables ($n=1105$).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Median</th>
<th>Mean</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADT [1000 vehicles/d]</td>
<td>1.8</td>
<td>328.0</td>
<td>74.0</td>
<td>104.7</td>
<td>92.4</td>
</tr>
<tr>
<td>ADF [d]</td>
<td>0.2</td>
<td>234.0</td>
<td>9.0</td>
<td>13.5</td>
<td>19.7</td>
</tr>
<tr>
<td>Rainfall [mm]</td>
<td>1.5</td>
<td>324.7</td>
<td>16.9</td>
<td>25.3</td>
<td>27.6</td>
</tr>
<tr>
<td>Max. 5-minute intensity [mm/h]</td>
<td>0.04</td>
<td>112.8</td>
<td>12.2</td>
<td>17.3</td>
<td>15.7</td>
</tr>
<tr>
<td>Antecedent event rain [mm]</td>
<td>0.25</td>
<td>206.2</td>
<td>9.9</td>
<td>19.6</td>
<td>26.1</td>
</tr>
</tbody>
</table>
detected’ was substituted with values of half the detection limit, as is often done (Wilkie et al., 1996; Kayhanian et al., 2007).

4. Results

Within each target pollutant category all possible combinations of explaining variables were trained and verified using the MT–GA algorithm. The resulting models’ performance was evaluated by their squared correlation coefficient for the complete data set, including training and verification subsets ($R^2_{\text{tr}}$). This measure of performance is similar to that proposed by Preis and Ostfeld, for DDMs of quantity and quality flow in watersheds (Preis, 2004; Preis and Ostfeld, 2008). A single model per category, presenting the best accuracy of prediction, was selected. The five best models, one for each pollutant (Table 4), vary in length, in the set of explanatory variables and in accuracy of prediction.

Table 4 presents a summary of model attributes and performance of the developed models for each target variable. Basic MT performance is brought as a comparison to the improved MT–GA models’ performance (each represented by the corresponding $R^2$ and $R^2_{\text{tr}}$). Apparently, ADP was signalled out as the only explaining variable which has a major role in explaining the variance of all five target pollutants. AADT is the second most common variable, taking part in all but the model for TOC. Rainfall and maximum rainfall intensity appear alternately in the models, one of the two in each one. The partial correlation which exists between these two variables seems to result in redundancy when they are both included in a model. It may therefore be said that length of the dry period and characteristics of the current storm (either rainfall volume or maximum intensity) are the two factors found to significantly affect all five modeled pollutants’ concentrations in the runoff.

Each model is assembled as a sequence of if–then deductive rules, introducing numeric conditions for the values of the explaining variables and providing a formula for calculating the forecast EMC of the target variable, through mathematical operations on these values. Variable names in the model 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. Table 5 demonstrates the form of model rules for Pb before and after this substitution. Table 6 shows the attribute coefficients as calibrated by the MT–GA model.

5. Discussion

5.1. The MT–GA algorithm

In most cases the MT–GA coupled methodology worked well, in that it resulted in an optimized set of model coefficients which improved the performance of the original model, constructed using the MT algorithm alone (i.e. a model with linear sub-models in the leaf nodes and no coefficients for calibration).

Only nine (6%) of the 130 different MT–GA models did not display a higher $R^2_{\text{tr}}$ than their MT counterparts. Six of these nine are models containing only two coefficients, a fact that may be ascribed to an inferior efficiency of the GA optimization mechanisms on small genomes. The crossover between a pair of two-gene genomes may produce only two different child genomes (with only a single possible point for crossover), therefore introducing only minimal diversity to the population of the next generation.

5.2. Model attributes

AADT (or similar parameters), ADP and rainfall volume (or runoff flow) are the three most commonly used explaining variables in highway runoff pollutant models previously reported (Irish et al., 1998; Kayhanian et al., 2003, 2007). In the current study AADT and ADP proved to be the most essential of the five potential attributes tested (Table 4). Rainfall volume and maximum rainfall intensity are used alternately in the models, implying that the effect of the current storm on the degree of pollutant washoff, either through continuous flushing and erosion, as represented by volume, or through shear force applied by the rain drops, is also among the most important factors in determining pollutant EMCS.

TOC is exceptional in that its model does not incorporate AADT as an explaining variable. Nevertheless, it achieves the highest result of the five models. Interestingly, the second best model for TOC ($R^2=0.67$) also does not include AADT, while all other tested combinations result in far lesser correlations ($R^2\leq0.34$). This may seem surprising, as organic compounds in highway runoff include VOCs, PAHs and other aromatic compounds, originating largely in gasoline, fuel additives and engine oils. The insensitivity of TOC to substitution.
AADT may insinuate either that organic compounds on the highway surface are mostly natural substances, originating from surrounding vegetation, or that the motor-related substances are abundantly present on the highway surface, regardless of traffic count. In the latter case factors affecting the extent of washoff (such as rainfall volume or maximum intensity) or the availability of surface adherence sites (such as antecedent rainfall) may be more significant in determining concentration in the runoff than is traffic density.

The washoff rate of particulate contaminants is generally expected to be more affected by the force of raindrops hitting the road surface and by the shear force of runoff flow than that of dissolved constituents, which are more readily transferred to the aqueous phase, regardless of these forces (Irish et al., 1995; Furumai et al., 2002). This trend has been demonstrated in studies modeling highway runoff pollutographs (Massoudieh et al., 2008). In the current study, modeling event mean concentrations using whole highway runoff pollutographs (Massoudieh et al., 2008). In the current study, modeling event mean concentrations using whole storm attributes, these different washoff behaviors may be obscured by the contradicting effects of rainfall intensity on particle detachment and that of total runoff volume in diluting the overall pollutant concentration. It is, therefore, not surprising that no consistent pattern is evident in the inclusion of the attribute of maximum intensity in the models for particle-bound constituents, such as Pb, TOC and TSS.

### 5.3. Model performance

All models display a tendency for underestimation, as may be seen in the predicted vs. observed plots of two of the models (Fig. 3). Some measure of underestimation of the higher-range inputs is an inherent deficiency of a data-driven technique, whose accuracy primarily depends on the dataset used for training. Since the high-range observations are naturally infrequent in the dataset, their weight in the training process is small and their characteristics cannot be thoroughly “studied” or trained upon.

Compared with models previously reported (Table 7), the currently developed MT–GA models give satisfactory results, which are better than most others. The modeling methodology used in each study is specified in row 2 in the table. Note that $R^2$ values presented here are those of $R^2_1$, since this is the only available measure for comparison with the other reported studies.

### 5.4. Model sensitivity and applicability

Sensitivity analysis was performed to examine the extent of the MT–GA models’ sensitivity to changes in input data. The models’ responses to a two-fold change of input variables displayed minimal sensitivity to all variables, with most predictions unchanged, or with an indecisive response direction (e.g. the $\text{Cr}_{\text{Total}}$ model’s response to an input of 2 $\times$ ADP was expressed by 16.3% of $\text{Cr}_{\text{Total}}$ concentration predictions being increased and 2.2% of predictions decreased, while the remaining 81.5% remained unchanged). The responses to a 50% decrease in the input values of AADT were more notable. For example, 69.9% of $\text{Pb}_{\text{Total}}$ concentration predictions for a halved AADT value were decreased relative to the original model’s predictions, while only 6.6% of predictions increased and 23.5% remained unchanged. All in all the models’ sensitivity appears quite low.

An additional aspect of sensitivity is the measure of impact that the presence or absence of a certain attribute has on the performance of the model when making predictions on the raw data. This analysis was carried out using statistics of the $R^2_1$ for each target variable category group, e.g. statistics of the performance of all TSS models containing the attribute ‘rainfall volume’ was compared with those of all TSS models which do not contain this attribute. The results, as demonstrated in Fig. 4, show that for each pollutant modeled the presence or absence of only one of the attributes has a dramatic influence on the models’ predicted–observed correlations. This may be observed by the overlapping of the ranges of the $R^2$ values in four out of the five box pairs in each section of the figure, while in the pair of boxes corresponding to the remaining attribute the two ranges are distinct.
For Cr, Pb, Zn and TSS (not shown in the figure) this single attribute is AADT. The TOC model, on the other hand, while apparently insensitive to the presence of AADT, shows similar sensitivity to the presence or absence of the rainfall volume attribute.

In examining the applicability of the MT–GA models as aids in environmental risk assessment it is likely that, like all data-driven models, these too will perform considerably less reliably when applied to input data which is out of the range of the training data used in their construction. Applying these models to runoff-related data from geographical regions with climatic patterns significantly different than those found in California, would require a simple calibration process to achieve a better fit.

6. Conclusions

A model for each of five common highway runoff pollutants was developed using the hybrid data-driven methodology of MT–GA. Every such model consists of a set of deductive rules which represents an optimized model tree structure. The altered MT algorithm incorporates some knowledge of physical processes involved through non-linear equations. The resulting tree, therefore, contains non-linear formulas at its leaves, unlike standard model trees, which carry a linear sub-model at each leaf.

The models’ correlation coefficients between predicted and observed pollutant EMCs are within the range of previously reported models of other modeling disciplines, or higher. As in the case of modeling methodologies based on regression analysis, another benefit of the MT–GA, apart from the estimation of pollutant concentrations in the runoff, is its emphasis of the most influential explaining variables, which appear to be ADP, AADT and a characteristic of the current storm (either rainfall volume or maximum rainfall intensity). In cases of relatively low model performance, such as those of CrTotal and TSS, there may be other significant factors which have not been included in this modeling effort.

Models’ sensitivity to changes in input values was low, while sensitivity to the incorporation of a certain attribute was very notable for all target variables. Each model is sensitive to a single attribute and quite indifferent to the other four. This one attribute is, in most cases that of daily traffic, with the exception of TOC, which responds to the presence of event rainfall rather than that of traffic count.

Available data is insufficient in determining the preferable modeling methodology, yet the results obtained demonstrate that data-driven modeling is certainly worth further refinement and examination as a solution for highway runoff pollution modeling.

Models of this sort may function as a tool for environmental risk assessment in seeking the BMP (best management practice) for handling highway runoff waters which have been confined off-road. According to the modeled EMCs appropriate action may be taken. When planning the route of a new highway or in debating the need for runoff containment facilities alongside it, models such as these may be used with a few approximations of the yet unknown input data. Future storm characteristics such as total rainfall, maximum intensity or length of dry period must then be based on multi-annual data.

Acknowledgments

This research was funded by the Grand Water Research Institute (GWRI).

The authors would like to acknowledge the assistance of the California State University Sacramento, Office of Water Programs, in acquiring the data used in this project.
References


Irish LBJ, Lesso WG, Barrett ME, Malina JFJ, Charbeneau RJ. An evaluation of the factors affecting the quality of highway runoff in the Austin, Texas area (technical report CRWR 264); 1995. 248 pages.


Quinlan JR. Learning with continuous classes. 5th Australian Joint Conference on Artificial Intelligence, World Scientific, Singapore; 1992.


Solomatine DP. Data-driven modeling: paradigm, methods, experienced. 5th International Conference on Hydroinformatics, Cardiff, UK; 2002.


