Multi-objective evolutionary polynomial regression paradigm for the selection of relevant input variables in stormwater quality modelling

Paradigme de la régression polynomiale évolutionnaire multi-objectif pour la sélection de variables d'entrée pertinentes de modèles de qualité des eaux pluviales

E. Creaco[1], L. Berardi[2], Siao Sun[3], O. Giustolisi[4], D. Savic[5]

1 Dept. of Civil Engineering and Architecture, University of Pavia, v. Ferrata 3, 27100 Pavia, Italy. Email: creaco@unipv.it; previously at College of Engineering, Mathematics and Physical Sciences, University of Exeter, North Park Road, Exeter EX4 4QF, UK. Email: E.F.Creaco@exeter.ac.uk
2 Dept. of Civil Engineering and Architecture, Technical University of Bari, v. E. Orabona 4, 70123 Bari, Italy. Email: luigi.berardi@poliba.it
3 Institute of Geographical Sciences and Natural Resource Research, Chinese Academy of Sciences, Beijing, 100101, People's Republic of China. Email: siao.sun05@163.com
4 Dept. of Civil Engineering and Architecture, Technical University of Bari, v. E. Orabona 4, 70123 Bari, Italy. Email: o.giustolisi@poliba.it
5 College of Engineering, Mathematics and Physical Sciences, University of Exeter, North Park Road, Exeter EX4 4QF, UK. Email: D.Savic@exeter.ac.uk

RÉSUMÉ

Cet article présente une procédure pour la sélection des variables d'entrée pertinents en utilisant le paradigme de la régression polynomiale évolutionnaire multi-objectif (EPR-MOGA). La procédure est basée sur l'examen des variables explicatives qui apparaissent à l'intérieur de l'ensemble des expressions symboliques de modèles de l'EPR-MOGA; de complexité et de qualité de calage avec la sortie cible croissantes. La stratégie permet également à la sélection d'être validée par un jugement technique. L'application de la procédure proposée sur la modélisation de la qualité des eaux pluviales sur deux bassins versants français montre qu'elle était en mesure de réduire considérablement le nombre de variables explicatives lors d'analyses successives. Enfin, les modèles EPR-MOGA obtenus après la sélection des variables d'entrée sont comparés avec ceux obtenus en utilisant la même technique sans bénéficier de sélection des variables d'entrée et à ceux obtenus dans des travaux précédents où d'autres techniques de modélisation des données ont été utilisées sur les mêmes données. La comparaison met en évidence l'efficacité à la fois de EPR-MOGA et de la procédure de sélection des variables d'entrée.

ABSTRACT

This paper presents a procedure for the selection of relevant input variables using the multi-objective evolutionary polynomial regression (EPR-MOGA) paradigm. The procedure is based on scrutinizing the explanatory variables that appear inside the set of EPR-MOGA symbolic model expressions of increasing complexity and goodness of fit to target output. The strategy also enables the selection to be validated by engineering judgement. The application of the proposed procedure on modelling storm water quality parameters in two French catchments shows that it was able to significantly reduce the number of explanatory variables for successive analyses. Finally, the EPR-MOGA models obtained after the input selection are compared with those obtained by using the same technique without benefitting from input selection and with those obtained in previous works where other data-modelling techniques were used on the same data. The comparison highlights the effectiveness of both EPR-MOGA and the input selection procedure.

KEYWORDS

Data driven modelling, input selection, EPR, stormwater quality, pollutants

1 INTRODUCTION

In the framework of data driven models, identifying the most relevant variables to describe a given phenomenon is of relevant interest from both the data collection and effective modelling viewpoints. On the one hand, understanding which information should be collected first can help in prioritizing what should be measured under budget constraints. On the other hand, the preliminary selection of inputs is essential for system identification through data-modelling. Irrelevant and redundant input variables used in data-modelling routines might result in poor modelling. As highlighted by Galelli et al. (2014), irrelevant input variables are uninformative about the underlying process and their eventual inclusion would cause noise and complexity to be added to the model. The inclusion of redundant, but relevant, input variables would instead increase the dimensionality of the model identification without providing any additional predictive benefit. One of the possible undesirable consequences of including irrelevant and redundant input variables is the construction of models that over-fit training data, while showing poor generalization capabilities on other similar contexts. Such a drawback is actually undesirable in engineering fields where the transferability of identified data-driven models from one case to another (new) is required.

The issue of input variable selection in non-linear models of stormwater quality is addressed in this paper. In particular, the procedure for relevant input selection developed and described by Creaco et al. (2015) is applied in this work. This procedure is based on the use of the multi-objective evolutionary polynomial regression (EPR-MOGA), a nonlinear regression technique developed by Giustolisi and Savic (2009) which has been widely used in the field of Hydroinformatics

The remainder of the paper is organized as follows. First, the main aspects of the methodology are described. Then, the applications of the methodology to the modelling of stormwater quality in two French sewer systems show how the input selection procedure works with a wide set of available data. Finally, EPR-MOGA is applied to construct stormwater quality models based on the selected input variables in the two sewer systems. The constructed models are then compared with those developed on the same data in a previous work (Sun and Bertrand-Krajewski, 2011). Further details about methodology and applications can be found in Creaco et al. (2016).

2 METHODOLOGY

The input selection procedure is based on the use of the nonlinear regression technique EPR-MOGA, which is able to yield data driven models taking on the following form:

$$\mathbf{Y} = a_0 + \sum_{j=1}^{m} a_j \times (\mathbf{X}_1)^{\mathbf{ES}(j,1)} \cdots (\mathbf{X}_k)^{\mathbf{ES}(j,k)} \times f\left((\mathbf{X}_1)^{\mathbf{ES}(j,k+1)} \cdots (\mathbf{X}_k)^{\mathbf{ES}(j,2k)}\right), \tag{1}$$

where **Y** and **X** are the output and input variables, respectively; a_j is the j-th model parameter, **ES**(j,1,…,k) are the exponents of the inputs, and function f is a user-defined function that has 5 possible settings: (1) no function, (2) natural logarithm, (3) exponential, (4) tangent hyperbolic and (5) secant hyperbolic; m and k represent the maximum number of polynomial terms and the number of input variables (details can be found in Giustolisi and Savic, 2009). In all the applications of the work, the first setting is used for function f, i.e., no function.

The optimal values of parameters a and exponents **ES** are searched for inside EPR-MOGA in such a way as to obtain a trade-off between mutually contrasting objectives (Giustolisi and Savic, 2009). In particular, in the applications of the work, the objective functions calculated for the generic model proposed by EPR-MOGA were: 1 – coefficient of determination (CoD), calculated as a function of simulated and observed values for the output; 2 – number N_x of inputs in Eq. (1); 3 – number N_a of monomials in Eq. (1). In fact, objective function 1 is representative of the goodness of fit of the model to data whereas objective functions 2 and 3 are representative of model complexity.

As a result of the application of EPR-MOGA to a certain training dataset, a Pareto front of optimal models featuring various levels of goodness of fit and complexity is obtained. The goodness of fit of the models can then be assessed also in a testing dataset.

The input selection procedure is based on counting the number of occurrences of the generic input in the optimal set of EPR-MOGA models. After fixing a certain occurrence threshold (equal to 0 in the applications of the work), the generic input is then considered relevant if it features a number of occurrences larger than the threshold.

In the case of multiple available case studies, the total number of occurrences of the generic input is obtained as the sum of its numbers of occurrences in the various case studies.

3 APPLICATIONS

3.1 Case studies

Two urban catchments, located in the West and East part of Lyon (France) respectively, namely the Ecully and Chassieu catchments (Métadier and Bertand-Krajewski, 2012), were considered for the applications. These catchments were two experimental sites in the OTHU project (Field Observatory for Urban Hydrology - www.othu.org). In particular, the Ecully catchment is a low density residential catchment, with a surface area of 245 ha and an imperviousness coefficient of 42%. The Chassieu catchment is an industrial catchment, with a surface area of 185 ha and an imperviousness coefficient of 75%. In the Ecully and Chassieu catchments, numerous variables were measured during 239 and 263 rain events, respectively, in the period from 2004 to 2008.

A typical stormwater quality indicator, i.e., the total amount (kg) of chemical oxygen demand (COD) recorded during a certain rain event, was chosen as target variable (model output). Other 56 variables, characterizing rainfall or runoff information, were considered potential explanatory variables (model inputs). These variables, whose symbols appear in the x axis in the graphs in Figure 1, are thoroughly described by Métadier and Bertand-Krajewski (2012).

In the applications, the first 2/3 of the events were used for selecting the most relevant variables. Afterwards, the same set of events was used to develop and train the data driven models; the second 1/3 of the events were used for testing the data driven models.

3.2 Results

The application of EPR-MOGA to the Ecully and Chassieu catchment yielded the results reported in the graphs in Figure 1, in terms of occurrences of each of the 56 input variables. The application of the input selection procedure enabled identification of 29 relevant inputs and, therefore, a substantial simplification of the problem.

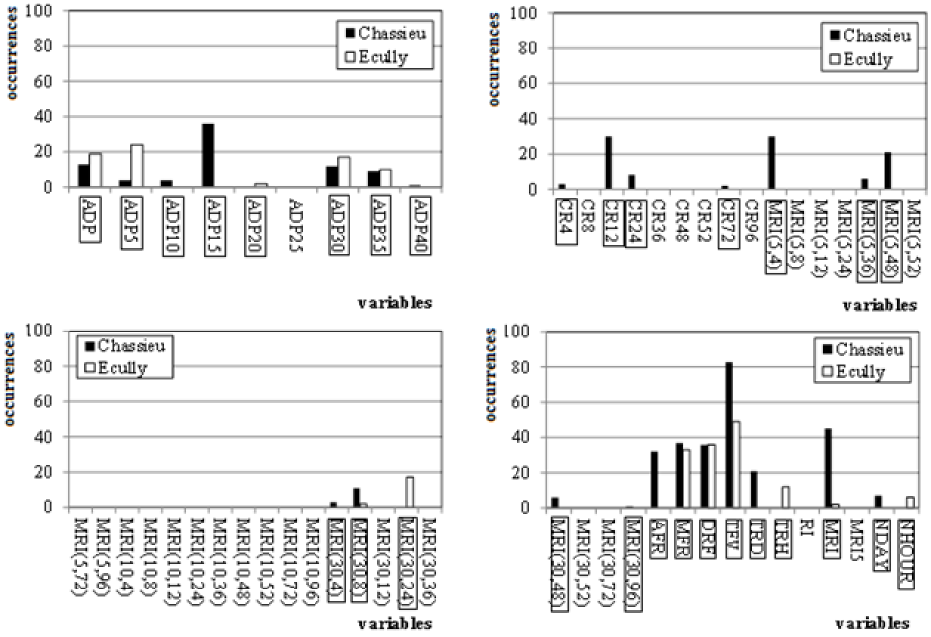

Figure 1 Modelling COD in the two catchments. Occurrences of the explanatory variables in EPR-MOGA models and selected variables highlighted with a rectangle.

A subsequent application of EPR-MOGA to the set of relevant inputs is able to produce models with higher goodness of fit to data in both the training and testing phases, than those that are obtained by applying EPR-MOGA to the whole set of 56 inputs. EPR-MOGA models obtained after relevant input identification are also more robust, since they show smaller performance decay when passing from the training to testing phases. This is evident in the graphs in Figure 2, relative to the simulation of COD in the Ecully catchment. The CoD of the models obtained by applying EPR-MOGA to the relevant inputs were then reported in the graphs in Figure 3 and compared with the CoD values of the models obtained by Sun and Bertrand-Krajewski (2011) using the genetic programming (GP) technique. Overall, the graphs show that various models with larger CoD values than the GP models in both the training and testing phases can be obtained thanks to the fitting performance of EPR and to the adoption of the input selection procedure.

In the paper of Creaco et al. (2015), these results are described more accurately and other aspects, such as the better simplicity and interpretability of the EPR-MOGA models, are highlighted.

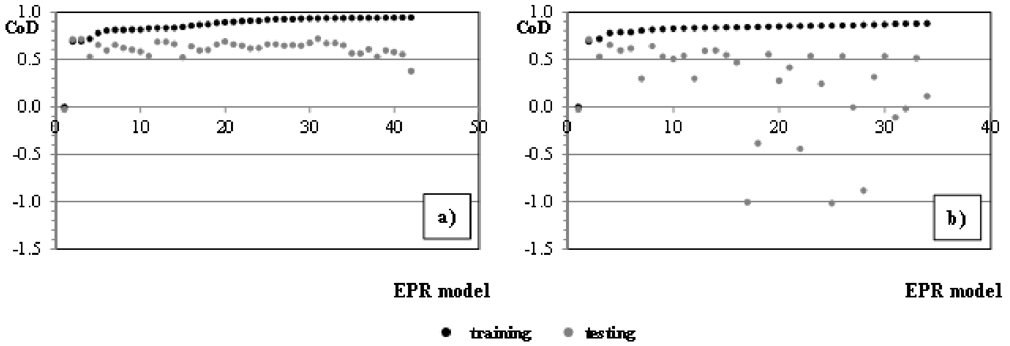

Figure 2 Values of coefficient of determination CoD obtained by the EPR models in the Ecully catchment in the training (black dots) and testing (grey dots) phases for the representation of COD (Kg); results of a) application of EPR following input selection procedure; b) application of EPR alone. In the graphs EPR models are sorted according to ascending values of CoD.

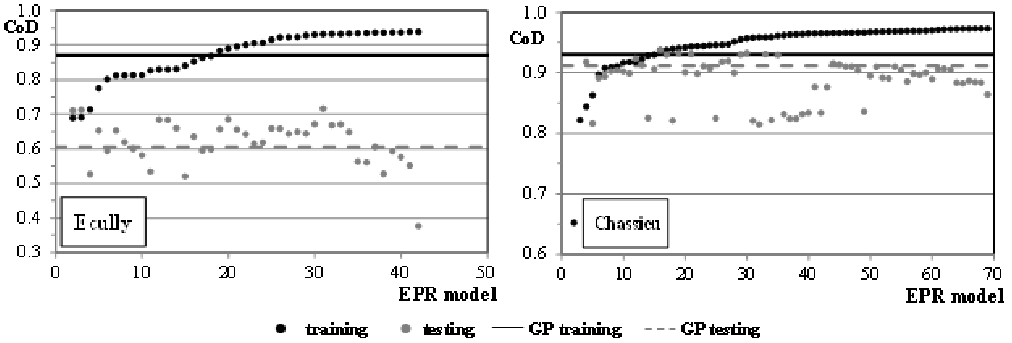

Figure 3 Values of coefficient of determination CoD obtained by the EPR models for the representation of COD (Kg) in the Ecully and Chassieu catchments in the training and testing phases. CoD values of the models obtained by Sun and Bertrand-Krajewski (2011) using the GP technique in the testing and training phases.

ACKNOWLEDGEMENTS

The data sets of the Chassieu and Ecully catchments have been collected and made available by OTHU – Field Observatory on Urban Hydrology (www.othu.org) with the financial support of the Greater Lyon and the Rhône-Méditerranée & Corse Water Agency.

LIST OF REFERENCES

Creaco, E., Berardi, L., Sun, S., Giustolisi, O. and Savic, D. (2016). *Selection of relevant input variables in stormwater quality modelling by multi-objective evolutionary polynomial regression paradigm*. Water Resources Research, doi: 10.1002/2015WR017971.

Galelli, S., Humphrey, G.B., Maier, H.R., Castelletti, A., Dandy, G.C., and Gibbs, M.S. (2014). *An evaluation framework for input variable selection algorithms for environmental data-driven models*. Environmental Modelling & Software, 62, 33-51.

Giustolisi, O., and Savic, D.A. (2009). *Advances in data-driven analyses and modelling using EPR-MOGA*. J. Hydroinformatics, 11, 225-236.

Metadier, M., and Bertrand-Krajewski, J.-L. (2012). *The use of long-term on-line turbidity measurements for the calculation of urban stormwater pollutant concentrations, loads, pollutographs and intra-event fluxes*. Water Research, 46(20), 6836-6856.

Sun, S., and Bertrand-Krajewski, J.-L. (2011). *The calibration of urban storm water quality models using genetic programming (GP)*. In: Urban Water Management: Challenges and Oppurtunities. Exeter, UK, 5-7 September 2011, Eds. D.A. Savic, Z. Kapelan, D. Butler, Centre for Water Systems, University of Exeter, pp. 663-668.