Data-mining approach to investigate sedimentation features in combined sewer overflows

M. Carbone, L. Berardi, D. Laucelli and P. Piro

ABSTRACT

Sedimentation is the most common and effectively practiced method of urban drainage control in terms of operating installations and duration of service. Assessing the percentage of suspended solids removed after a given detention time is essential for both design and management purposes. In previous experimental studies by some of the authors, the expression of iso-removal curves (i.e. representing the water depth where a given percentage of suspended solids is removed after a given detention time in a sedimentation column) has been demonstrated to depend on two parameters which describe particle settling velocity and flocculation factor. This study proposes an investigation of the influence of some hydrological and pollutant aggregate information of the sampled events on both parameters. The Multi-Objective (EPR-MOGA) and Multi-Case Strategy (MCS-EPR) variants of the Evolutionary Polynomial Regression (EPR) are originally used as data-mining strategies. Results are proved to be consistent with previous findings in the field and some indications are drawn for relevant practical applicability and future studies.

Key words | CSOs (combined sewer overflows), data-mining techniques, Evolutionary Polynomial Regression, urban drainage, water pollutant

M. Carbone
P. Piro (corresponding author)
Department of Soil Conservation,
University of Calabria,
Italy
E-mail: *patpiro@dds.unical.it*

L. Berardi
D. Laucelli
Department of Civil and Environmental Engineering,
Technical University of Bari,
Italy

NOTATION

a	settling velocity of discrete particles [m/s]
a_j	$(j = 0, \ldots, m)$ regression parameters of EPR models
b	flocculation factor
CoD	coefficient of determination
CoD_{MCS}	coefficient of determination in MCS-EPR
E_i	efficiency corresponding to the i-th iso-removal curve
EMC	event mean concentration
EPR	Evolutionary Polynomial Regression
EPR-MOGA	Multi-Objective Evolutionary Polynomial Regression
EX	user-defined set of exponents in EPR models
$\mathbf{ES}(j,i)$	exponent of the ith candidate input in the jth term of EPR models
f	user defined function in EPR models
h	water depth in sedimentation column [m]
hp	rainfall depth [mm]
i_{avg}	average rain intensity recorded during the rain event [mm/h]
$i_{\max}$	maximum rain intensity recorded during the rain event [mm/h]
LC	the Liguori Channel
m	maximum number of terms in EPR models
M	mass fraction removed by sedimentation according to Stoke's law
MCS-EPR	Multi-Case Strategy for Evolutionary Polynomial Regression
t	detention time in sedimentation column [min]
TSS	initial concentration of total suspended solids [mg/l]
PDD	previous dry days
v_0	settling velocity of particles larger than d_0 [m/s]
v_i	settling velocity of particles smaller than d_0 [m/s]
$\mathbf{X}_i$	$(i = 1, \ldots, k)$ column vector of the ith candidate input

doi: 10.2166/hydro.2011.003

$\hat{\mathbf{Y}}$	vector of model predictions
Θ_H	detention time assumed for design purposes [min]
WWTP	waste water treatment plant

INTRODUCTION

The wide variety of pollutants contained in urban wastewater represents one of the most critical reasons for the long-term persistence of poor quality waters. Among these contaminants, inorganic substances, such as heavy metals (Pettersson 2002; Vaze & Chiew 2004; Characklis *et al.* 2005; Vallet *et al.* 2010), are difficult to treat without resorting to costly chemical-physical procedures. Their quantity depends on different factors, such as land use, population density, traffic intensity (Butler & Davies 2000). Such substances are present under particulate form because they are attached to the solid particles (TSS) in wastewater. Therefore, the removal of TSS is an essential procedure in order to reduce pollutant contents of receiving water bodies (Peavy *et al.* 1985).

In a previous study (Piro *et al.* 2007) the variability of dissolved chemical demand fraction for events observed in the Liguori Channel (Italy) provided considerable implications with regards to treatment design. In particular, the results showed that the selection of a treatment strategy involving a physical unit operation is required in order to remove particulates through sedimentation and clarification. The efficiency of these units depends, to a great extent, on the flow behavior through them (Maus & Uhl 2010) and on the settling characteristics of the suspended solids in the wastewater to be treated. In particular, it is necessary to determine the settling velocity of solid particles (Chebbo *et al.* 1998) in order to decide the detention time corresponding to the desired level of solid removal.

Among the TSS characteristics, particle terminal settling velocities are the key factors for design and are determined experimentally using a variety of procedures and devices, which can be classified as: (a) quiescent settling devices (for example, various types of settling columns) with liquid at rest; and (b) dynamic settling devices, in which liquid can flow or be subject to mechanically generated turbulence (Marsalek *et al.* 2006).

Actually, sedimentation is the natural method of removing suspended particles from wastewater since all solids requiring removal are heavier than water. Therefore, using gravity as the natural dividing force is the cheapest and most common separation (sedimentation) technique (Peavy *et al.* 1985).

To determine the sedimentation characteristics of a suspension and measure the settling velocities of discrete particles in diluted suspensions, an indirect method was devised by Camp (1946), who first introduced the concept of the settling column test procedure, some detailed descriptions of which can be found in common environmental engineering handbooks (for example, Metcalf & Eddy 2003).

Nevertheless, few studies have dealt with settling column tests (Weber 1972; Zanoni & Blomquist 1975; Berthouex & Stevens 1982; Eckenfelder 1989; Oke *et al.* 2006) or with the relationship between experimental results and design criteria of sedimentation tanks.

Piro *et al.* (2011a) described the iso-removal curves (which represent the water depth where a given percentage removal of suspended solids is achieved after a given detention time) in a settling column for combined sewer overflows (CSOs) in both wet-weather and dry-weather conditions. The resulting model expression is a power function depending on two parameters that entail settling velocity of discrete particles (a) and flocculation factor (b). They suggested using the column tests to determine these parameters as features of the diluted suspension.

The aim of this paper is to investigate the possible dependence between some aggregate information on the sampled event and the two parameters (a) and (b). Such information pertains to some aggregate hydrological and pollution data which are supposed to affect re-suspension of solids during the runoff and settling in sedimentation columns.

From a practical standpoint, the possible dependence of settling characteristics from such aggregate information might provide useful advice for design purposes (e.g. residence time and thus volume of sedimentation tanks) based on some typical values of the analyzed area. On the other hand, for existing treatment plants, it might support decisions about possible enhancement works such as construction of some additional sedimentation tanks, or more

efficient management practices (e.g. opening/closure of additional sedimentation units to tune the detention time) based on a few and easily measurable variables. The investigation reported here is also aimed at supporting future research on modeling sedimentation features based on easily retrievable effluent information.

The intended analysis includes the description of sedimentation features as result of very complex mechanisms involving washing of pollutants from surface, re-suspension of sediments, transport and sedimentation by using simple aggregate information. Such relationships can be hardly deduced by using classical physical modeling since they imply a number of concurrent phenomena under different possible boundary conditions. In addition, not all available information might be useful in describing the target but rather just a subset of them, thus implying a combinatorial nature of the analysis. As a consequence, this work leverages a data-mining approach to achieve additional knowledge about the relationship between available aggregate information and sedimentation features assumed as input/output data from the system in hand.

Currently, a number of data-driven techniques are available for developing models from data; from among them the Evolutionary Polynomial Regression (EPR) (Giustolisi & Savic 2006) has been proved effective to identify patterns in various applications entailing the exploration of a combinatorial space of possible alternatives (Berardi *et al.* 2008; Markus *et al.* 2010; Rezania *et al.* 2010). The analysis performed here exploits two recent variants of the EPR, namely the Multi-Objective EPR (EPR-MOGA) (Giustolisi & Savic 2009) and the Multi-Case Strategy for EPR (MCS-EPR) (Berardi & Kapelan 2007; Giustolisi *et al.* 2009). Furthermore, this paper presents an original way to use the Multi-Objective EPR modeling paradigm for analyzing possible dependence between candidate explanatory variables and the target attribute. Accordingly, a methodology to use the EPR for data-mining purposes rather than developing complete model expressions is provided.

Background on data

Data used for the analyses have been sampled from the catchment of the Liguori Channel (LC) which is located in the town of Cosenza (southern Italy). The catchment has an area of 414 Ha, 48% of which is densely urbanized, while the remaining part is largely covered with natural vegetation. The relevant urban area has a population of 50,000 inhabitants. Figure 1 shows a plan view of the catchment.

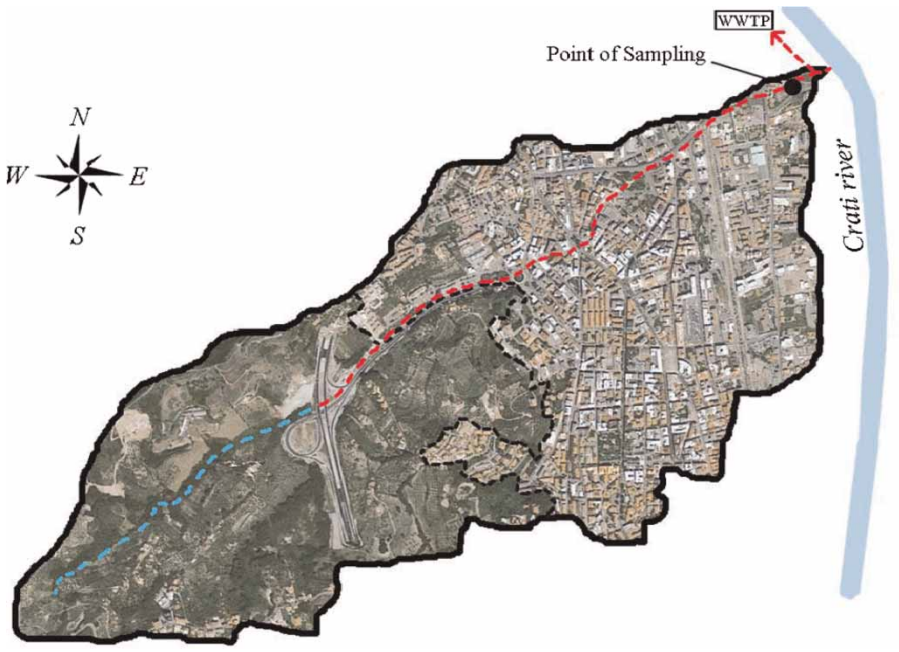

Figure 1 | Monitored catchment of Liguori Channel (LC).

The catchment is drained by a combined sewer system that conveys the dry-weather flows to the waste water treatment plant (WWTP) at Montalto Uffugo, a small town near Cosenza.

During heavy rainfall events, the sewer flow exceeds the capacity of the sewer system and WWTP; the excess flow is discharged directly into the Crati River (see Figure 1) through overflow drop structures, without any treatment.

The sampling campaign which this research work refers to was carried out between autumn 2007 and autumn 2009. In particular, the analysis refers to nine wastewater samples that were collected during as many events in wet-weather conditions, and for each one a settling column test was performed.

Each sample was collected at the outlet of the overflow drop structures during the rainfall event (although at different times for different events) and consisted of 400 L. The column tests were performed only after mixing the whole volume, thus, for each sample the event mean concentration (EMC) was measured while detailed information about possible concentration variability during the same event (e.g. the first flush phenomenon) was not recorded.

Since a minimum diameter of 12.7 cm is recommended to minimize wall effects (Eckenfelder 1989), the column used in this work was a stationary settling column of 150 mm in diameter and 3 m in height. Such height value was assumed to reflect typical depth adopted for sedimentation tanks in the analyzed area (Piro *et al*. 2011b).

Each settling test consisted of determining the residual concentration of suspended solids in the wastewater sampled every 5 min at five orifices equally spaced along the column. These concentrations are then used to compute percentage of mass fraction removed at each depth and for each detention time. The percentage removal values obtained from the test data are plotted at the appropriate depths and times, and the iso-lines of percentage removal (i.e. the iso-removal curves) (Zanoni & Blomquist 1975) are constructed by interpolating plotted values (see Figure 2). Thus, such curves represent the limiting or maximum settling path for the indicated percentage (Eckenfelder 1989).

Piro *et al*. (2011a) recently found that the pattern of iso-removal curves can be described by using a power law,

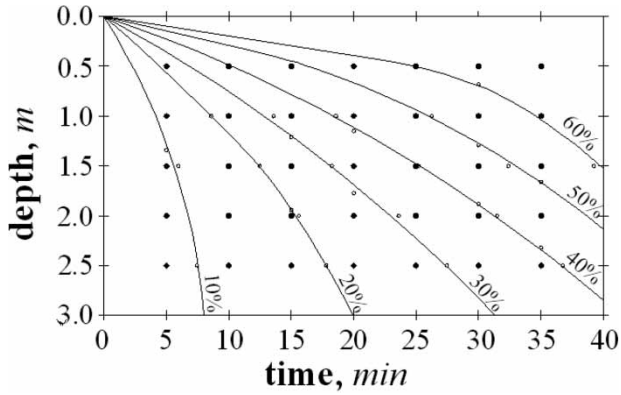

Figure 2 | Example of iso-removal curves from settling analysis.

represented in the following equation:

$$h = at^b \tag{1}$$

where h is the depth, t the residence time, a the particle settling velocity and b represents the flocculation factor. Those analyses demonstrated that the simple Equation (1) results in an exceptionally accurate description of iso-removal curves on both experimental and literature data. In particular, those experiments were carried on samples coming from the LC, the same catchment considered here during both wet-weather and dry-weather conditions.

Although the mathematical definition of the iso-removal curves allows easy and accurate calculation of the removal efficiency of a sedimentation unit, a relatively large variability of settling velocity a and flocculation factor b was observed among different sampled events. Table 1 reports the values of both parameters a and b estimated for the iso-removal curves corresponding to 10% up to 50% removal of suspended solids for nine wet-weather events. Data missing at 10 and 50% removal reflect the lack of relevant experimental data as a consequence of the space-time grid used to estimate the experimental point of the removal efficiency (depth step 0.5 m; sampling every 5 min). In the case of 10% removal, this was due to the faster early sedimentation which resulted in erroneous measurements performed at the top of the column; in the case of 50% removal the omission was the result of the short duration of the test (40 min) thus preventing the estimations of points referring to the lower side of the column.

Therefore, the aim here is to provide an analysis of possible dependence between such parameters and some

Table 1 | Settling velocity (*a*) and flocculation factor (*b*) parameters of iso-removal curves

	Removal rate									
	10%		20%		30%		40%		50%	
Event	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
1	0.0395	1.4965	0.0040	1.9742	0.0021	1.9223	2.2×10^{-11}	6.6917	–	–
2	–	–	0.0053	2.1410	0.0006	2.5419	4.2×10^{-05}	3.1203	1.2×10^{-07}	4.5408
3	0.0700	1.2731	0.0008	2.6361	1.5×10^{-5}	3.5775	4.5×10^{-07}	4.2904	0.0054	2.8841
4	0.1290	1.0000	0.0350	1.2273	0.0054	1.6423	0.0005	2.2123	–	–
5	0.0106	2.7938	0.0307	1.4787	0.0168	1.4534	0.0022	1.8681	–	–
6	0.0001	3.4840	1.9×10^{-7}	5.3090	1.1×10^{-09}	6.5461	4.3×10^{-12}	7.8619	1.2×10^{-14}	9.1882
7	–	–	0.0795	1.1427	0.0004	2.6517	0.0005	2.2804	3.2×10^{-06}	3.5537
8	0.1694	1.1031	0.0542	1.2656	0.0239	1.344	0.0221	1.2490	–	–
9	–	–	0.0005	3.0623	0.0002	3.1302	0.0011	2.1359	–	–

common and easily measurable hydrological and pollutant variables which characterize the different rainfall–runoff events and relevant effluent.

In fact, it is well known that both settling velocity and flocculation are mainly affected by particle size. However, the comparison between the grain size found in wastewater samples in dry-weather and wet-weather conditions in the LC catchment confirms the general observation that the range of variation of particle size in combined sewer systems is likely to increase during rainfall events (Piro *et al*. 2010). This is due to the combination of two concurrent phenomena: (1) sediment transport from the catchment into the sewer system; and (2) sediment scratching from pipes (and re-suspension) due to the increased flow rate during runoff.

From such observation it can be hypothesized that the heavier the rainfall, the sharper the runoff peak flow and the re-suspension of solids is expected to be. In addition, the longer the number of previous dry days, the larger is the mass of pollutant expected to be flushed off from catchment surface and from culverts. All these considerations motivated the analysis of those nine samples collected during wet-weather conditions for which only the following aggregate variables were available whose meanings are reported in the notation section: TSS, previous dry days (PDD), hp, i_{max}, i_{avg}. Table 2 reports relevant values for the nine events.

From Table 2, it is evident that samples pertain to a wide range of rainfall events. In particular, events numbered 2, 5 and 7 are the most severe in terms of maximum and average intensity as well as rainfall depth. As reported above, for some of them (events 2 and 5), the number of PDD seems to lead to the highest TSS values. Nonetheless, the opposite cannot be said; in fact, for events 7 and 9 large values of TSS have been registered after only 1 PDD; thus, it is impossible to determine any trivial univocal trend from these data.

Table 2 | Values of event-specific aggregate variables

Event	Date day/month/year	TSS [mg/l]	i_{avg} [mm/h]	i_{max} [mm/h]	hp [mm]	PDD
1	02/04/2008	49.5	1.13	3.0	3.4	5
2	13/01/2009	138.5	2.13	7.4	36.2	2
3	18/02/2009	86.0	1.73	3.2	13.8	1
4	11/03/2009	82.0	1.56	3.2	17.2	4
5	20/03/2009	237.0	2.58	6.2	20.6	4
6	21/04/2009	32.0	1.67	3.6	5.0	1
7	28/04/2009	102.5	2.45	9.4	36.8	1
8	22/09/2009	65.0	0.30	0.4	0.6	1
9	23/10/2009	131.0	0.49	1.0	3.4	1

It is also worth noting that events have been recorded in different seasons and not in consecutive days; thus, events are independent from each other as well as the aggregate variables to be used for next analysis. The only exception holds for average and maximum rainfall intensity (i_{avg} and i_{max}) that show a similar behavior in eight out of nine events (the only exception is event number 8 where rain intensity was almost uniform). Nonetheless, i_{max} and

i_{avg} are far from being proportional over all samples and might be hypothesized to independently affect the size of particles re-suspended during the runoff, thus the aggregate data used includes both i_{max} and i_{avg}.

The Evolutionary Polynomial Regression (EPR) framework

Starting from the eighties, the advent of information technology as well as the pervasive use of personal computers has made available a number of data in many different scientific areas. A key issue is to effectively extract information from stored data. Nowadays, a number of methodologies and tools are available that resort too many different approaches ranging from classical statistical inference to artificial intelligence techniques. All of them are usually referred to as 'data-mining' techniques since they are aimed at extracting information from data (Fayyad *et al.* 1996). Nonetheless, in many practical applications such as civil engineering, the physical interpretation of some phenomena is almost completely known even as mathematical models, while the lack of knowledge about systems/phenomena is put into numerical parameters. In such cases information from data might provide additional elements of knowledge and/or support future data collection strategies.

Among data-mining techniques, data-driven modeling entails the development of mathematical models based on data. The main driving criterion for developing such models is the accuracy in reproducing recorded data, while just a few data-driven modeling techniques aims at providing mathematical expression which can also be interpreted from a physical point of view. Actually, the interpretation of mathematical expressions is one of the main validation criteria to be accounted for when selecting from among different models describing the same phenomenon (Domingos 1999; Ljung 1999). Such an observation motivated the development of some techniques based on the so called Genetic Programming (GP) (Koza 1992). It consists of developing symbolic expressions by using evolutionary algorithms to search among a population of possible combinations of mathematical operators, candidate arguments (variables) and parameters. Moving from the classical GP paradigm, in recent years the EPR has been adopted in a few applications (Giustolisi *et al.* 2007; Berardi *et al.* 2008; Markus *et al.* 2010; Rezania *et al.* 2010).

In brief, the expressions achievable by EPR are basically made of a number of additive terms multiplied by as many coefficients (i.e. as with polynomials), as reported in the following general expression:

$$\hat{\mathbf{Y}} = a_0 + \sum_{j=1}^{m} a_j \cdot (\mathbf{X}_1)^{\mathbf{ES}(j,1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,k)} \times f((\mathbf{X}_1)^{\mathbf{ES}(j,k+1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,2k)}) \quad (2)$$

where m is the maximum number of additive terms, $\mathbf{X}_i$ and $\hat{\mathbf{Y}}$ are model input and output variables, function f is chosen by the user, and exponents of variables [i.e. $\mathbf{ES}(j,i)$, see previously referenced works for details (e.g. Giustolisi & Savic 2006)] are selected from a set EX of candidates defined by the user. The additive terms that constitute the mathematical structure of models are obtained by means of a genetic algorithm which selects exponents from EX, while parameters are estimated using classical numerical regression (e.g. least squares). Thus, the search for non-linear models is based on an integer coding of possible alternatives, while final expressions are linear with respect to coefficient a_j. It is worth noting that, if the set of exponents contains zero and $\mathbf{ES}(j,i) = 0$, the relevant input disappears from the final expression; thus, although simple, structures like Equation (2) are quite versatile and flexible enough to reproduce patterns in data.

A recent upgrade of the EPR encompasses a multi-objective optimization strategy (i.e. EPR-MOGA) where accuracy of data reproduction and parsimony of model structures are simultaneously maximized. Accuracy is evaluated in terms of coefficient of determination (CoD):

$$\mathrm{CoD} = 1 - \frac{\sum_N (\hat{y} - y_{exp})^2}{\sum_N (y_{exp} - \mathrm{avg}(y_{exp}))^2} \quad (3)$$

where N is the number of samples, $\hat{y}$ is the value predicted by the model, and $\mathrm{avg}(y_{exp})$ is the average value of the corresponding observations (evaluated on the N samples).

Parsimony refers to the number of variables and/or additive terms involved in the mathematical expressions and its minimization is assumed to result into more general

description of the phenomenon while allowing its physical readability.

The EPR-MOGA paradigm is based on a global search within the space of model expressions, such space is defined by the user in terms of base structure of mathematical expressions (e.g. as in Equation (2), function f and maximum number of additive terms m), the cardinality of set EX of candidate exponents and number of candidate explanatory variables. The number of generations set for the multi-objective genetic algorithm used (i.e. OPTIMOGA, see Laucelli & Giustolisi (2011) for details) is proportional to all these factors in order allow for a sufficient exploration of the space. However, the actual number of function evaluations is not necessarily proportional to the number of generations due to the efficient management of a dynamic archive of optimal individuals performed by OPTIMOGA.

The advantages of the EPR-MOGA are that: (1) it allows developing a Pareto set of models with different accuracy and parsimony in a unique modeling run; (2) the possible similarities between returned expressions allow for discussing and interpreting the description of the phenomenon; and (3) the set of models is aimed at supporting the user to select the expression suited for the peculiar intended analysis.

The most recent version of EPR entails the MCS-EPR. It adopts the same evolutionary strategy of EPR-MOGA for developing mathematical expressions, while the assessment of model parameters (i.e. $a_{j,s}$ with $s = 1, \ldots, C$) and the evaluation of accuracy refer to C separate cases/experiments simultaneously. Actually, such cases/experiments reflect user-defined partitioning of the available data based on the hypothesis that all of them refer to the same phenomenon. Thus, resulting model expressions (i.e. sets of exponents) actually hold for every individual case, although different parameters take charge of different error realization in each subset of data.

The following measure of fitness to data is used in MCS-EPR instead of Equation (3):

$$\mathrm{CoD}_{\mathrm{MCS}} = 1 - \frac{\sum_{s=1}^{C} \sum_{N_s} (\hat{y}_s - y_{\exp})^2}{\sum_{N} (y_{\exp} - \mathrm{avg}(y_{\exp}))^2} \tag{4}$$

where N_s is the number of samples in the s-th case/experiment (i.e. $N = \Sigma N_s$), C is the number of cases, $\hat{y}_s$ is the model prediction using coefficients $a_{j,s}$ and $y_{\exp}$ is the corresponding observation.

Such formulation is particularly valuable when very few data are available for each case/experiment or data partitioning is unbalanced among different cases. This is also consistent with the observation that larger data subsets (i.e. large N_s) allow being more confident on the final model (Ljung 1999). Similarly to the CoD in EPR-MOGA, the closer $\mathrm{CoD}_{\mathrm{MCS}}$ is to 1 the more suitable the model structure is in describing the overall observed data.

It is worth noting that the same measure of parsimony as EPR-MOGA is adopted in MCS-EPR (i.e. number of variables and/or additive terms involved in the mathematical expressions).

Figure 3 provides a comparison of EPR-MOGA and MCS-EPR algorithms; grey boxes emphasize the key differences.

Mining data by EPR-MOGA and MCS-EPR

Although both EPR-MOGA and its MCS variant have been proved to be useful for developing models in different applications, this paper proposes exploiting their paradigm to analyze the relative influence of each variable in describing the output, without necessarily achieving a final model expression. The practical implication of this approach is twofold: on one hand it investigates which are the most meaningful variables (if any) from among those available; on the other hand it might support next more effective data collection to model the phenomenon in hand.

To this end, suppose that a hypothetical model has to be developed made of one (polynomial) term only (i.e. $m = 1$),

$$\hat{\mathbf{Y}} = a_0 + a_1 \cdot (\mathbf{X}_1)^{\mathbf{ES}(1,1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(1,k)} \tag{5}$$

Then, EPR-MOGA is run so that accuracy to target output is maximized while minimizing the number of input variables (i.e. $\mathbf{X}_i$) involved in the final mathematical expression. The resulting Pareto set of models is expected to show a progressively increasing number of input variables and accuracy to training data. It is worth noting that such analysis does not require any prior assumption about the candidate exponents of variables but just some values

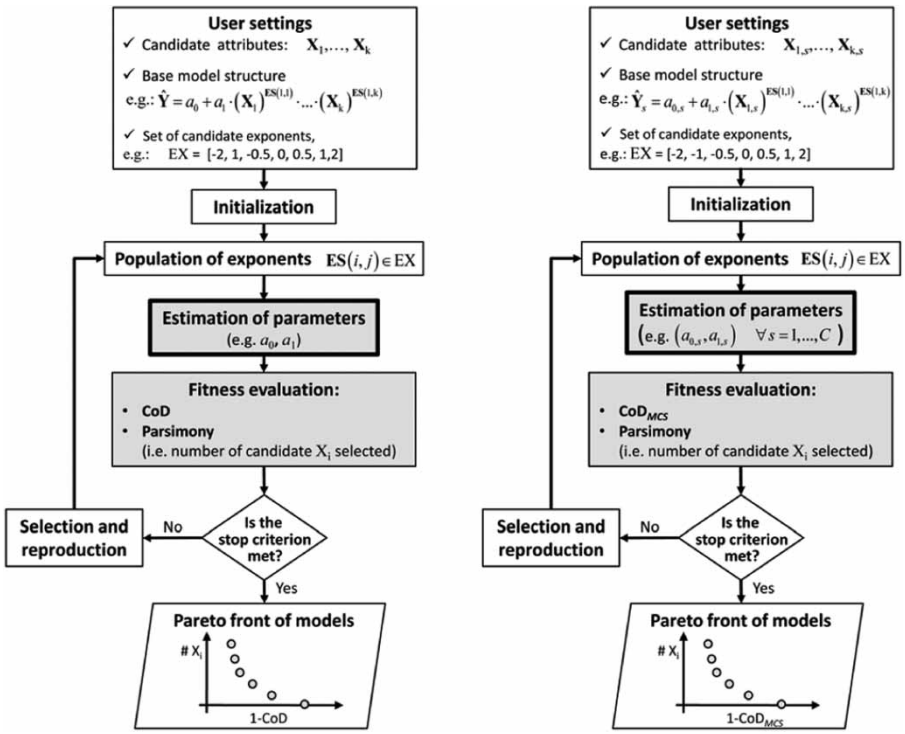

Figure 3 | EPR-MOGA (left) and MCS-EPR (right) flow charts.

entailing direct/inverse and linear/non-linear relationships. This way, resulting models allow arguing the relative importance of each explanatory variable in reproducing the target attribute (model output) and the type of possible relationship.

From a data-mining standpoint, the variable(s) selected first (i.e. in the most parsimonious expression) and, in particular, that (those) leading to a significant improvement of accuracy is (are) likely to be the most meaningful to explain the target variable among all candidates. The persistence of one or more variables in the same direct/inverse fashion among different models is a further point to assert its (their) influence. On the contrary, those variables selected in the most accurate (and less parsimonious) models only and/or showing both direct and inverse dependence from output, or even appearing in few models, only can be assumed to be not that informative about the phenomenon.

Confidence about such conclusions descends also from the observation that different models are obtained independently from each other during the global exploration of the search space.

A similar analysis can be repeated by using the MCS-EPR. In fact, it permits discovering direct/inverse relationships of quite general validity without incurring possible misleading conclusions due to the particular error realization of the single experiments. Indeed, the use of MCS-EPR for data-mining purposes allows the confirming/discussing of previous findings coming from separate EPR-MOGA analyses on individual cases/experiments. When non-unique relationships are obtained from both single EPR-MOGA and MCS-EPR analyses, then variables considered are likely to be non-correlated and, eventually, this indicates that a different set of candidate explanatory variables should be considered. Moreover, the model accuracy achieved on different cases also allows confirmation of such conclusions.

ANALYSES AND RESULTS

Two analyses are carried out considering parameters a and b separately. The aggregate data reported above (i.e. TSS, PDD, hp, i_{max}, i_{avg}) are assumed to be potential explanatory

variables for describing the variation of a and b among the nine samples (events). In order to do this, five different runs of the EPR-MOGA have been performed for a and b, respectively, each pertaining an iso-removal curve (i.e. 10, 20, 30, 40 and 50%). Thus, for each iso-removal curve two hypothetical models are developed which describe $a = a(\text{TSS, PDD, hp, } i_{max}, i_{avg})$ and $b = b(\text{TSS, PDD, hp, } i_{max}, i_{avg})$. Each run is performed considering Equation (5) (i.e., $m = 1$ in Equation (3), no function f), and a set of candidate exponents $ES = [-3, -2.5, -2, -1.5, -1, -0.5, 0, 0.5, 1.5, 2, 2.5, 3]$. Although they do not descend from any prior physical consideration, the exponents in EX here are representative of direct/inverse relationships only while allowing an acceptable degree of accuracy in the final expression. The estimation of the offset parameter a_0 is also allowed at this stage as it is assumed that other variables might be considered for achieving a more accurate description of a and b.

Afterwards, two MCS-EPR runs are performed considering the values of parameters a and b of 10, 20, 30, 40 and 50% iso-removal curves as different cases/experiments governed by the same underlying relationships with candidate explanatory variables. The scope of this analysis is twofold. On one hand, it allows to confirm/controvert findings from previous EPR-MOGA analyses on individual cases. On the other hand, it aims at discovering the most informative aggregate variables to explain settling velocity and flocculation factor over all removal stages during sedimentation (as represented by the five iso-removal rate curves). The same search settings of the individual EPR-MOGA runs have been applied for MCS-EPR analyses.

Models obtained from both EPR-MOGA and MCS-EPR are discussed in the following by neglecting numerical coefficients (i.e. a_0 and a_1) since the scope of this analysis is not to provide a final model expression but rather to mine information from data.

Flocculation factor (b)

The expressions returned for the flocculation factor b of the five iso-removal curves are represented in Figure 4. Each diagram reports a number of possible expressions each containing an increasing number of variables and showing increasing accuracy. All diagrams reports on the lower-right corner a point representing the average value of parameters b of the iso-removal curves over all samples: it is the less accurate but, of course, the most parsimonious model structure (none inputs selected). It is evident that including one or more input variables results into more accurate reproduction of b.

The most informative explanatory variable is TSS for almost all iso-removal curves. It is reported in all models but the second model of the 10% case, where i_{max} and hp are selected.

From the analysis of the Pareto fronts of expressions, it is also evident that TSS allows for a significant improvement of accuracy (i.e. reproduction of the target b) in all iso-removal cases. On the contrary, the inclusion of additional variables beyond TSS results into a marginal accuracy improvement. From a data-driven perspective this means that the additional variables do not actually improve the main description of the target (i.e. model output) but are selected to slightly improve the fit to data. This general advice is somehow confirmed by the alternation of direct and inverse dependence on the remaining variables over the different iso-removal cases.

The MCS-EPR analyses returned the following model expressions (beyond the trivial constant value) which confirm the previous EPR-MOGA finding about the influence of TSS.

$$
\begin{aligned}
b &= a_0 + \frac{a_1}{\text{TSS}^3} \\
b &= a_0 + a_1 \frac{i_{max}^{0.5}}{\text{TSS}^2} \\
b &= a_0 + a_1 \frac{i_{max}^3}{\text{TSS}^2 \cdot \text{hp}^{1.5}} \\
b &= a_0 + a_1 \frac{i_{avg}^{1.5} \cdot i_{max}^{1.5}}{\text{TSS}^{1.5} \cdot \text{hp}^{1.5}}
\end{aligned}
\tag{6}
$$

It is worth noting that the flocculation factor b is found to be inversely dependent on the concentration of TSS. Clear explanations of this can be found in literature about the wet-weather flow characteristic (Lin *et al.* 2009). In fact, the particulate matter transported in rainfall-runoff processes at the urban surface is largely inorganic, with a specific gravity in the range of 2.3 to 2.7 g/cm^3, and a volatile fraction generally less than 30% (Sansalone *et al.* 1998;

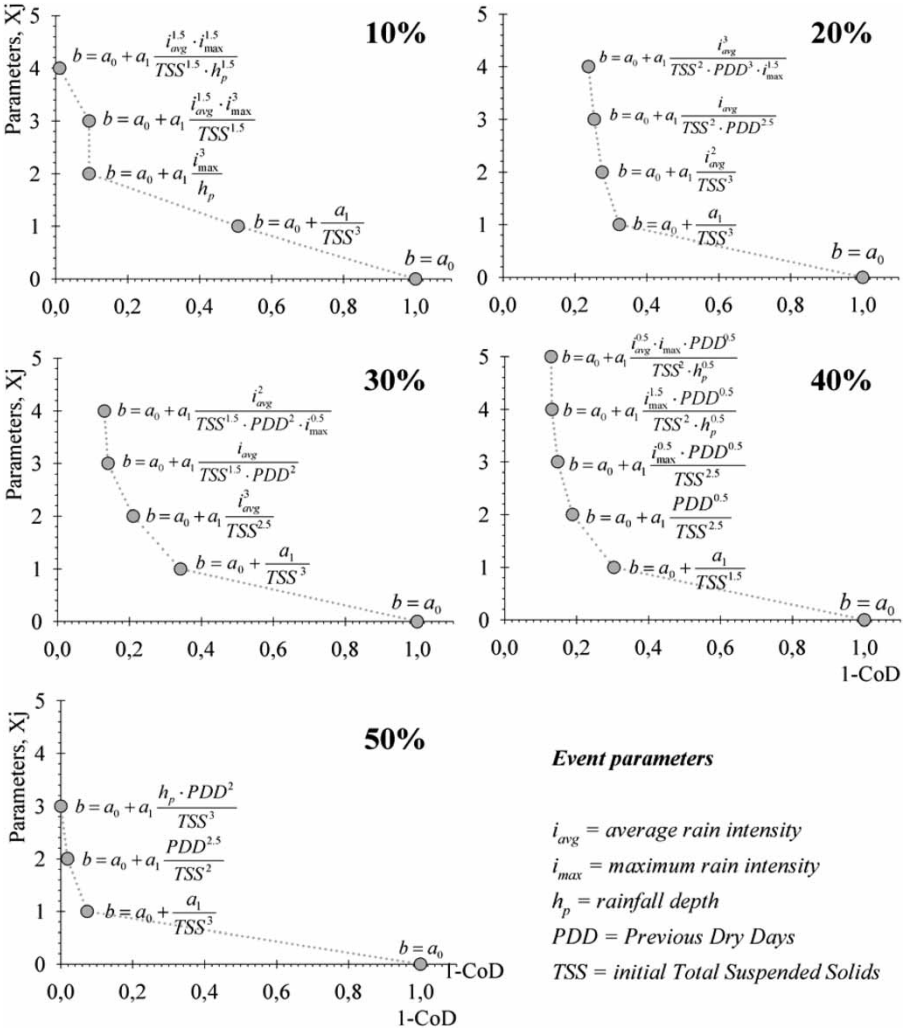

Figure 4 | Pareto's fronts of models for flocculation factor (*b*).

Ying & Sansalone 2008). These conditions generate a very hetero-disperse and inorganic gradation of particulate matter for source area discharges, which is in turn less prone to flocculation. The increase in TSS from one event to another is, therefore, mainly due to the contribution of particles that come from the surfaces scouring, which decreases the propensity to flocculation of the water solution.

The MCS-EPR analysis emphasizes that the number of PDD = does not provide any information if all removal rates (i.e. the whole settling process) are analyzed together.

As final remark, the same analyses have been repeated by forcing $a_0 = 0$ and results mainly confirm the same conclusions about the inverse dependence between TSS and *b*, although with different exponents and less accurate models. For the sake of brevity such further results are not reported herein, but they somehow confirm that additional information, beyond those provided by the available data, might improve the description of the target.

Settling velocity (*a*)

About the settling velocity parameter, the analysis of returned expressions does not allow a unique conclusion for all iso-removal curves to be drawn (see Figure 5). On one hand, the 10, 40 and 50% iso-removal cases shows the

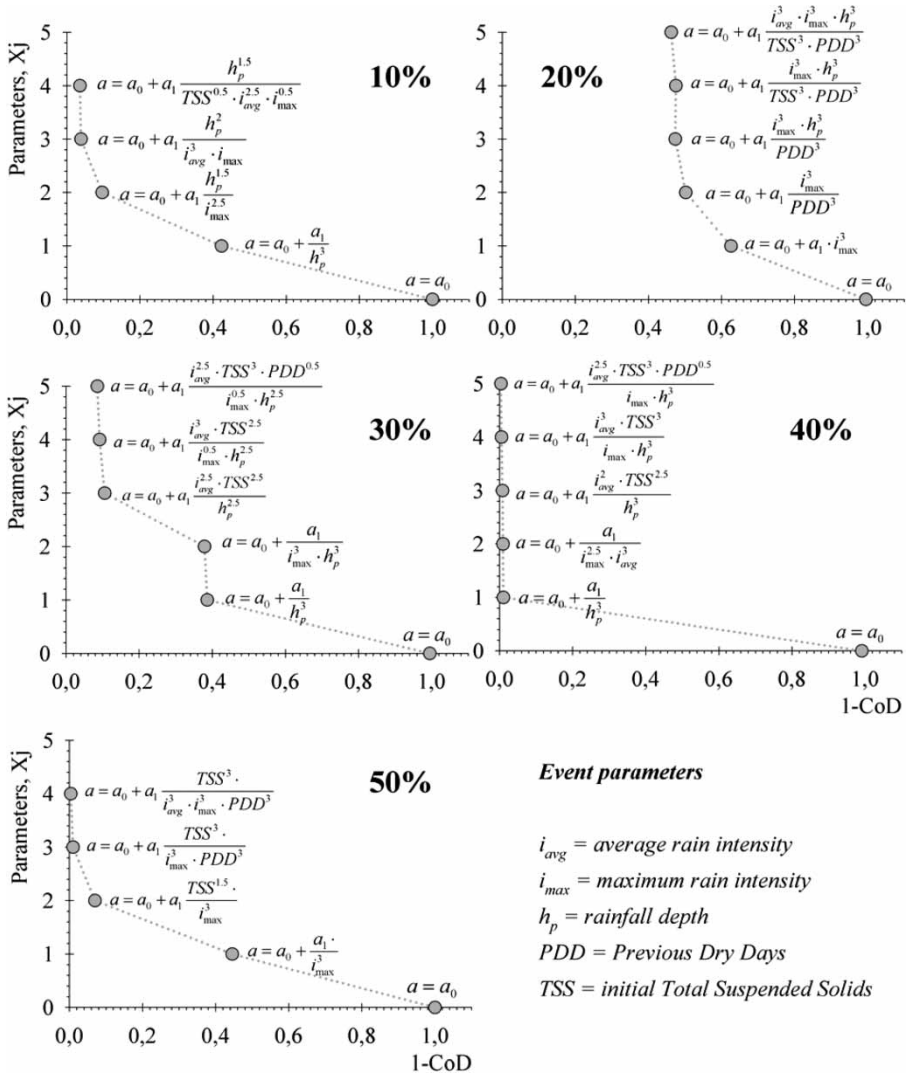

Figure 5 | Pareto's fronts of models for settling velocity (*a*).

inverse dependence on maximum rainfall intensity $i_{\max}$ as the most important explanatory variables. This would lead us to hypothesize that during severe rainfall events the flow through the channel scratches out large size sediments, which in turn settles down quickly. The relation between *a* and hp is quite ambiguous even among different models of the same iso-removal case. However, somehow hp also seems to be informative about the variation of *a* since it is selected as first in three out of five cases.

On the other hand, the 20% iso-removal case describes the opposite situation, and the number of PDD seems to be also important. Also in this case, it could be argued that the higher the number of dry days, the more heavy sediments (i.e. with higher settling velocity) are likely to be settled along the pipes. Nonetheless, the analysis of the accuracy of returned models shows very inaccurate reproduction of parameter *a* for 20% iso-removal case. From a data-mining perspective this suggests that the available information (i.e. aggregate parameters) is not sufficient to describe the variation of settling velocity. Moreover, it should be remarked that, from present analysis the initial concentration of TSS is clearly one of the less informative aggregate variable for explaining the variation of settling velocity.

The MCS-EPR models obtained by considering the five iso-removal cases for parameter a are different from previous EPR-MOGA models, as reported in Equation (5). In addition, the dependence from hp is not univocal and they are quite inaccurate on relevant subsets of data (i.e. on each iso-removal case). Vice-versa, $i_{\max}$ was found to be inversely proportional to a, although it is not the most significant variable.

$$\begin{aligned} a &= a_0 + \frac{a_1}{\mathrm{hp}^3} \\ a &= a_0 + a_1 \frac{\mathrm{hp}^{1.5}}{i_{\max}^{2.5}} \\ a &= a_0 + a_1 \frac{\mathrm{hp}^2}{i_{\mathrm{avg}}^3 \cdot i_{\max}} \\ a &= a_0 + a_1 \frac{\mathrm{PDD}^{0.5} \cdot \mathrm{hp}}{\mathrm{TSS}^{0.5} \cdot i_{\max}^2} \end{aligned} \tag{7}$$

From a physical perspective, this behavior might suggest that the rainfall intensity (both average and maximum), the rainfall depth and the number of PDD do not entail a unique explanation for particle size distribution. In fact, the parameter a is the settling velocity of the particles with a specific size at time zero; thus, to correctly estimate its value, knowledge of the specific physical and geometric characteristics of the particles is required (Metcalf & Eddy 2003). Actually such characteristics are better described by particle size distribution and the particles weight rather than by aggregate parameters like the TSS. In fact, for example, in sedimentation Type I (every particle settles independently) the particle size distribution is required to compute the settling velocity distribution by using Stoke's law; consequently the total mass fraction removed by sedimentation is computed as:

$$M = 1 - x_0 + \int_0^{x_0} \frac{v_i}{v_0} \mathrm{d}x \tag{8}$$

where $(1 - x_0)$ is the fraction of particles with settling velocity greater than v_0 (corresponding to size larger than d_0), and the integral is the fraction of particles removed according to the ratio v_i/v_0, with v_i settling velocity corresponding to the particles smaller than d_0.

From a modeling perspective, this analysis confirms that further investigation are needed to achieve meaningful relations between parameter a and particle size distribution to better describe the variation of the iso-removal curves by event.

As a side achievement, this result shows that the combined use of EPR-MOGA and MCS-EPR allows for robust data-mining that helps to avoid misleading conclusions. Also in this case, the analysis performed by imposing $a_0 = 0$ neither improved the description of the target a nor provided any additional insight.

Practical implications

It was observed that the CSOs involve flocculating particles and that the Stoke's equation cannot be used to design clarifiers as flocculating particles are continually changing in size and shape. Thus, the criteria adopted to design clarifiers have evolved in both practice and theory in order to account for many factors which contribute to the flocculation process. Usually the settling column test (Peavy *et al.* 1985) is used to estimate the removal efficiency of TSS and the detention time of the flocculating solution, but this approach is often time consuming and expensive for practical design purposes.

The iso-removal lines allow determining the fractions of particles that are completely removed from the column, i.e., the particles with diameters $d \geq d_0$, for a given detention time $t = \Theta_H$. Nevertheless, the total removal efficiency will be greater because also finer particles (with dimensions $d < d_0$ and settling velocity $v < v_0$) are partially removed. The results of the column test can be used to assess the total removal efficiency (E_{tot}) of sedimentation process (Metcalf & Eddy 1991), as:

$$E_{\mathrm{tot}} = E(\Theta_H) + \sum_i (E_i - E_{i+1}) \cdot \frac{h_{i,i+1}}{H} \tag{9}$$

where H is the column depth, $E(\Theta_H)$ is the constant percent removal curve passing through point (Θ_H, H), E_{i} and $E_{\mathrm{i+1}}$ are the iso-removal efficiency greater than $E(\Theta_H)$ and $h_{i,i+1}$ is the depth of the middle point of the segment joining E_i

and E_{i+1} curves at $t = \Theta_H$. Figure 6 shows how single contributions are obtained for iso-removal curves.

It is evident that this approach is time consuming and expensive for practical design purposes, especially taking into account the high variability of the settling process in single events. The knowledge of the analytical relationships for each iso-removal curve (i.e. $h_i = a_i t^{bi}$), allows us to write Equation (9) as:

$$E_{\text{tot}} = E(\Theta_H) + \sum_i (E_i - E_{i+1}) \cdot \frac{h_i + h_{i+1}}{2H} =$$

$$= E(\Theta_H) + \frac{1}{2H} \sum_i (E_i - E_{i+1}) \cdot (a_i \Theta_{\text{H}}^{b_i} + a_{i+1} \Theta_{\text{H}}^{b_{i+1}}) \qquad (10)$$

where the parameters a_i and b_i are assumed to take charge of the variability of the settling behavior due to effluent characteristics.

The data-mining methodology used herein allows investigating the possible influence of some aggregate hydrological and pollution indicators on the variability of a_i and b_i. In addition, it might support the development of some concise mathematical relationships of a_i and b_i for each iso-removal curve which can be easily included in Equation (10).

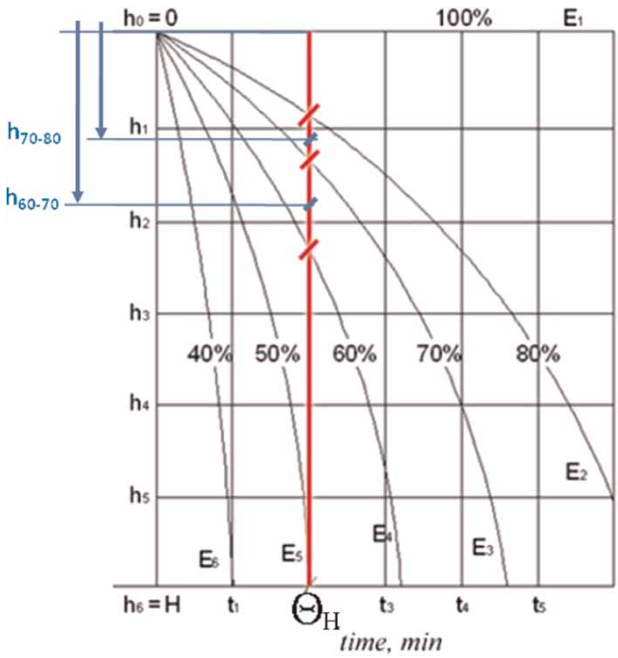

Figure 6 | Schematics methodology to compute the total removal efficiency for the fixed detention time Θ_H.

The results showed the presence of relationship between the sedimentation characteristics of the CSOs and the more general hydrological and pollutant aggregate information, such as rainfall intensity and the concentration of TSS. However, future experimental studies on different catchments are expected to improve the mathematical relationships between the used parameters in order to adapt the methodology to different local contexts.

From design perspective, Equation (10) can easily be used to assess the total removal efficiency as a function of the detention time (i.e. flow rate) and incoming pollutant aggregate information (i.e. TSS concentration), once the parameters of the iso-removal curves are estimated from the peculiar characteristics of the drained catchment. The assessment of the detention time required to achieve a sufficient particle removal, allows determining the volume of clarifiers and, in turn, permits choosing the best treatment type.

From a treatment plant management perspective, the TSS removal efficiency can be enhanced by using coagulant additives, such as polymers or aluminum salts (Li *et al.* 2003). As a consequence, the possible assessment of expected total removal efficiency based on the conditions of detention time and concentration of pollutants (TSS) is a crucial point for developing automatic systems for setting (e.g. in real time) the concentration of clotting agents based on the peculiar event. Similarly, this approach can be used to assess the air to blow for air bubble generation in the clarifier or to adjust the rotation speed of the stirring in the mixing reactor. Such systems would ultimately allow savings of energy and coagulant additives.

CONCLUSIONS

The treatment and management of the CSOs need the knowledge of sedimentation characteristics; when retained for clarification, larger sediment and settleable particles are mainly influenced by gravitational forces, while the suspended particles are subject to coagulation phenomena. A relationship between sedimentation characteristics and easily measurable parameters represents a powerful tool in design and management of treatment processes.

The present work proposes an investigation into the relationships between some hydrological and pollution aggregate variables, gravitational forces and flocculation indicators based on a recent modeling approach for the sedimentation analysis developed by some of the authors.

The data-driven modeling paradigm of EPR is proposed as an original way to perform the analysis of available information. Such data-mining approach takes advantage from the multi-objective strategy underlying both EPR-MOGA and MCS-EPR. It allows information retrieval about the influence of candidate explanatory variables on the target (i.e. model output) by analyzing a Pareto set of simple (monomial) expressions. While doing so, two aspects need to be taken into account in order to improve the generalization of results (Ljung 1999): the data available and the introduction of field expert knowledge. Data should refer to typical system conditions in order to be representative of actual system behavior and increasing the number of samples may improve the general validity of conclusions. On the other hand, the data-mining methodology should allow the analyst to clearly read results in order to facilitate elicitation of expert knowledge. In this study, the few data available are representative of typical system behavior since they neither pertain to any extreme event nor a singular catchment condition. About the flocculation factor b, the expressions found have been explained from a technical standpoint. Upcoming data to be collected on the same catchment are expected to confirm these conclusions. In contrast, in the case of settling velocity a, the lack of consistent and understandable relationships between variables basically prevents us from drawing similar conclusions, and the harvesting of additional data (even involving different type of information) is recommended for future studies.

The results of the analysis are mostly consistent with previous studies on urban CSOs where a strong relationship between the TSS removal efficiency and the TSS initial concentration was observed (Rossini *et al.* 1999; Li *et al.* 2003). Nonetheless, in common wet-weather sewage and CSO, higher initial TSS concentration results in an increase in particle size and particle density and, in turn, in an increase of removal efficiency (Lenhart 2008). On the contrary, the analyses reported here, as well as previous studies on the LC catchment, emphasize that an inverse relationship holds. This is explained by considering that the matter flushed from the catchment surface during rainfall events is mostly inorganic and consequently the particle coagulation decreases.

Regarding the settling velocity a, it is known to depend on particle size and gravity distribution which were not available among the analyzed data. A follow-up study is currently investigating the particles size and gravimetric changes to better define the variability of the gravitational forces indicator among different events.

Despite the complexities and challenges normally associated with the measure and evaluation of the physical and chemical sedimentation characteristics in CSOs this study highlights the concrete possibility of a relationship between some of these characteristics and easier measureable and appraisable parameters. The usefulness on such models to improve design and managements practices of treatment plants has been also delineated.

REFERENCES

Berardi, L. & Kapelan, Z. 2007 Multi-Case EPR strategy for the development of sewer failure performance indicators. In: *Proceedings of the World Environmental and Water Resources Congress – Restoring Our Natural Habitat, Tampa, Florida, USA* (K. C. Kabbes, ed.). ASCE Publisher, pp. 1–12. (CD-ROM).

Berardi, L., Kapelan, Z., Giustolisi, O. & Savic, D. 2008 Development of pipe deterioration models for water distribution systems using EPR. *Journal of Hydroinformatics* **10**, 113–126.

Berthouex, P. M. & Stevens, D. K. 1982 Computer analysis of settling test data. *Journal of Environmental Engineering* **108**, 1065–1069.

Butler, D. & Davies, J. W. 2000 *Urban Drainage*. E&FN SPON, London.

Camp, T. R. 1946 Sedimentation and the design of settling tanks. *Transactions ASCE* **111**, 895–958.

Characklis, G. W., Dilts, M. J., Simmons III, O. D., Likirdopulos, C. A., Krometis, L.-A. H. & Sobsey, M. D. 2005 Microbial partitioning to settleable particles in stormwater. *Water Research* **39**, 1773–1782.

Chebbo, G., Lucas-Aiguier, E., Bertrand-Krajewski, J. L., Gagnè, B. & Hedges, P. 1998 Analysis of the methods for determining the settling characteristics of sewage and stormwater solids. *Water Science and Technology* **37** (1), 53–60.

Domingos, P. 1999 The role of Occam's Razor in knowledge discovery. *Data Mining and Knowledge Discovery* **3**, 409–425.

Eckenfelder, W. W. 1989 *Industrial Water Pollution Control*. McGraw-Hill, New York.

Fayyad, U. M., Piatetsky-Shapiro, G. & Smyth, P. 1996 From data mining to knowledge discovery: an overview. In: *Advances in Knowledge Discovery and Data Mining* (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy, eds). AAAI Press and the MIT Press, Cambridge, MA, Ch. 1, pp. 1–34.

Giustolisi, O. & Savic, D. A. 2006 A symbolic data-driven technique based on evolutionary polynomial regression. *Journal of Hydroinformatics* **8**, 207–222.

Giustolisi, O. & Savic, D. A. 2009 Advances in data-driven analyses and modelling using EPR-MOGA. Special Issue on *Advances in Hydroinformatics, Journal of Hydroinformatics* **11**, 225–236.

Giustolisi, O., Doglioni, A., Savic, D. & Webb, B. W. 2007 A multi-model approach to analysis of environmental phenomena. *Environmental Modelling & Software* **22**, 674–682.

Giustolisi, O., Savic, D. & Laucelli, D. 2009 Asset deterioration analysis using multi-utility data and multi-objective data mining. *Journal of Hydroinformatics* **11**, 211–224.

Koza, J. R. 1992 *Genetic Programming: On the Programming of Computers by Natural Selection*. MIT Press, Cambridge, MA.

Laucelli, D. & Giustolisi, O. 2011 Scour depth modelling by a multi-objective evolutionary paradigm. *Environmental Modeling & Software* **26**, 498–509.

Lenhart, J. H. 2008 BMP Performance expectation functions – a simple method for evaluating stormwater treatment BMP performance data. *Florida Water Resources Journal* **7**, 28–39.

Li, J. G., Dhanvantari, S., Averill, D. & Biswas, N. 2003 Windsor combined sewer overflow treatability study with chemical coagulation. *Water Quality Research Journal of Canada* **38**, 317–334.

Lin, H., Ying, G. & Sansalone, J. J. 2009 Granulometry of non-colloidal particulate matter transported by urban rainfall-runoff. *Water, Air and Soil Pollution* **198**, 269–284.

Ljung, L. 1999 *System Identification: Theory for the User*, 2nd edition. Prentice-Hall Inc., Upper Saddle River, NJ.

Markus, M., Hejazi, M., Bajcsy, P., Giustolisi, O. & Savic, D. A. 2010 Prediction of weekly nitrate-N fluctuations in a small agricultural watershed in Illinois. *Journal of Hydroinformatics* **12**, 251–261.

Marsalek, J., Krishnappan, B. G., Exall, K., Rochfort, Q. & Stephens, R. P. 2006 An elutriation apparatus for assessing settleability of combined sewer overflows (CSOs). *Water Science and Technology* **54** (6–7), 223–230.

Maus, C. & Uhl, M. 2010 Tracer studies for the modelling of sedimentation tanks. In: *Proceedings of the 7th International Conference on: Sustainable Techniques and Strategies in Urban Water Management*. Lyon, France, June 27–July 1.

Metcalf & Eddy 1991 *Wastewater Engineering: Treatment, Disposal, Reuse*, 2nd edition. McGraw-Hill, New York.

Metcalf & Eddy 2003 *Wastewater Engineering: Treatment and Reuse*, 4th edition. McGraw-Hill, Boston.

Oke, I. A., Oladepo, K. T., Olajumoke, A. M. & Ajayi, E. O. 2006 Settlement properties of solids in a domestic-institutional wastewater. *Journal of Applied Sciences Research* **2**, 385–390.

Peavy, H. S., Rowe, D. R. & Tchobanoglous, G. 1985 *Environmental Engineering*. McGraw-Hill, New York.

Pettersson, T. J. R. 2002 Characteristics of suspended particles in a small stormwater pond. In: *Proceedings of the Ninth International Conference on: Urban Drainage*. American Society of Civil Engineers, Portland, OR, USA, pp. 1–12.

Piro, P., Carbone, M., Garofalo, G. & Sansalone, J. 2007 CSO treatment strategy based on constituent index relationships in a highly urbanized catchment. *Water Science and Technology* **12** (56), 85–91.

Piro, P., Carbone, M., Garofalo, G. & Sansalone, J. 2010 Size distribution of wet weather and dry weather particulate matter entrained in combined flows from an urbanizing sewershed. *Water Air and Soil Pollution* **206** (1–4), 83–94.

Piro, P., Carbone, M., Penna, N. & Marsalek, J. 2011a Characterization of the sedimentation process for wastewater in combined sewer systems. *Water Research* **45**, 6615–6624.

Piro, P., Carbone, M. & Tomei, G. 2011b Assessing settleability of dry and wet weather flows in an urban area serviced by combined sewers. *Water Air and Soil Pollution* **214**, 107–117.

Rezania, M., Javadi, A. A. & Giustolisi, O. 2010 Evaluation of liquefaction potential based on CPT results using evolutionary polynomial regression. *Computers and Geotechnics* **37**, 82–92.

Rossini, M., Garcia Garrido, J. & Galluzzo, M. 1999 Optimization of the coagulation-flocculation treatment: influence of rapid mix parameters. *Water Research* **33**, 1817–1826.

Sansalone, J. J., Koran, J. M., Smithson, J. A. & Buchberger, S. G. 1998 Physical characteristics of urban roadway solids transported during rain events. *Journal of Environmental Engineering* **124**, 427–440.

Vallet, B., Muschalla, D., Lessard, P. & Vanrolleghem, P. A. 2010 A new dynamic stormwater basin model as a tool for management of urban runoff. In: *Proceedings of the 7th International Conference on: Sustainable Techniques and Strategies in Urban Water Management*. Lyon, France, June 27–July 1.

Vaze, J. & Chiew, F. H. S. 2004 Nutrient loads associated with different sediment sizes in urban stormwater and surface pollutants. *Journal of Environmental Engineering* **130**, 391–396.

Weber, W. J. 1972 *Physicochemical Processes for Water Quality Control*. John Wiley & Sons, New York.

Ying, G. & Sansalone, J. J. 2008 Granulometric relationships for urban source area runoff as a function of hydrologic event classification and sedimentation. *Water Air and Soil Pollution* **193**, 229–246.

Zanoni, A. E. & Blomquist, M. W. 1975 Column settling tests for flocculent suspensions. *Journal of Environmental Engineering* **101**, 309–318.

First received 10 January 2011; accepted in revised form 29 July 2011. Available online 24 October 2011