Title

A hybrid multi-step sensitivity-driven evolutionary polynomial regression enables robust model structure selection

Authors

Gomes, Ruan GS
Gomes, Guilherme JC
Vrugt, Jasper A

Publication Date

2022-11-01

DOI

10.1016/j.engappai.2022.105421

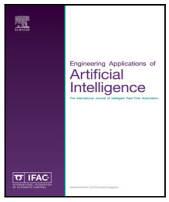
Peer reviewed

Engineering Applications of Artificial Intelligence 116 (2022) 105421

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



A hybrid multi-step sensitivity-driven evolutionary polynomial regression enables robust model structure selection

Ruan G.S. Gomes [a], Guilherme J.C. Gomes [b,c,*], Jasper A. Vrugt [d,e]

[a] Department of Civil and Environmental Engineering, Pontifical Catholic University of Rio de Janeiro, Brazil
[b] Graduate Program in Geotechnics, School of Mines, Federal University of Ouro Preto, Ouro Preto, Minas Gerais, Brazil
[c] Department of Environmental Engineering, Federal University of Ouro Preto, Minas Gerais, Brazil
[d] Department of Civil and Environmental Engineering, University of California, Irvine, CA, USA
[e] Department of Earth System Science, University of California, Irvine, CA, USA

ARTICLE INFO

Keywords:
Evolutionary polynomial regression
Model selection
Monte Carlo simulations
Sensitivity analysis
Optimum moisture content
Creep index

ABSTRACT

Evolutionary Polynomial Regression (EPR) has found widespread application and use for model structure development in engineering and science. This hybrid evolutionary approach merges real world data and explanatory variables to generate well-structured models in the form of polynomial equations. The simple and transparent models produced by this technique enable us to explore, via sensitivity analysis, the robustness of the derived models. Yet, existing EPR frameworks do not make explicit use of sensitivity analysis in the selection of robust and high-fidelity model structures. In this paper, we develop a multi-step sensitivity-driven method which combines the strengths of differential evolution and model selection via Monte Carlo simulation to explore the input–output relationships of model structures. In the first step, our hybrid approach automatically determines the optimum number of terms of the polynomial equations. In a subsequent step, our algorithm explores the mean parametric response of each explanatory variable used in the mathematical formulation to select a final model structure. Finally, in our selection of the most robust mathematical structure, we take explicit consideration of the prediction uncertainty of the simulated output. We illustrate and evaluate our EPR method for different engineering problems involving modeling and prediction of the moisture content and creep index of soils. Altogether, our results demonstrate that the use of sensitivity analysis as an integral part of model structure search and selection will lead to robust models with high predictive ability.

1. Introduction

Complex engineering systems are commonly derived from first principles or closed-form equations (white-box models), data-driven techniques (black-box models), or conceptual mathematical structures (gray-box models) (Giustolisi and Savic, 2006). White-box models assume algebraic and ordinary (or partial) differential equations to model the dynamics of intricate engineering processes at distinct spatial dimensions and temporal scales (Vrugt, 2016). These models can accurately characterize the underlying physical meaning of the process being investigated. Yet, many white-box models are often unable to precisely describe complex, real-world engineering systems. Black-box models, such as artificial neural networks (ANN), have the advantage of dealing with a significant amount of information to produce complex model functions. These approaches can mimic engineering processes by learning from numerous examples, that is, by analyzing input and output data. While popular, these techniques also have their own drawbacks (Giustolisi and Savic, 2006). For instance, parameter estimation and overfitting problems are some of the disadvantages of model development by black-box models (Giustolisi and Laucelli, 2005). What is more, these techniques do not allow us to explicitly incorporate knowledge obtained from physical processes into the model search. This makes it very difficult to detect if the model can reproduce theoretically relevant parts of the system behavior. Consequently, many resort to gray-box techniques such as Evolutionary Polynomial Regression (EPR), which provides well-structured, transparent, and physically based mathematical expressions. Additionally, EPR methods allow us to explore, through sensitivity analyses (parametric study), the generalization ability (robustness), and the physical meaning of each input data in the model (Shahin, 2015). Sensitivity analysis is an essential process for differentiating (gray) EPR models from black-box approaches. For these reasons, EPR models have found its way into engineering practice (Ahangar-Asr et al., 2011b; Alani and Faramarzi, 2014; Balf et al., 2018; Berardi et al., 2008; Bruno et al., 2018; Costa et al., 2020; Doglioni et al., 2010; Doglioni and Simeone, 2021;

* Correspondence to: Department of Environmental Engineering, School of Mines, Federal University of Ouro Preto, 35400-000 Ouro Preto, Minas Gerais, Brazil.
E-mail addresses: ruangomes@puc-rio.br (R.G.S. Gomes), guilhermejcg@ufop.edu.br (G.J.C. Gomes), jasper@uci.edu (J.A. Vrugt).

https://doi.org/10.1016/j.engappai.2022.105421
Received 18 January 2022; Received in revised form 21 July 2022; Accepted 1 September 2022
Available online 1 October 2022

Faramarzi et al., 2012; Fiore et al., 2012, 2016; Giustolisi et al., 2007, 2008; Gomes et al., 2021a; Jin and Yin, 2020; Laucelli and Giustolisi, 2011; Montes et al., 2020; Rezania et al., 2008, 2010, 2011; Shahin, 2015).

EPR is a useful two-stage hybrid regression technique that performs (i) model structure identification, and (ii) parameter estimation to fit simple polynomials in the input–output process. Traditionally, the EPR framework uses simple genetic algorithm (GA) and linear least-squares (LS) for model structure identification and parameter estimation, respectively. The model structure search strategy using single-objective genetic algorithm (SOGA) has been widely applied (Ahangar-Asr et al., 2010, 2011a,b, 2012; El-Baroudy et al., 2010; Faramarzi et al., 2012; Shahin, 2015; Shahnazari et al., 2013). In these approaches, the objective function relies on statistical metrics, such as the minimization of the sum of squared errors (SSE). Still, overfitting and lack of generalization ability are some of the drawbacks involved in SOGA-based EPR modeling schemes (Giustolisi and Savic, 2009; Savic et al., 2009; Laucelli and Giustolisi, 2011; Jin et al., 2019b). In contrast, the multi-objective (MO) strategy enhances the classical SOGA-base EPR techniques for multiple reasons (Marasco et al., 2021). Most notably, MO-based EPR enables us to handle multiple objectives within the search strategy. For instance, several adopted MO-based EPR to maximize the model fitness to data and to minimize the number of polynomial terms (Alani and Faramarzi, 2014; Balf et al., 2018; Berardi et al., 2008; Creaco et al., 2016; Giustolisi and Savic, 2009; Rezania et al., 2008). The MO procedure thus returns a Pareto-efficient subset of feasible non-dominated solutions, that is, optimal model structures based on different criteria (Giustolisi and Savic, 2009). However, a key task is then to select one representative model from the set of Pareto optimal solutions. Hence, many researchers worldwide are trying to improve the different building blocks of EPR to select optimum model structures for empirical engineering models (Jin et al., 2019b; Jin and Yin, 2020; Gomes et al., 2021a; Marasco et al., 2021; Marasco and Cimellaro, 2021). A review of Gomes et al. (2021a) discusses challenges and research gaps on model selection within the EPR framework.

Recently, different methodologies have been proposed to improve search strategy and model selection within the EPR framework. These attempts have successfully developed multi-step automatic model selection schemes. For instance, Jin and Yin (2020) developed an EPR process that consisted of two steps: (i) model selection using a multi-objective differential evolution algorithm (MODE) that handles multiple objectives (model accuracy, complexity and robustness) and (ii) delicacy identification, in which a set of candidate models are ranked according to the coefficient of determination (R^2), number of EPR terms, number of input variables, robustness ratio and monotonicity. Their results highlighted that the MODE-based EPR technique can efficiently model soil properties. Still, they also reported that innovative optimization algorithms or advanced model selection schemes should enhance the EPR performance. In another attempt to create an intelligent multi-step automatic model selection, Jin et al. (2019b) proposed a single-objective differential evolution (SODE) EPR procedure. In their approach, two optimal models were selected to predict the creep index of clays based on the predictive ability, model complexity, robustness and monotonicity. However, to select one of the two models as the optimum, sensitivity analysis was performed on the physical properties used as explanatory variables. Despite both formulations presented excellent predictive ability, the parametric study showed quite different (one of them unrealistic) physical meaning. In fact, from a practical perspective, EPR models must interpret the underlying physical meaning of the system (Shahin, 2015). This raises a question of how uncertain the sensitivity analyzes are if independent simulations are carried out. Therefore, an EPR method that quantifies such information and automatically preserves the theoretical underpinning of the system behavior during the search strategy would hence be desirable to engineering practice. In a similar line of research, Gomes et al. (2021a) have proposed a new EPR method that differs from previous attempts in three different procedures: dual search-based using GA and differential evolution (DE) as model structure exploration engine, self-adaptive evolution of new population and compromise programming as a model selection tool. The study has demonstrated that it is possible to nicely predict dependent variables within the EPR framework with accuracy, physical meaning, and reduced number of parameters and input data in the model structure.

These previously published works did not resolve, however, other questions concerning the optimum model selection and uncertainty quantification within the EPR framework. First, it is unclear if the robustness of single EPR runs is warranted since distinct models might be produced using the same input data and algorithmic parameters. In fact, Oparaji et al. (2017) highlighted that different ANN models might be obtained utilizing equal training data due to the random initialization of the (weights and biases) parameters in each network, which leads to unavoidable uncertainty in the selection of the best performing model. Accordingly, EPR models with different structures can produce close predictive capability, but they could also have quite different generalization and parametric responses (Jin et al., 2019b). On the one hand, simple models without sufficient explanatory variables can potentially overlook components of the system. Alternatively, care should be exercised not to derive complex EPR models with too many parameters in lieu of overfitting, therefore decreasing substantially generalization ability. As a consequence, EPR models that provide excellent predictions and generalization abilities, maintain accuracy and robustness for predicting real-world phenomenon (Jin et al., 2019a; Marasco et al., 2021; Gomes et al., 2021a). Moreover, robust EPR models can preserve features of the physical process that are likely to be revealed in the sensitivity analysis (Shahin, 2015). Second, while sensitivity analysis could provide several advantages to the model search strategy, existing publications only adopt such parametric study subsequently to the definition of the optimum mathematical model (e.g., among others, (Ahangar-Asr et al., 2011a, 2012; Alzabeebee, 2020; Javadi et al., 2012; Rezania et al., 2010; Shahin, 2015; Jin et al., 2019b; Gomes et al., 2021a)). Therefore, a natural question arises whether sensitivity analysis can effectively delineate the space of feasible solutions, i.e., if and how the parametric study can be used as an integral part of the model structure search. This paper addresses these research questions.

This essay introduces and tests a novel multi-step sensitivity-driven EPR. We build on the hypothesis that sensitivity analysis can drive our search strategy toward improved model structure selection. We use our previous multi-objective differential evolution and genetic algorithm EPR (EPR-MODEGA) to explore the search space in pursuit of models that have a trade-off between goodness of fit and model complexity (Gomes et al., 2021a). By coupling two different optimization algorithms (DE and GA) in a self-adaptive evolutionary scheme and a compromise programming tool, the method has shown benefits in the decision-making of optimal EPR models. Here, we extend the usefulness and general applicability of EPR-MODEGA with a new model selection procedure that merges Monte Carlo simulations and a parametric study to investigate how input data are propagated through the EPR models. Monte Carlo simulations provide a simple way to quantify the average trend of model predictions and their corresponding uncertainty ranges (Dao et al., 2020; Naserim et al., 2020; Pham et al., 2019; Tian et al., 2014; Cunha et al., 2014; Oparaji et al., 2017). We are particularly interested in the impact of the uncertainty sources of explanatory data on the simulated output and its associated 95% confidence interval, rather than only looking at its deterministic results. The framework presented herein is illustrated using real-world data, involving the prediction of two complex geotechnical engineering variables, the optimum moisture content and the creep index of clayey soils.

The remaining of this paper is structured as follows. Section 2 briefly describes the classical EPR approach and the new developments of this contribution. In Section 3, we discuss our methodology, including the database, case studies and details of our computational

setup for the EPR simulations. Then, Section 4 highlights the results of the application of our method to real-world data and illustrates the advantages of our proposed EPR framework. Finally, Section 5 concludes this paper with a summary of our main discoveries.

2. EPR method

2.1. Classical approach

The key feature of evolutionary polynomial regression is to assume that the mathematical structure of a given physical phenomenon can be approximately described by:

$$\boldsymbol{y} = \sum_{j=1}^{m} f(\boldsymbol{X}, g(\boldsymbol{X}), a_j) + a_0, \tag{1}$$

where $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$ is a n-vector of simulated data of the physical process, m is the number of terms in the polynomial expression, f represents a polynomial function developed by the process, $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ is the matrix of explanatory data (input data) with k explanatory variables, g denotes an optional function (e.g., $\exp, \log, \cos, \sin$) determined by the user, which extends the polynomial search into a pseudo-polynomial search strategy, a_j is an adjustable parameter for the jth term and a_0 is an optional bias parameter. It is mathematically convenient to transform Eq. (1) into the following vector form (Giustolisi and Savic, 2006):

$$\begin{aligned} \boldsymbol{y}_{n\times 1}(\boldsymbol{\theta}, \boldsymbol{Z}) &= [\boldsymbol{I}_{n\times 1} \ \boldsymbol{Z}^j_{n\times m}][a_0 \ a_1 \ \ldots \ a_j]^{\mathrm{T}} \\ &= \boldsymbol{Z}_{n\times d} \times \boldsymbol{\theta}^{\mathrm{T}}_{d\times 1}, \end{aligned} \tag{2}$$

where $\boldsymbol{y}_{n\times 1}$ represents the least-squares (LS) estimator vector of n target values, $\boldsymbol{\theta}_{d\times 1} = \{a_0, a_1, \ldots, a_j\}$ is a d $(= m + 1)$ vector of regression parameters, $[\ ]^{\mathrm{T}}$ denotes transpose, $\boldsymbol{Z}_{n\times d}$ is a matrix composed by a unitary vector $\boldsymbol{I}_{n\times 1}$ for an optional bias, a_0, and m-vectors of explanatory variables $\boldsymbol{Z}^j$. As an example, the j-term of Eq. (2) can be written as follows:

$$\boldsymbol{Z}^j_{n\times 1} = [\boldsymbol{x}_1^{\mathbf{ES}(j,1)} \cdot \boldsymbol{x}_2^{\mathbf{ES}(j,2)} \cdot \boldsymbol{x}_3^{\mathbf{ES}(j,3)} \cdot \ldots \cdot \boldsymbol{x}_k^{\mathbf{ES}(j,k)}] \tag{3}$$

where, $\boldsymbol{Z}^j$ is the jth column vector whose elements are products of candidate-independent inputs and **ES** comprises a $k \times m$ user-defined matrix of candidate exponents. The central question in the EPR problem is to optimize, by evolutionary computing, the matrix $\mathbf{ES}_{k\times m}$ of exponents for a certain number of terms (m) that will produce the polynomial equation. By adopting linear least squares, it is then possible to tune the vector of regression parameters $\boldsymbol{\theta}$ in Eq. (2).

Fig. 1 illustrates the EPR workflow. The classical approach consists of two main steps: (i) model structure identification using simple genetic algorithm (GA) and (ii) parameter estimation by the linear least-squares (LS) method. Initially, the user must provide a matrix with input data $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$, i.e., explanatory data. Next, a matrix of exponents ($\mathbf{ES}_{k\times m}$) is randomly assigned, from pre-specified power values. After that, the EPR framework proceeds with repeated application of three main steps. First, the exponents created are assigned to the columns of input data, generating then a polynomial equation. Second, the standard least squares method is used to estimate the vector of regression parameters, $\boldsymbol{\theta}_{d\times 1}$ in Eq. (2). Then, the method provides a model structure in the form of a polynomial equation, which can be used to simulate the dependent variable, $\boldsymbol{y}$. In the third step, the vector of simulated data, $\boldsymbol{y}$, is compared to the observed (calibration) data using standard statistical metrics, such as the sum of squared errors (Giustolisi and Savic, 2006). At this point, the first generation of the proposed equations is thus created and tested. If the stopping criterion is not satisfied, GA is used in the evolutionary process to create a new matrix of exponent vectors.

In practice, the evolutionary process can be implemented using any global search algorithm (Jin et al., 2019b). The classical method was widely used in the literature not only because it was the first EPR method available, but also notably because of its simple implementation. Examples are SOGA methods, which generally have the maximization of the model's accuracy as their objective function, but can produce complex model structures, often increasing the chances of overfitting and lack of generalization ability (e.g., Faramarzi et al., 2012; Shahin, 2016). Other methods have been proposed to improve the building blocks of the classical approach shown in Fig. 1. A summary of the main characteristics of the available EPR methods is presented in Table 1. The fitness functions of multi-objective EPR methods are considerably more sophisticated in pursuit of more parsimonious models, with fewer polynomial terms and number of explanatory variables incorporated into the model structure. In addition to handling multiple objectives, some available methods can offer automatic model selection, such as MODE (e.g., Jin and Yin, 2020), while other approaches require inspection of the Pareto front to select the optimum model (e.g., Balf et al., 2018). However, sensitivity analysis must be carried out independently to confirm the robustness of the parametric response of each model. In this work, we draw inspiration from the newly developed MODEGA approach (Gomes et al., 2021a) to develop our gray-box models. The main features of this last method, detailed below, will be

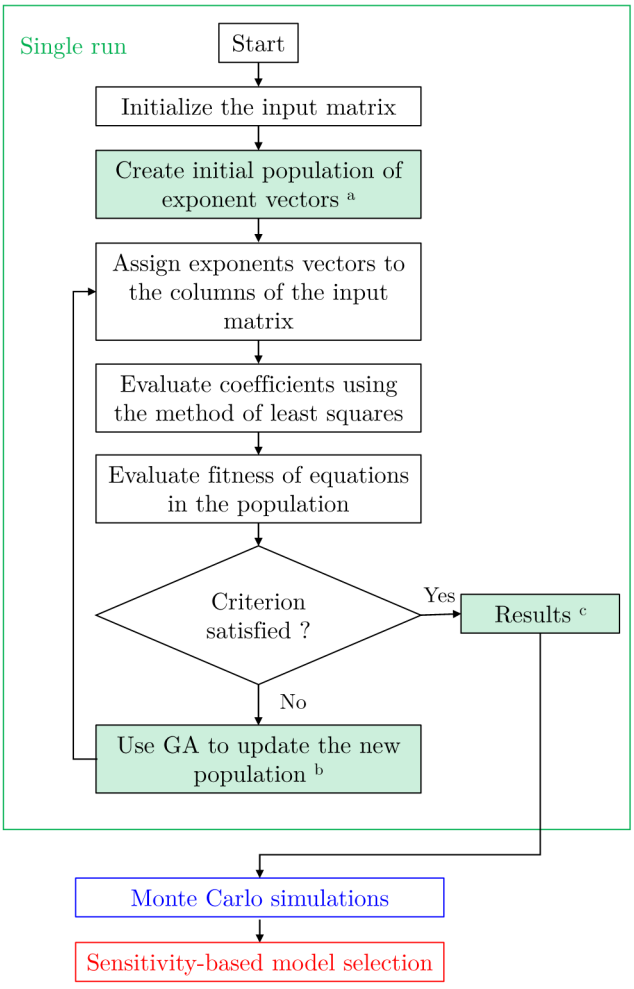


Fig. 1. Overview of the EPR approach. The top (single run) panel represents the flowchart of the classical EPR procedure. The filled green boxes indicate the developments of the EPR-MODEGA, while the final elements of the flowchart illustrate the main steps involved in the sensitivity-driven model structure search strategy.

Table 1
Main characteristics of available EPR methods: general fitness functions, central advantages and significant challenges.

Method	Fitness function	Advantages	Challenges
SOGA	Maximization of model accuracy	Simple implementation	Overfitting and lack of generalization ability
MOGA	Maximization of model accuracy, minimization of polynomial terms and/or minimization of inputs	Handles multiple objectives	Non-automated model selection
SODE	Maximization of model accuracy with complexity penalization	Adaptive process for selecting the combination of involved variables	Requires user inspection of ranked models
MODE	Maximization of model accuracy, minimization of polynomial terms and maximization of model robustness	Handles multiple objectives with automatic model selection	Generalization ability not guaranteed
MODEGA	Maximization of model accuracy, minimization of polynomial terms and minimization of input combinations	Dual search optimization, self-adaptive offspring creation and automatic model selection	Generalization ability not guaranteed

further enhanced to improve model selection using sensitivity analysis, while quantifying modeling uncertainties, which have been overlooked in the literature related to EPR (Jin and Yin, 2020).

2.2. MODEGA approach

EPR-MODEGA differs fundamentally in three elements from the classical approach. First, a multi-objective optimization procedure is implemented using both DE and GA to enhance the search efficiency. In fact, its dual search-based method has shown to outperform individual algorithms such as MOGA and MODE (Gomes et al., 2021a). Second, the self-adaptive offspring creation was specifically designed to select the most efficient search method for population evolution. This technique updates the new population based on the reproductive success of both DE and GA in the previous generation. The proposition adaptively changes the contribution of each algorithm and ensures that the "best" optimization method contributes the most offspring to the next generation (Vrugt et al., 2009). Third, the compromise programming method embedded in the EPR procedure facilitates the decision-making stage, since it enables us to select models preferred statistically from a set of Pareto optimal models with different polynomial terms. The selection of an optimal number of terms is particularly important to avoid overfitting as additional terms introduce unnecessary complexity, hence producing models more sensitive to the noise of the training set that do not generalize to other data sets (Giustolisi and Savic, 2006). During the search procedure and population evolution, two objectives are minimized: minimization of SSE and minimization of the number of explanatory variables in the model structure. These improvements of the EPR-MODEGA method are shown with filled green boxes in Fig. 1.

The EPR-MODEGA approach provides to the user one model structure with m-terms. In fact, in this model search strategy, the fittest models (with different numbers of polynomial terms, m) are stored and then further evaluated using a compromise programming tool. Thus, model structure selection is conducted according to the modeler's viewpoints, who assigns the relative importance to five objectives: minimization of the number of EPR terms (m) and root mean squared error (RMSE), and maximization of the coefficient of determination (R^2), coefficient of correlation (r), and relative efficiency (E_{rel}). This process is illustrated in the top panel of Fig. 2, which summarizes schematically how these statistical metrics vary with m. The solid red squares in each plot denote a hypothetical optimum number of terms ($m = 2$) provided by a single EPR run with the MODEGA method. Overall, the MODEGA approach can help EPR modelers incorporate different objectives, in a relatively simple way, to select the optimum EPR model. Consequently, by adopting such multi-criteria technique for model selection, the approach selects the best-compromised EPR model more efficiently. However, since a single EPR run (see Fig. 1) for a specific value of m can provide different model structures, it has yet to be established whether we can develop a broader, uncertainty-based approach, for model structure selection within the EPR framework.

2.3. New multi-step sensitivity-driven model search

We propose a new EPR framework for model structure selection consisting of two major blocks: (i) Monte Carlo simulations and (ii) sensitivity-driven model selection, which benefits from the Monte Carlo simulations to select one model with the best adjustment to the mean parametric response. Our sensitivity-driven multi-objective differential evolution and genetic algorithm (MODEGA-SD) is coded in MATLAB and integrates the EPR-MODEGA method, Monte Carlo simulations and sensitivity analysis so that the users do not need to port data between the different modeling steps, thereby simplifying substantially sensitivity analyzes and model selection. Furthermore, our code has post-processing features to help visualization of the results, specially the prediction uncertainty ranges of the output variables with respect to each independent variable. The following subsections detail the different steps involved in our methodology.

2.3.1. Monte Carlo simulations

Now that the number of polynomial terms, m, of our EPR framework has been defined, we are left with the final model structure. Thus, we resort to Monte Carlo simulations to quantify the uncertainty of the sensitivity analysis. To execute our Monte Carlo method, users must supply the number of Monte Carlo runs, w. We note here that additional inputs, such as algorithmic EPR parameters for both DE and GA, training and testing datasets, the number of polynomial terms, m, and the set of exponents, which will compose the matrix $\mathbf{ES}_{k\times m}$, are also required information. Table 2 provides a brief description of the different inputs and outputs of our EPR framework. Details of input/output information required for the MODEGA-SD framework will be discussed in Section 3.

Through multiple EPR runs, Monte Carlo simulations provide us to store much more information than single EPR runs. These include a variety of statistical metrics on model performance and sensitivity analysis. Such useful information will be used to investigate the predictive capability and generalization ability of w models with m polynomial terms, and thus providing a basis for model selection using the MODEGA-SD method. As schematically illustrated in Fig. 3, the output information of Table 2 is stored in several matrices, including automatic model parametric responses with respect to each explanatory variable. Thus, with the sensitivity analysis concluded, in the next subsection, we describe how such information can be used as a formal component of the model structure selection.

2.3.2. Model selection

Sensitivity analysis is an important component of the EPR process, since it can provide information on the underlying physics of the derived model (Ahangar-Asr et al., 2011a, 2012; Alzabeebee, 2020; Javadi et al., 2012; Rezania et al., 2010; Shahin, 2015). If we denote $\boldsymbol{y}_i^{(x_u)}$ as a vector that stores the parametric response of the i^{th}-model ($i = \{1, 2, \ldots, w\}$), with respect to the explanatory variable u

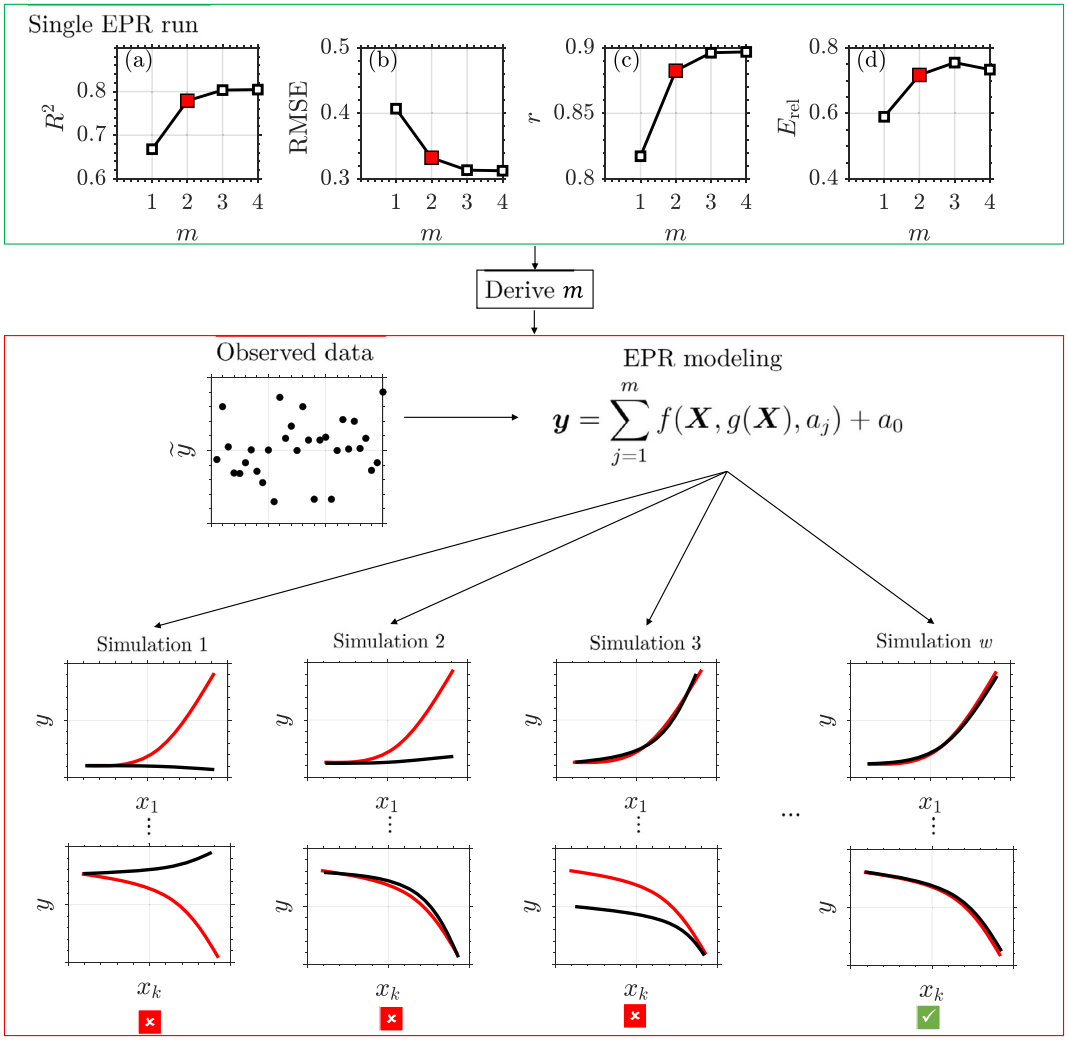


Fig. 2. Summary of our sensitivity-based model selection. The top panel illustrates how the EPR-MODEGA method can be used to derive the optimum number of polynomial terms, m, within a single EPR run. The bottom panel exemplifies the sensitivity analysis of multiple EPR runs with m-terms using Monte Carlo simulations. Red lines denote the mean parametric response of w simulations, while black lines correspond to the parametric response of each model simulation. The green check indicates a model whose parametric response is consistently close to the mean.

Table 2
Description of the input and output variables and algorithmic information of the MODEGA-SD approach, including their size (dimension) or type (scalar or vector).

Inputs	Type/ Size	Outputs	Type/ Size
Training and testing data		Model performance	
Dependent variable, $\boldsymbol{y}$	$n \times 1$	Simulated dependent variable, $\boldsymbol{Y}$	$n \times w$
Explanatory variables, $\boldsymbol{X}$	$n \times k$	Root mean squared error (RMSE)	$n_g \times w$
Algorithmic parameters		Coefficient of determination, R^2	$n_g \times w$
Number of EPR terms, m	scalar	Coefficient of correlation, r	$n_g \times w$
Number of Monte Carlo runs, w	scalar	Relative efficiency, E_{rel}	$n_g \times w$
Set of EPR exponents	vector	Sum of squared errors (SSE)	$n_g \times w$
Number of generations, n_g	scalar	Algorithmic information	
Population size	scalar	Number of offspring points, N_o	$n_g \times w$
Offspring diversity (DE)	scalar	Parametric study	
Crossover rate (DE and GA)	scalar	Model parametric response, $\boldsymbol{Y}^{(X)}$	$n_p \times k \times w$
Mutation rate (GA)	scalar	Mean model response, $\overline{\boldsymbol{Y}}^{(X)}$	$n_p \times k$

$(u = \{1, 2, \ldots, k\})$, then we can write the mean parametric response as follows:

$$\overline{\boldsymbol{y}}^{(x_u)} = \frac{\sum_{i=1}^{w} \boldsymbol{y}_i^{(x_u)}}{w}, \tag{4}$$

where $\overline{\boldsymbol{y}}^{(x_u)}$ identifies the parametric vector that computes the mean model response for the explanatory variable x_u. Since in our sensitivity analysis the model is simulated at n_p points of the explanatory variable (x_u), the size of the array $\overline{\boldsymbol{y}}^{(x_u)}$ is $n_p \times 1$. This process, executed for all k explanatory variables, characterizes the first step of our sensitivity-driven model selection. We can next store the mean model parametric response in a matrix, as follows:

$$\overline{\boldsymbol{Y}}^{(\boldsymbol{X})} = \left[\overline{\boldsymbol{y}}^{(x_1)} \quad \overline{\boldsymbol{y}}^{(x_2)} \quad \ldots \quad \overline{\boldsymbol{y}}^{(x_k)}\right], \tag{5}$$

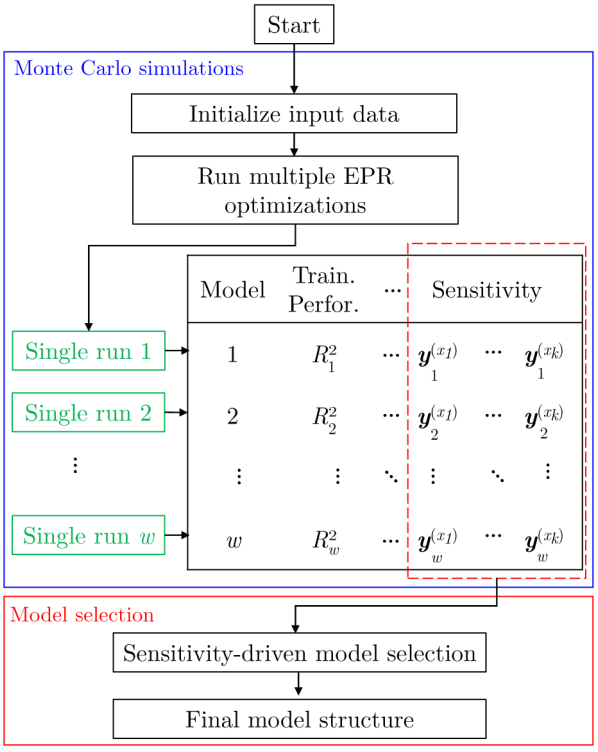


Fig. 3. Schematic overview describing the main steps of the Monte Carlo simulations within the MODEGA-SD approach. By storing a variety of statistical metrics of each model generated and automatically producing parametric responses for these models, our new EPR provides an arsenal of data to be further explored in model selection.

where $\overline{\boldsymbol{Y}}^{(\boldsymbol{X})}$ corresponds to the $n_p \times k$-matrix that stores the mean model parametric responses of all k explanatory variables, $\overline{\boldsymbol{y}}^{(x_1)}$, $\overline{\boldsymbol{y}}^{(x_2)}$ and $\overline{\boldsymbol{y}}^{(x_k)}$ are the mean model parametric values of the first, second and kth explanatory variable, respectively. When an explanatory variable is propagated forward through different models, it is possible that different responses can be produced. In fact, simulation 1 in Fig. 2 depicts that the variables x_1 and x_k produce quite distant responses (solid black lines) from the mean (solid red lines). Simulations 2 and 3 reveal that at least one parametric response of each model varies significantly from the ensemble mean. However, one can see that simulation w, for instance, returns excellent match with the mean model parametric responses. This last model will thus provide a nice generalization ability, since its responses are close to the mean of w model structures derived using an optimum number of EPR terms. Once the matrix $\overline{\boldsymbol{Y}}^{(\boldsymbol{X})}$ has been derived, all the model parametric responses must be statistically investigated, so that they can be ranked accordingly.

The next step is to rank the w models considering their fit to the mean, $\overline{\boldsymbol{y}}^{(x_u)}$. For this purpose, we therefore resort to two standard statistical metrics, the R^2 and RMSE indicators, whose mathematical formulations are expressed as follows:

$$R_i^2 = 1 - \frac{(\overline{\boldsymbol{y}}^{(x_u)} - \boldsymbol{y}_i^{(x_u)})^2}{(\overline{\boldsymbol{y}}^{(x_u)} - \overline{\boldsymbol{y}}_i^{(x_u)})^2}, \tag{6}$$

$$\mathrm{RMSE}_i = \sqrt{\frac{(\overline{\boldsymbol{y}}^{(x_u)} - \boldsymbol{y}_i^{(x_u)})^2}{w}}, \tag{7}$$

where R_i^2 and RMSE_i correspond to the R^2 and RMSE-values of the i^{th}-model, respectively. Of course, other statistical metrics could be considered to evaluate these model parametric responses. These values are then conveniently stored into reference matrices (Eqs. (8) and (9)), as follows:

$$\mathbf{D}_{w\times k}^{(R^2)} = \begin{bmatrix} R_1^{2,(x_1)} & \cdots & R_1^{2,(x_k)} \\ \vdots & \ddots & \vdots \\ R_w^{2,(x_1)} & \cdots & R_w^{2,(x_k)} \end{bmatrix}, \tag{8}$$

$$\mathbf{D}_{w\times k}^{(\mathrm{RMSE})} = \begin{bmatrix} \mathrm{RMSE}_1^{(x_1)} & \cdots & \mathrm{RMSE}_1^{(x_k)} \\ \vdots & \ddots & \vdots \\ \mathrm{RMSE}_w^{(x_1)} & \cdots & \mathrm{RMSE}_w^{(x_k)} \end{bmatrix}, \tag{9}$$

where $\mathbf{D}_{w\times k}^{(\,)}$ is a matrix that stores the statistical performance of w models. For example, $R_1^{2,(x_1)}$ and $R_w^{2,(x_1)}$ correspond to the R^2 values of the first and w^{th} models, respectively, of the first parametric response, that is, the sensitivity analysis of the model response with respect to the variable x_1.

The following procedure consists of sorting the matrices $\mathbf{D}_{w\times k}^{(R^2)}$ and $\mathbf{D}_{w\times k}^{(\mathrm{RMSE})}$ in descending and ascending order, respectively. This ranking method is convenient since it allows us to accommodate the best statistical performances at the first row of the matrix (Eqs. (10) and (11)). We note here that an ascending or descending ranking depends on the statistical indicator being considered. For instance, if we consider the kth explanatory variable, the maximum R^2-value ($R_{\max}^{2,(x_k)}$), the closest to one, would reflect a model parametric response very close to the mean model response. Alternatively, low RMSE-values ($\mathrm{RMSE}_{\min}^{2,(x_k)}$) would then represent the best performances in terms of deviation from the mean model parametric response. Again, it is important to stress that one model can produce an excellent agreement with the mean model parametric response for one explanatory variable, but the adjustment to other variables can deviate considerably from the average (e.g., simulations 2 and 3 in the bottom panel of Fig. 2). These ranked statistical performances are then multiplied by a $w\times k$-matrix, which is filled with values from 1 to w. The resulting award matrices $\mathbf{A}_{w\times k}^{(R^2)}$ and $\mathbf{A}_{w\times k}^{(\mathrm{RMSE})}$ will translate into numbers the sensitivity analysis of the Monte Carlo simulations that are closest to the mean model parametric response. In fact, the smallest values of both matrices will appear in the first row. In our sensitivity-driven model selection, these "best" simulations stored in the matrix $\mathbf{A}_{w\times k}^{(\,)}$ can be schematically observed in the bottom panel of Fig. 2 (simulation w). To facilitate our next step, at this stage, a sort index is stored to link elements of $\mathbf{D}_{w\times k}^{(\,)}$ with those of the sorted award matrix $\mathbf{A}_{w\times k}^{(\,)}$.

$$\mathbf{A}_{w\times k}^{(R^2)} = \begin{bmatrix} R_{\max}^{2,(x_1)} & \cdots & R_{\max}^{2,(x_k)} \\ \vdots & \ddots & \vdots \\ R_{\min}^{2,(x_1)} & \cdots & R_{\min}^{2,(x_k)} \end{bmatrix} \cdot \begin{bmatrix} 1 & \cdots & w \\ \vdots & \ddots & \vdots \\ 1 & \cdots & w \end{bmatrix}, \tag{10}$$

$$\mathbf{A}_{w\times k}^{(\mathrm{RMSE})} = \begin{bmatrix} \mathrm{RMSE}_{\min}^{(x_1)} & \cdots & \mathrm{RMSE}_{\min}^{(x_k)} \\ \vdots & \ddots & \vdots \\ \mathrm{RMSE}_{\max}^{(x_1)} & \cdots & \mathrm{RMSE}_{\max}^{(x_k)} \end{bmatrix} \cdot \begin{bmatrix} 1 & \cdots & w \\ \vdots & \ddots & \vdots \\ 1 & \cdots & w \end{bmatrix}. \tag{11}$$

The optimum model structure in our approach will be the one with better model performance involving all k explanatory variables. If we create an array with values computed by the award matrix $\mathbf{A}_{w\times k}^{(\,)}$ for each model of the Monte Carlo simulation, we can thus compare each model performance with a single scalar. This relatively simple but efficient idea is represented mathematically by Eqs. (12) and (13):

$$\mathbf{s}_{w\times 1}^{(R^2)} = \begin{bmatrix} \mathrm{A}_1^{(R^2,x_1)} + \cdots + \mathrm{A}_1^{(R^2,x_k)} \\ \vdots \\ \mathrm{A}_w^{(R^2,x_1)} + \cdots + \mathrm{A}_w^{(R^2,x_k)} \end{bmatrix}, \tag{12}$$

$$\mathbf{s}_{w\times 1}^{(\mathrm{RMSE})} = \begin{bmatrix} \mathrm{A}_1^{(\mathrm{RMSE},x_1)} + \cdots + \mathrm{A}_1^{(\mathrm{RMSE},x_k)} \\ \vdots \\ \mathrm{A}_w^{(\mathrm{RMSE},x_1)} + \cdots + \mathrm{A}_w^{(\mathrm{RMSE},x_k)} \end{bmatrix}, \tag{13}$$

where $\mathbf{s}_{w\times 1}^{(R^2)}$ and $\mathbf{s}_{w\times 1}^{(\mathrm{RMSE})}$ depict the scoring arrays that compute the sum of elements of $\mathbf{A}_{w\times k}^{(\,)}$ for each model considering the R^2 and RMSE metrics, respectively. $\mathrm{A}_1^{(R^2,x_1)}$ is, for example, the award value

attributed to the first Monte Carlo model for its position in the matrix $\mathbf{A}^{(\ )}_{w\times k}$ when R^2 is used to compare the model response with respect to the variable x_1. One should note that the sum of elements of $\mathbf{A}^{(\ )}_{w\times k}$ for each model (for model 1, $\{\mathrm{A}_1^{(R^2,x_1)}+\cdots+\mathrm{A}_1^{(R^2,x_k)}\}$) will lead to a scalar that will summarize the fit of each model to the mean model parametric response. Eq. (14) generalizes the procedure above described:

$$\mathbf{s}^{(\ )}_{w\times 1} = \begin{bmatrix} \sum \mathrm{A}_1^{(\ ,\boldsymbol{X})} \\ \vdots \\ \sum \mathrm{A}_w^{(\ ,\boldsymbol{X})} \end{bmatrix}, \tag{14}$$

where $\mathbf{s}^{(\ )}_{w\times 1}$ is the scoring vector for a generic $^{(\ )}$ statistical indicator, $\sum \mathrm{A}_1^{(\ ,\boldsymbol{X})}$ sums the awards for the first model considering all explanatory variables, and the same stands for the w^{th} model, $\sum \mathrm{A}_w^{(\ ,\boldsymbol{X})}$. To conclude this step, both statistical indicators are combined by simply adding the scoring vectors, as demonstrated by Eq. (15):

$$\mathbf{s}^{(R^2+\mathrm{RMSE})}_{w\times 1} = \left[\mathbf{s}^{(R^2)}_{w\times 1} + \mathbf{s}^{(\mathrm{RMSE})}_{w\times 1}\right]. \tag{15}$$

The last step relies on the model structure selection, which can be done by sorting the elements of $\mathbf{s}^{(\ )}_{w\times 1}$ in ascending order. The first index that connects the elements of $\sum \mathrm{A}^{(\ ,\boldsymbol{X})}$ into $\mathbf{s}^{(\ )}_{w\times 1}$ along the sorted dimension will be the index of the i^{th}-optimum model. This process can also be performed using the index of the minimum value of $\mathbf{s}^{(\ )}_{w\times 1}$ ($\mathrm{s}^{(R^2+\mathrm{RMSE})}_{\min}$). Finally, the selected (i^{th}-)model is the used to simulate the output variable, $\boldsymbol{y}$, which will be further confronted with observed data. Algorithm 1 provides a step-by-step procedure on how to store all the required information for model structure selection using sensitivity analysis within the MODEGA-SD method.

Algorithm 1 Sensitivity-driven model structure selection.

Require: input data of Table 2
1: **for** $i \leftarrow 1$ to w **do**
2: Compute $\overline{\boldsymbol{y}}^{(x_u)}$ using Eq. (4)
3: Evaluate statistically sensitivity analysis using Eqs. (6) and (7)
4: Store R^2 and RMSE values using Eqs. (8) and (9)
5: Compute the award matrices using Eqs. (10) and (11)
6: Evaluate the models' performances using Eqs. (12) and (13)
7: Merge the models' performances using Eq. (15)
8: $i \leftarrow$ index of $\mathrm{s}^{(R^2+\mathrm{RMSE})}_{\min}$
9: **return** i ▷ i is the index of the optimum EPR model

3. Methodology

The MATLAB framework with its different elements and settings related to the sensitivity-driven model selection was used to develop new EPR models of two complex engineering problems. We below detail the dataset and the input information adopted within our EPR framework. Additionally, we also discuss how the predictive capability of the models was investigated and our treatment of uncertainty in the sensitivity analysis.

3.1. Database

Two case studies are used to test our method: the modeling of optimum moisture content and the modeling of creep index of clays. The modeling of optimum moisture content is based on measurements provided by Ahangar-Asr et al. (2011a), consisting of 57 discrete values. The data predict optimum moisture content (OMC, %), by using values of fineness modulus (F_{m}), coefficient of uniformity (U) and plastic limit (PL, %). The modeling of creep index is established on the data presented by Jin et al. (2019b), comprising a database of 147 measurements. The data predict the creep index (C_α) of soils, by computing the clay content (CI, %), liquid limit (LL, %), plastic index ($\mathrm{I_P}$, %), and void ratio (e). The strength and direction of a linear relationship between each observed input and the target output were quantified using the coefficient of correlation (r). The databases used in this paper are presented as supplementary material (Table S1). All models (developed and cited) were trained and tested using the database mentioned above. This enables a direct comparison between the mathematical structures proposed in this work and those of the two case studies. A complete description of the databases used here is given in the cited publications, and interested readers are referred to these works for further details.

3.2. MODEGA-SD method input information

Table 2 lists the input information required to execute our framework. These include training and testing data of the dependent, $\boldsymbol{y}$ and k-explanatory variables, $\boldsymbol{X}$, deemed important to explain the underlying physical process. As previously detailed, the optimum number of polynomial terms is automatically defined with a single EPR run of the EPR-MODEGA method. Users also have to provide the number of Monte Carlo runs, w, and several algorithmic parameters. In this work, 100 Monte Carlo simulations were considered for each case study. We follow previous work (Berardi et al., 2008; Creaco et al., 2016) and utilize a vector of exponents with a step of 0.1 ($\mathbf{ES} = [-2, -1.9, \ldots, 1.9, 2]$). This step size provides a good compromise between the CPU costs of our EPR method and the corresponding accuracy of the optimal model structure. In fact, if a smaller step size is adopted, a higher number of generations (n_{g}) or population may be necessary to adequately explore the complete model search space. Conversely, the larger the exponent step, the lower the accuracy of the model (Marasco et al., 2021). Here, a default population size of $20m$ was used for 300 generations. The remaining input parameters of both DE and GA optimization algorithms were identical to those reported by Gomes et al. (2021a).

3.3. Predictive capability and robustness

Monte Carlo simulations provided by the MODEGA-SD code allow us to store a series of statistical metrics that can be used to investigate the performance of the models and robustness of our methodology. When the EPR process stops, after the desired number of generations is reached, the simulated outputs, $\boldsymbol{Y}$, of w-EPR runs are stored in a $n \times w$-matrix. At each generation, the values of different statistical metrics such as RMSE, R^2, r, and E_{rel} are stored in $n_{\mathrm{g}} \times w$ matrices for both training and testing data sets (see outputs of Table 2). Despite our efforts to address the performance of the models using common statistical metrics adopted in the EPR framework (e.g., Ahangar-Asr et al., 2011a; Shahin, 2015; Alzabeebee, 2020; Gomes et al., 2021a) some other metrics can be equally useful. To assess the predictive ability and robustness of our procedure, these statistical metrics were evaluated over generations using box plots. The evolution of the mean and corresponding 95% uncertainty ranges of SSE and the number of offspring points, N_{o} were also investigated from one generation to another. N_{o} is a routine of the MODEGA approach that indicates which optimization algorithm, GA or DE, exhibits the greatest reproductive process. To evaluate the performance of the optimum model selected with the sensitivity-driven method, we compared its performance with other single-objective and multi-objective EPR models for both case studies analyzed.

To provide a closer inspection of the differences in bias and model complexity, the Percent bias (PBIAS) (Yapo et al., 1996) and the corrected Akaike's Information Criterion ($\mathrm{AIC_c}$) (Akaike, 1974) applied for small sample sizes (Hurvich and Tsai, 1989), were used to refine model evaluation. PBIAS measures the tendency of the predictions to be larger or smaller than their observed counterparts. PBIAS value of zero is considered optimum, while positive values express a tendency to overestimation, and negative values express a tendency to underestimation. The AIC considers model complexity (parameter dimensionality) and goodness of fit, providing a basis for measuring the quality of each model relative to other models. Models with lower AIC values should be statistically preferred. Mathematical formulas for both performance metrics are found in these cited publications.

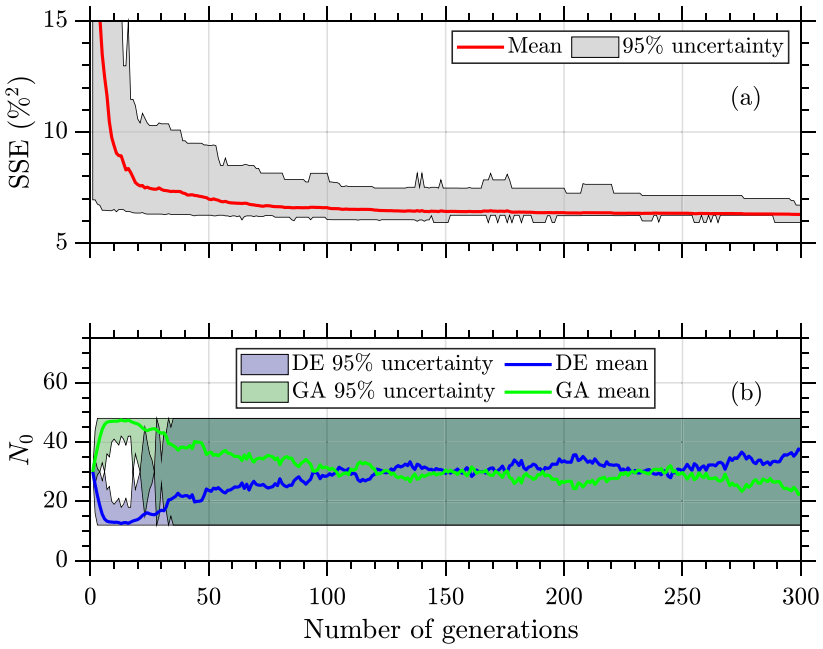


Fig. 4. Case study 1: Evolution of the sum of squared errors (SSE) and its associated 95% uncertainty ranges (a). The bottom plot displays the number of offspring points (N_o) through the generations (b). The mean N_o-value for each individual optimization algorithm is coded with a different color. Light blue and green colors represent the 95% uncertainty intervals of N_o for DE and GA, respectively.

3.4. Sensitivity-driven model selection and its associated uncertainty

During the sensitivity analysis, one explanatory variable varied between the specified minimum and maximum values, while the others were kept constant at their mean values (see Table S1). The mean, minimum, maximum and standard deviation of the values obtained with the matrices $\mathbf{A}_{w\times k}^{(R^2)}$ and $\mathbf{A}_{w\times k}^{(\mathrm{RMSE})}$ were further used to compare the generalization ability of our results with those obtained with different EPR models. Finally, we also showed how large can be the uncertainty in the model simulations obtained using the parametric study provided by the Monte Carlo simulations.

4. Modeling of soil properties

We describe below results obtained with the proposed EPR approach for two different case studies. Modeling of optimum moisture content (OMC) has important implications for the compaction characteristics of soils (Omar et al., 2018; Gomes et al., 2021b), while the modeling of creep index (C_α) is fundamental for a variety of constitutive models used in engineering (Karim and Lo, 2020; Yin et al., 2011).

4.1. Case study 1: Modeling of optimum moisture content (OMC)

Fig. 4(a) illustrates the evolution of the stored SSE (solid red line) and its associated 95% confidence intervals (gray area) for case study 1. Results show that about 100 generations are required for MODEGA-SD to converge adequately to a stable mean SSE-value. The 95% confidence bounds appear quite large during the first generations, but get progressively narrower with the advance of the evolutionary search. The results showed in Fig. 4(a) could be conveniently adopted to guide users toward a sufficient number of generations during the Monte Carlo simulations. Fig. 4(b) depicts the evolution of the number of offspring points of both DE and GA for modeling of OMC. Initially, the GA algorithm demonstrates a highest reproductive success due to the ability of its standard genetic operators for crossover and mutation along the optimization process (Vrugt and Robinson, 2007). After about 100 generations, one can see the adaptive strategy of switching algorithms. Indeed, both DE and GA algorithms have a similar number of offspring points. Yet, at the end of the generations, DE then outperforms GA in terms of reproductive success, a finding that was also previously supported (Vrugt and Robinson, 2007; Gomes et al., 2021a). The low uncertainty ranges of SSE in Fig. 4(a) support the robustness of our methodology, while the rather high confidence intervals in Fig. 4(b) highlight that the combination of global optimization methods appears to be effective in enhancing the richness of solutions along the generations, thereby reflecting the benefits of the dual-search based method.

Now we investigate in Fig. 5 the evolution of (a) R^2, (b) RMSE, (c) r and (d) E_{rel} through generations. The box plots show summary statistics of both training (black) and testing (red) data sets. The right plots are zoomed insets of the data obtained in the final generation and much better exhibit the median values and the corresponding 25$^{\mathrm{th}}$ and 75$^{\mathrm{th}}$ percentiles. Our Monte Carlo simulations demonstrate that multiple different summary statistics improve as evolution is on course of action and quickly converges to the closest fit to the observed OMC data. In fact, from a practical perspective, the R^2-value of both training and testing data are higher than 0.8, which indicates that the performances of the EPR models are adequate for empirical formulations. The low mean RMSE of 2.5 and 2.0 for both training and testing data, respectively, and values of r and E_{rel} greater than 0.9 and 0.8, respectively, further verify the excellent predictive capability of the models derived with the Monte Carlo simulations. Furthermore, the relatively small differences between the statistical metrics of the training and testing data sets reflect the robust nature of our EPR method.

The Monte Carlo simulations thus far have demonstrated excellent agreement between the predicted and observed OMC. We now show in Table 3 summary metrics of the optimum model selected using our sensitivity-based approach and compare with other model structures derived with single runs of the MODEGA algorithm (Gomes et al., 2021a) and with an EPR model proposed by Ahangar-Asr et al. (2011a) obtained using a single objective genetic algorithm (SOGA). MODEGA-SR1 and MODEGA-SR2 are two different (single run) simulations performed with the method presented in Section 2.2. It is clear from Table 3 that the fine performance statistics of the models derived with MODEGA are quite similar. While the summary statistics for the training data of the MODEGA-SD model are slightly lower than those of the MODEGA-SR1, the statistical performance for the testing data of the

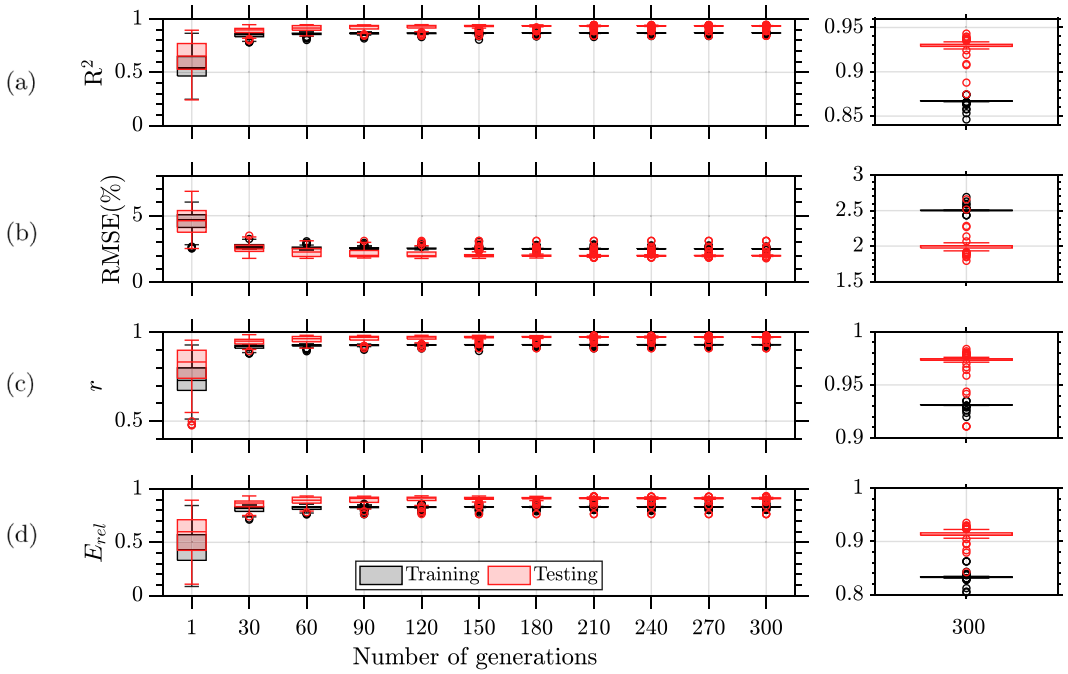


Fig. 5. Evolution of the performance metrics for modeling the optimum moisture content within the Monte Carlo approach. The summary metrics (a) R^2, (b) RMSE, (c) r and (d) E_{rel} include values for both training (black) and testing (red) data sets.

Table 3
Summary statistics of the optimal EPR models selected using the MODEGA-SD, MODEGA-SR1, MODEGA-SR2 and SOGA-SR algorithms for modeling of optimum moisture content. The listed values encompass results for training and testing data.

Metrics	MODEGA-SD		MODEGA-SR1		MODEGA-SR2		SOGA-SR	
	Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.
R^2	0.867	0.929	0.876	0.881	0.865	0.933	0.785	0.842
RMSE (%)	2.501	2.005	2.416	2.588	2.520	1.951	2.920	2.369
r	0.931	0.973	0.936	0.941	0.930	0.976	0.905	0.973
E_{rel}	0.833	0.912	0.867	0.859	0.838	0.918	0.748	0.824
PBIAS (%)	<0.001	–	−0.078	–	0.011	–	−0.388	–
AIC_c	219.2	–	216.1	–	219.9	–	235.7	–

MODEGA-SD model marginally outperforms MODEGA-SR1. Comparison of the summary statistics of the MODEGA-SD and MODEGA-SR2 models gives very similar results. The SOGA-SR EPR model provided a lower statistical performance. Since PBIAS for MODEGA-SD is zero, the model is considered unbiased. While MODEGA-SR2 indicates a slight tendency to overestimation, the remaining models express a small tendency to underestimation. Thus, MODEGA-SD and MODEGA-SR2 outperform the other models. The AIC_c also indicates that both models are less complex. Such results indicate that our approach maintains a good and unbiased predictive capability, an outcome that is consistent with the parsimonious nature of the developed models.

For completeness, Fig. 6 plots the observed and simulated OMC values obtained using the optimal MODEGA-SD model for the training (a) and testing (b) data. The solid black line is used as a reference mark to denote perfect fit. These graphs indicate that MODEGA-SD has a good predictive ability to model the optimum moisture content using the physical properties adopted as inputs. In fact, the regression plots illustrate that simulated OMC values track closely the 1:1 line and are within the 20% error line. The MODEGA models indicate consistency between their results, yet additional evaluations involving the model's parsimony and generalization ability are needed to provide sufficient support on the choice of the optimum model.

To allow a better understanding of the model structures obtained with the EPR approach, we next examine Table 4, which lists the optimum mathematical formulations selected for the optimum moisture content using the MODEGA-SD, MODEGA-SR1, MODEGA-SR2 and SOGA-SR algorithms. The second, third, and fourth columns indicate the number of m terms (obtained by MODEGA and fixed by the SOGA approach), the number of input variables used in the model, and how many input (explanatory) variables were assimilated by the models, respectively. The presence of at least one zero in the matrix $\mathbf{ES}_{k\times m}$ guarantees the ability to exclude some (not relevant) inputs (or input combinations) from the EPR equation in such a manner that the input will not be assimilated by the model (Giustolisi and Savic, 2006, 2009). For modeling the optimum moisture content, the 3 inputs were assimilated by the listed models. The models selected by single runs of the MODEGA approach were different, since the MODEGA-SR1 model included 5 variables in the mathematical formulation, which is somewhat higher than 3 of its MODEGA-SR2 counterpart. MODEGA-SD and MODEGA-SR2 models have very similar mathematical structures and parameter values. The model MODEGA-SR2 was proposed by Gomes et al. (2021a), yet, as in most EPR modeling approaches, it is not guaranteed that such structure will be obtained in the first (single run) optimization. In contrast, the optimum EPR model obtained using MODEGA-SD was obtained with a single Monte Carlo run, providing support for the use of the sensitivity-driven model selection. It is evident from Table 4 that the SOGA-SR model is more complex because of its larger structure. Additionally, the summary metrics shown in Table 3 indicated a reduced predictive ability. This can be explained in part by the single-objective nature of its search strategy and by the lack of a robust model structure selection.

The models proposed in Table 4, however, must be used with caution. EPR is a data-driven technique, strongly dependent on the amount and range of training data. Consequently, the more data, the

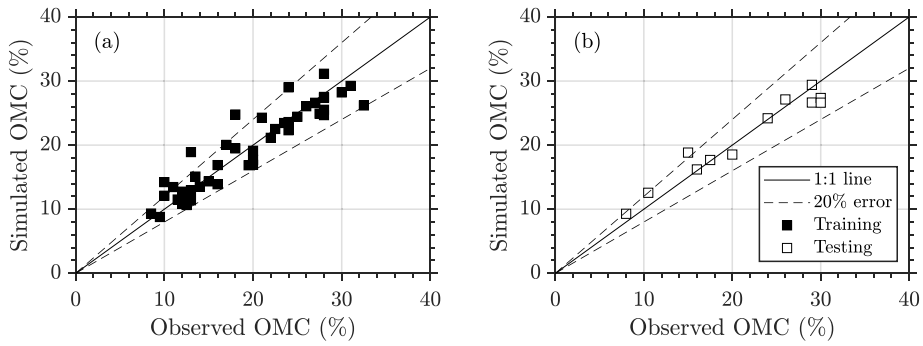


Fig. 6. Observed versus simulated OMC for the best model obtained with the MODEGA-SD approach. The plots display the results for training (a) and testing (b) data.

Table 4
Number of EPR terms (m), explanatory variables involved (inputs), assimilated inputs and the corresponding optimum EPR equations for the optimum moisture content using the MODEGA-SD, MODEGA-SR1, MODEGA-SR2 and SOGA-SR algorithms.

Model	m	Inputs	Assimilated inputs	Optimized equations
MODEGA-SD	3	3	3	$\text{OMC} = -1.86F_m^{1.6} - 8.32U^{0.1} + 0.43\text{PL}^{0.6} + 37.75$
MODEGA-SR1	3	3	3	$\text{OMC} = -4.94F_m^{1.2} + 13.43U^{-0.2} - 3.37F_m^{-3}U^{-0.6}\text{PL}^{-0.2} + 27.65$
MODEGA-SR2	3	3	3	$\text{OMC} = -1.84F_m^{1.6} - 2.52U^{0.2} + 0.42\text{PL}^{0.6} + 31.14$
SOGA-SR	5	3	3	$\text{OMC} = 9.47F_m^{-3}U^{-1} - 3.57\times10^{-5}F_m^{-2}\text{PL}^3 - 4.55\times10^{-3}F_m^{-1}U + 1.72\times10^{-3}\text{PL}^2 - 6.36F_m + 34.09$

better the understanding of the physical process and the more realistic the model structure. In addition, for the purposes of this paper, we have considered the same training and testing data as those of (Ahangar-Asr et al., 2011a). However, cross-validation approaches that use split sampling should be adopted to estimate the model performance when the EPR is trained on different data.

We can further investigate the generalization ability of the models listed in Table 4 by visualizing their sensitivity analysis. Fig. 7 depicts a parametric study of the three explanatory variables used to model OMC. For this parametric study, each variable analyzed ranged between its maximum and minimum values, while the remaining variables were fixed at their mean values (Table S1). The correlation coefficients are displayed in each plot of Fig. 7. The three plots show the ensemble mean (solid red line) and the 95% uncertainty bounds (shaded region) derived with the w-models of our Monte Carlo simulations after 300 generations. The parametric study of the models previously reported in Table 4 is also represented with dashed lines. As shown in Fig. 7(a), the four models predict that OMC is inversely proportional to the fineness modulus (F_m). This finding is in accordance with the observed data ($r = -0.88$), since granular soils have a lower specific surface, which decreases the optimum moisture content (Ahangar-Asr et al., 2011a; Gomes et al., 2021a). Such effect has been correctly captured by all models. The uncertainty ranges envelop a large majority of the OMC simulations, except model MODEGA-SR1, which for F_m values greater than 3, presented a larger deviation from the mean model parametric response. The sensitivity analysis for U-values shows a larger discrepancy of some models with respect to the mean (Fig. 7(b)). These are the cases of the simulations obtained with MODEGA-SR1 and SOGA. However, MODEGA-SD and MODEGA-SR2 closely mimic the mean model parametric response. The negative correlation between OMC and U ($r = -0.35$) is a strong indicator that the models represent the underlying signatures of soil properties (Mujtaba et al., 2013; Ahangar-Asr et al., 2011a). Indeed, the higher the values of U, the larger the range of particle sizes in the soil, and hence the lower the optimum moisture content. Apparently, our findings reveal small uncertainty ranges, which indicate that the models obtained using the Monte Carlo simulations are very similar in terms of the response of OMC to the variations in U. Finally, the mean model parametric response shown in Fig. 7(c) indicates that OMC increases for larger values of plastic limit, PL ($r = 0.63$). Similar results were obtained by Reddy and Grupta (2008) and Sridharan and Nagaraj (2005), who, through laboratory tests, demonstrated that by increasing PL, an increase in OMC is expected, which happens due to a raise in the specific surface of the soil grains. Large deviations of the mean model parametric response are observed by the SOGA model, which predicts OMC outside the uncertainty bounds provided by the Monte Carlo simulations. Furthermore, OMC-values appear insensitive to PL when this explanatory variable is propagated forward in the MODEGA-SR1 model. Therefore, MODEGA-SR1 exemplifies the main drawback of performing a single EPR simulation, that is, the possibility of obtaining an equation in which one explanatory variable does not explain the underlying physical process. This condition would then require one or multiple additional simulations to derive another model with a consistent parametric response for all candidate explanatory variables. Note that MODEGA-SD has been specifically designed to overcome this problem, since one of its functions is to automatically search for a model that has its explanatory variables with physical meaning.

To provide more insights into the newly developed model structure selection within the EPR method, consider Table 5, which lists R^2 and RMSE values obtained using Eqs. (6) and (7) for the Monte Carlo ensemble and for the MODEGA-SD, MODEGA-SR1, MODEGA-SR2 and SOGA-SR models. The values of both statistical metrics, stored in the $\mathbf{D}_{100\times3}^{(R^2)}$ and $\mathbf{D}_{100\times3}^{(\text{RMSE})}$ matrices, are now analyzed. For the Monte Carlo simulations, the maximum, minimum, average, and standard deviation values of each model with respect to the mean model parametric response are listed. Here, maximization of R^2 and minimization of RMSE are proposed as model performance indicators. Bold numbers are given special attention as they reveal the best performance indicators (R^2 and RMSE) that were achieved for specific models. For example, the reported maximum and minimum of R^2 for the explanatory variable F_m are 0.999 and 0.786, respectively. Notice that the average R^2-values obtained with our Monte Carlo simulations closely matched the corresponding maximum R^2-values for all explanatory variables. Similar findings are provided by analyzing the average RMSE values of these input variables, which exhibited similar results to those of the minimum RMSE-listed values. The EPR models show distinct performance metrics. Indeed, MODEGA-SD shows excellent agreement between the model parametric response and the mean model response, while the remaining models appear far less adjusted to the mean model response. This is specially the case of MODEGA-SR1 and SOGA-SR models, as previously highlighted in Fig. 7. In summary, one can see that our sensitivity-driven approach leads to a nice predictive capability

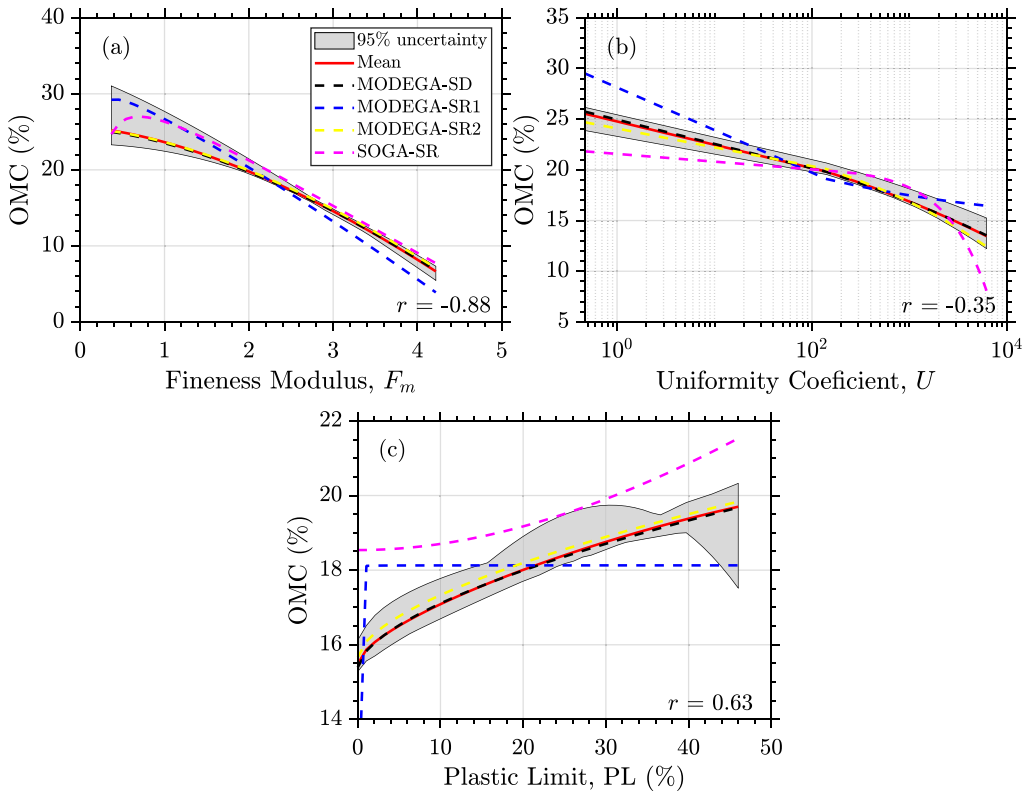


Fig. 7. Sensitivity analysis (parametric study) of the optimum moisture content derived with the best models of MODEGA-SD, MODEGA-SR1, MODEGA-SR2 and SOGA-SR algorithms. The simulated mean OMC is highlighted with solid red lines and the 95% uncertainty bounds (gray region) represent the variability of the output variable with respect to the explanatory variables (a) fineness modulus, F_m, (b) uniformity coefficient, U, and (c) plastic limit, PL. Correlation coefficients (r) are displayed in each plot.

Table 5
Comparison of the sensitivity analysis of OMC models against the mean model parametric response: R^2 and RMSE values of models generated by the Monte Carlo simulations and optimal MODEGA-SD, MODEGA-SR1, MODEGA-SR2 and SOGA-SR models.

Inputs	Metrics	Monte Carlo simulations				EPR models			
		Max.	Min.	Average	Std. Dev.	1[a]	2[b]	3[c]	4[d]
F_m	R^2	**0.999**	0.786	0.989	0.041	0.999	0.838	0.998	0.922
	RMSE (%)	2.617	**0.066**	0.308	0.513	0.088	2.255	0.225	1.563
U	R^2	**0.999**	0.495	0.956	0.115	0.999	0.061	0.911	< 0
	RMSE (%)	1.539	**0.049**	0.241	0.385	0.049	2.075	0.638	2.503
PL	R^2	**0.999**	0.010	0.950	0.137	0.999	< 0	0.974	< 0
	RMSE (%)	1.183	**0.028**	0.154	0.199	0.038	1.233	0.184	1.583

[a]MODEGA-SD.
[b]MODEGA-SR1.
[c]MODEGA-SR2.
[d]SOGA-SR.

(Table 3) and an excellent generalization ability (Fig. 7 and Table 5). These results provide support for the claim that our method produces robust model structure selection that is consistently stable in terms of sensitivity.

4.2. Case study 2: Modeling of creep index

Fig. 8(a) shows the behavior of SSE (solid red line) stored during the Monte Carlo simulations and its corresponding 95% confidence intervals (gray region) for case study 2. The average SSE-value decreases with simulations until it stabilizes after 100 generations. Unexpectedly, this result is in agreement with the data presented in case study 1 (see Fig. 4(a)). The width of the 95% confidence bounds also drops substantially during evolutionary search. Fig. 8(b) presents the performance of DE and GA within our multi-step optimization approach. Similarly to the first case study, it is evident that GA is most efficient at early generations as the number of offspring points is much larger. The reproductive success of both optimization algorithms is similar after 75 generations, but gets progressively different towards the end of the optimization process, when then DE acquires greater N_o.

Fig. 9 illustrates the performance of (a) R^2, (b) RMSE, (c) r and (d) E_{rel} through the Monte Carlo simulations. Similar to the first case study, the box-plots on the left-hand side demonstrate that the performance indicators improve as the evolution process occurs for the training and testing datasets. Overall, the differences between the best models obtained with the Monte Carlo simulations decrease across the generations for both data sets. Additionally, the performance indicators of the training data closely match those of the testing data, further supporting the robustness of our methodology. The box-plots of the last generation, displayed on the right-hand side of the plots, also emphasize similarities between the performance of the models using training and testing data.

Summary statistics of the optimum model obtained with the proposed MODEGA-SD method are listed in Table 6. Similar to the first case study, we use three additional models for comparison. Two model structures were derived with single runs of the MODEGA algorithm

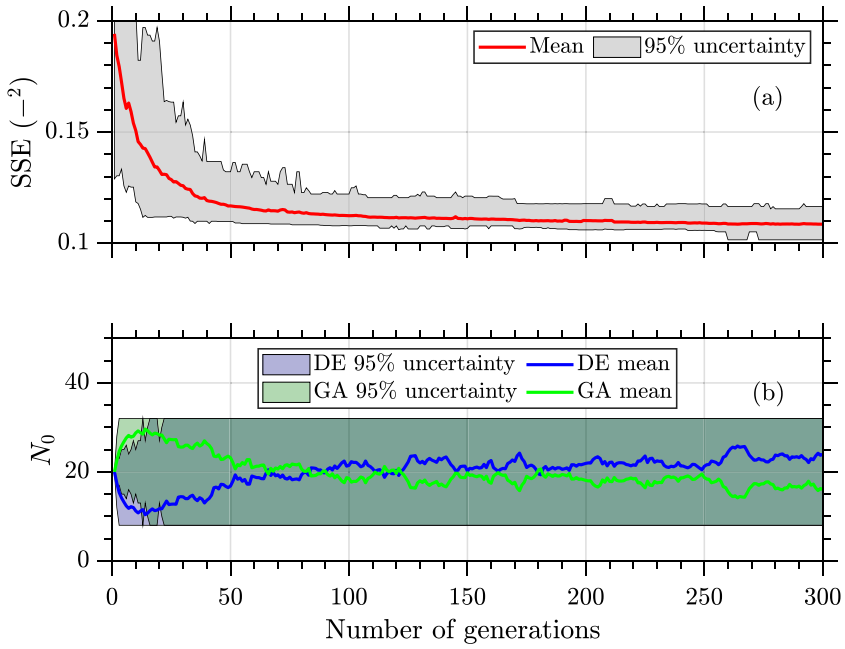


Fig. 8. Case study 2: Evolution of the sum of squared errors (SSE) and its associated 95% uncertainty ranges (a). The bottom plot displays the number of offspring points (N_o) through generations (b). The mean N_o-value for each individual optimization algorithm is coded with a different color. Light blue and green colors represent the 95% uncertainty intervals of N_o for DE and GA, respectively.

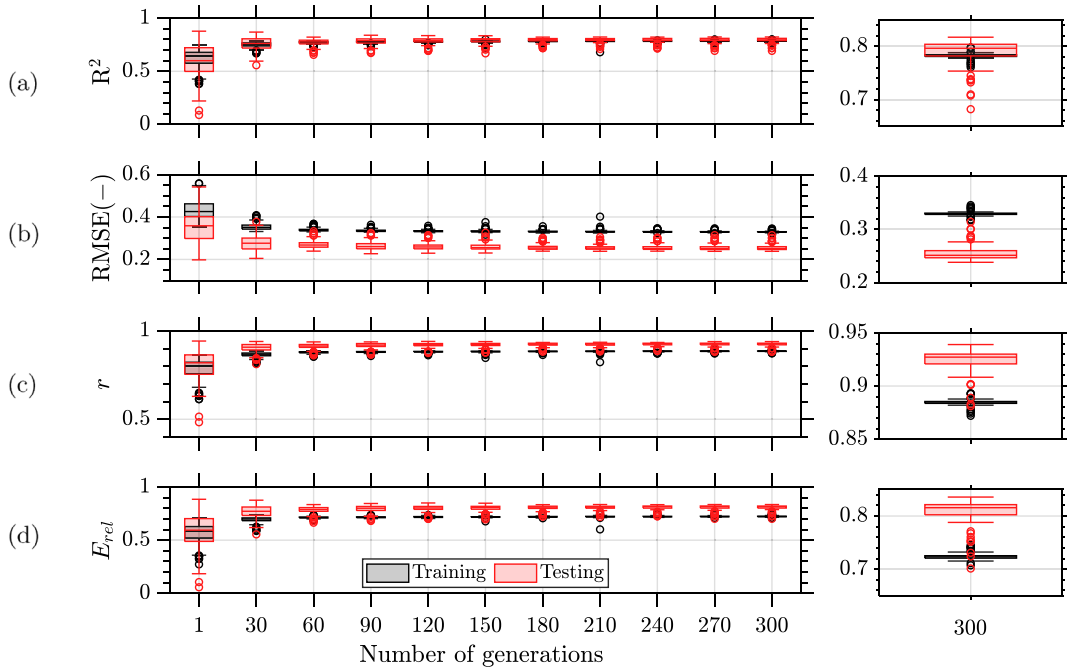


Fig. 9. Evolution of the performance metrics for modeling the creep index of soils within the Monte Carlo approach. Summary metrics (a) R^2, (b) RMSE, (c) r and (d) E_{rel} include values for both training (black) and testing (red) data sets.

(Gomes et al., 2021a) and another EPR model obtained using a SODE algorithm (Jin et al., 2019b), here assumed as a single optimization run (SODE-SR). It is remarkable from the summary statistics that MODEGA-SR1 and SODE-SR showed the best performances for the training and testing data, respectively. However, the statistical metrics of the training data for SODE-SR were not good as those of the MODEGA models. Instead, since MODEGA-SD metrics have the smallest performance distance between training and testing data, the model can be considered more consistent. From the point of view of predictive capability, all EPR models can be useful to obtain the creep index from soil physical properties. The performance of each model using PBIAS are satisfactory since PBIAS is very low, less than 10% (Yapo et al., 1996), so that the models can be considered unbiased. Although AIC_c values indicate that MODEGA models are less complex than SODE-SR for training data, no substantial differences were found between AIC_c values of the tested models. Furthermore, while such performances are very similar as those of the first case study, sensitivity analysis is still required to investigate their parametric response.

We now compare in Fig. 10 the observed and simulated creep index values of the optimal MODEGA-SD model on the training (a) and testing (b) sets. Again, the solid black line is used to denote a perfect fit. Similar to the results reported by Jin et al. (2019b), these

Table 6
Summary statistics of the optimal EPR models selected using the MODEGA-SD, MODEGA-SR1, MODEGA-SR2 and SODE-SR algorithms for modeling the creep index of soils. The listed values encompass results for of training and testing data.

Metrics	MODEGA-SD		MODEGA-SR1		MODEGA-SR2		SODE-SR	
	Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.
R^2	0.785	0.800	0.813	0.830	0.779	0.823	0.698	0.832
RMSE	0.328	0.249	0.305	0.229	0.332	0.235	0.342	0.214
r	0.885	0.928	0.902	0.940	0.883	0.937	0.877	0.932
E_{rel}	0.726	0.817	0.754	0.840	0.717	0.843	0.667	0.832
PBIAS (%)	<0.001	–	0.341	–	<0.001	–	−0.924	–
AIC_c	78.24	–	65.75	–	81.06	–	90.20	–

Table 7
Number of EPR terms (m), explanatory variables involved (inputs), assimilated inputs, and the corresponding optimum EPR equations for the creep index of soils using the MODEGA-SD, MODEGA-SR1, MODEGA-SR2 and SODE-SR algorithms.

Model	m	Inputs	Assimilated inputs	Optimized equations
MODEGA-SD	2	4	4	$\ln(C_\alpha) = 0.25CI^{-0.4}e^{1.9} - 0.23LL^{0.1}\ I_P^{-1.1} - 4.04$
MODEGA-SR1	4	4	4	$\ln(C_\alpha) = 0.06e^{-2.0} - 0.28LL^{1.6}I_P^{-2.0} - 0.65CI^{2.0}LL + 0.65LL\ I_P^{1.0}e - 4.35$
MODEGA-SR2	2	4	3	$\ln(C_\alpha) = 0.29CI^{-0.4}e^{1.8} - 0.04I_P^{-2.0} - 4.43$
SODE-SR	3	4	3	$\ln(C_\alpha) = (0.31CI^{-1.0}\ I_P^{2.0} - 0.12I_P^{-2.0} + 0.65I_P^{-1.0})e - 5.13$

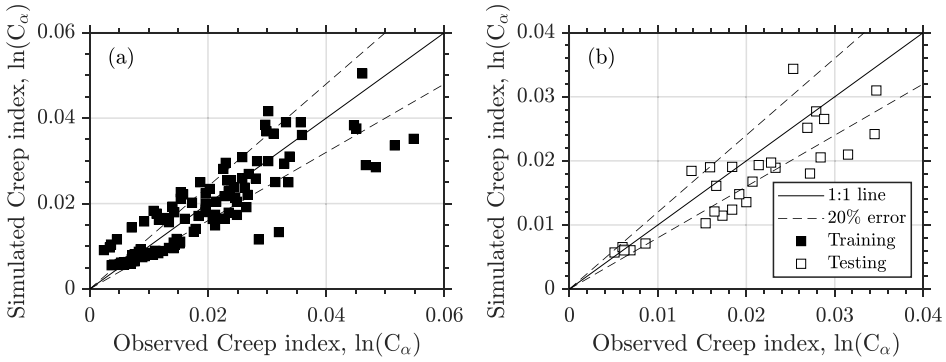


Fig. 10. Observed versus simulated creep index for the best model obtained with the MODEGA-SD approach. The plots display the results for training (**a**) and testing (**b**) data.

graphs indicate that MODEGA-SD predicts the data nicely. Table 7 lists the optimized equations. The second column indicates that MODEGA-SD and MODEGA-SR2 models have a smaller number of m terms. By analyzing the third and fourth columns, it is possible to note that not all models have assimilated all the number of inputs. Indeed, MODEGA-SR2 and SODE-SR models have excluded LL from the mathematical structure during the EPR procedure. We will revisit this question in our sensitivity analysis. While single EPR runs eventually exclude an explanatory variable, our approach benefits from the overall tendency of the multiple runs to decide if the input should be included in the final model structure. Note, for mathematical convenience, creep index is modeled using the natural logarithm. As in case study 1, single runs of the MODEGA approach generated different EPR models. The SODE-SR model fixed the variable e during the search strategy. Once the mathematical formulations have been defined, what is left now is to explore the effect of each input variable on the values of $\ln(C_\alpha)$.

Fig. 11 illustrates the main findings of the sensitivity analysis for the four explanatory variables. The parametric responses are displayed with the uncertainty bounds of the Monte Carlo simulations (shaded region), the mean model parametric response (solid red line), and the output of each model reported in Table 7 (dashed lines). Regardless of the explanatory variable used in the sensitivity analysis, the MODEGA-SD model (dashed black lines) closely follows the shape of the mean model parametric response. The clay content (CI), collected from different experimental papers by (Jin et al., 2019b) has a poor correlation with the creep index ($r = 0.04$), as indicated in Fig. 11(a). The models' responses indicate that a higher amount of CI reflects in a slight decrease in the creep index response for all models. While MODEGA-SD and MODEGA-SR2 models are similar to the ensemble mean, the remaining two EPR models deviate from the uncertainty ranges envelop due to their distinct model structures. As MODEGA-SR1 was not derived from a sensitivity-driven approach, its parametric response differs from the ensemble mean, and provides support for the claim that single EPR optimization runs might produce low-fidelity models. In contrast to the relatively high positive correlation ($r = 0.67$) between LL and $\ln(C_\alpha)$, the sensitivity analysis shown in Fig. 11(b) indicates small (almost negligible) correlation between these variables. As MODEGA-SR2 and SODE-SR did not assimilate LL into their model structures, the creep index value is insensitive to LL variation. The MODEGA-SR1 model has shown excellent predictive capability (Table 6), yet the presence of a more pronounced (negative) correlation between $\ln(C_\alpha)$ and LL leaves in doubt the generalization ability of this model. The correlation between LL and I_P was found to be noteworthy ($r = 0.91$), indicating that LL could be excluded from the modeling approach. Since the objective of this paper is on model structural selection rather than on the influence of explanatory variables on models' performance (e.g., Creaco et al., 2016), LL was kept in the modeling for purposes of comparison with previous work (Jin et al., 2019b). The model's response to I_P variation, available in Fig. 11(c) indicates that increases in I_P reflect high increases in the creep index response, results that are in line with the positive correlation between such variables ($r = 0.75$). The uncertainty ranges of $\ln(C_\alpha)$ are higher for greater values of I_P. The responses captured by MODEGA-SD and MODEGA-SR2 are well within the uncertainty bounds. The mean parametric response indicates that $\ln(C_\alpha)$ first strongly increases, but then when I_P reaches about 50%, there is a decrease in the increment of $\ln(C_\alpha)$. Model MODEGA-SR1, contrarily, predicts a decrease in $\ln(C_\alpha)$ for larger I_P-values. A more complex relationship between these variables is found using the

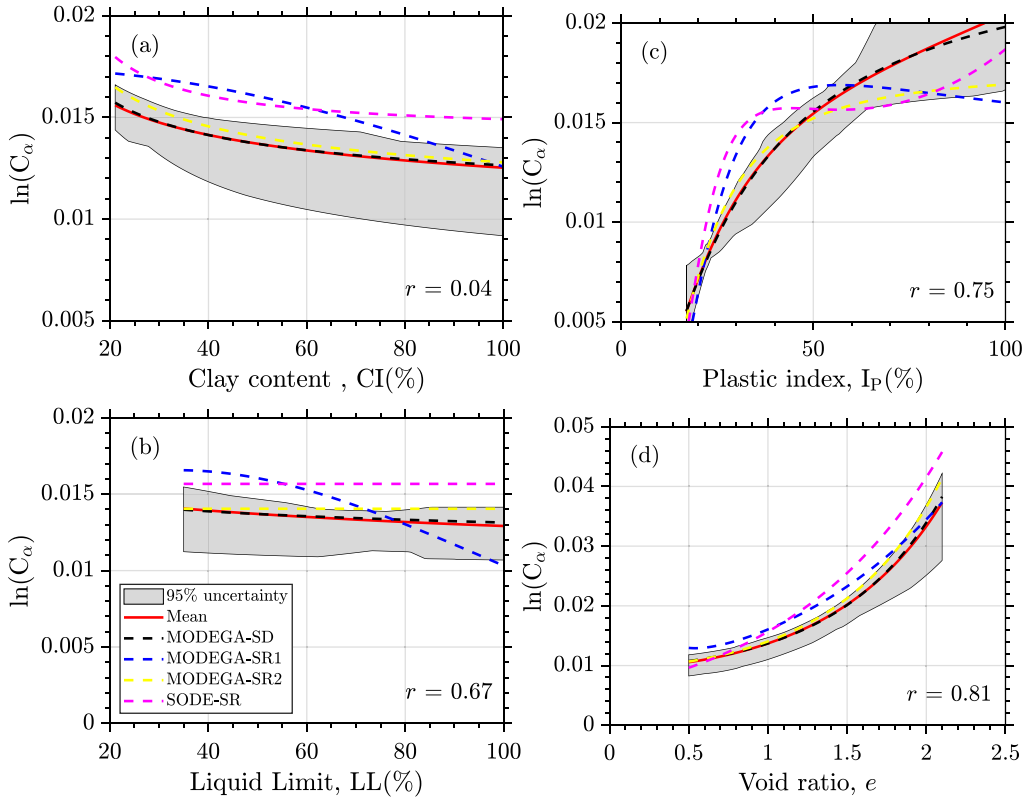


Fig. 11. Sensitivity analysis (parametric study) of the creep index derived with the optimum models of MODEGA-SD, MODEGA-SR1, MODEGA-SR2 and SODE-SR algorithms. The simulated mean $\ln(C_\alpha)$ is highlighted with solid red lines and the 95% uncertainty bounds (gray region) represent the variability of the output variable with respect to the explanatory variables (a) clay content, CI, (b) liquid limit, LL, (c) plastic index, I_P, and (d) void ratio, e. Correlation coefficients (r) are displayed in each plot.

SODE-SR model. Apparently, for I_P-values between 30 and 80%, $\ln(C_\alpha)$ appears insensitive to variations in I_P, but assumes a cubic format when the value of explanatory variable I_P is larger. Finally, the void ratio, e, is known to be positively correlated to the creep index (Zhu et al., 2016). Fig. 11(d) shows that the increase in e reflects an increase in creep index. All EPR models adequately capture this response, nevertheless MODEGA-SD and MODEGA-SR2 were closer to the mean parametric response.

We conclude this manuscript with Table 8, which provides information on the summary statistics (R^2 and RMSE) of the sensitivity analysis. Each EPR model of the Monte Carlo ensemble and the optimum MODEGA-SD, MODEGA-SR1, MODEGA-SR2 and SOGA-SR models are compared with the mean model parametric response. The first column displays the explanatory variables used in this case study. From the second to the fifth column, we list the summary statistics of the $\mathbf{D}^{(R^2)}_{100\times3}$ and $\mathbf{D}^{(\mathrm{RMSE})}_{100\times3}$ matrices. The results presented in Table 8 confirm that the MODEGA-SD model closely tracks the shape of the mean model parametric response. Indeed, R^2 and RMSE values of the four explanatory variables are far better than those obtained with the other models investigated and are similar to the best (marked in bold) statistical metrics of the Monte Carlo simulations. As a consequence, our approach enables us to select optimum EPR models with excellent generalization ability, while also maintaining predictive capability.

The results presented illustrate that our sensitivity-driven evolutionary polynomial regression is a powerful new approach for model structure selection. The robust model structure selection proposed herein significantly reduces the subjectivity of obtaining the optimal structure of a model within the EPR context, thus avoiding intensive and time-consuming efforts. This study appears to be the first to incorporate sensitivity analysis in the evolutionary process for model structure selection. As such, manual post-processing of EPR equations and subjective analysis of the physical meaning of each input data in the model can be avoided. Overall, the results of our case studies showed that the proposed method overcomes difficulties in the decision-making of optimal EPR models. Moreover, as many engineering systems lack a precise analytical theory or model for their solutions, empirical models are much needed in engineering practice. The present contribution has broader implications that go beyond the geotechnical problems studied here. For instance, the MODEGA-SD method should provide new opportunities and perspectives for model selection of other empirico-statistical, multivariate methods that predict engineering properties from covariates.

5. Conclusion

The multi-step sensitivity-driven evolutionary polynomial regression approach introduced in this work offers a coherent and integrated framework for consistent selection of model structures of engineering systems. Based on the dual search-based EPR with self-adaptive offspring creation and compromise programming model selection, our approach couples Monte Carlo simulations and sensitivity analysis to improve model structure selection within this regression-based framework. First, Monte Carlo simulations explores the model search space using an optimal number of polynomial terms. After that, sensitivity analysis of each explanatory variable is used to obtain an arsenal of statistical metrics, which describe relevant parts of the system behavior, including the mean model parametric response. Two real-world case studies involving predictions of optimal moisture content and creep index of soils are used to illustrate our method.

Altogether, results demonstrate that if a model structure produces similar physical meaning to the mean model parametric response obtained with the Monte Carlo framework, the selected model can maintain good predictive and generalization abilities. Statistical (performance) metrics were useful for general monitoring of the predictive ability of the models. Our findings revealed that performance metrics of the training data match closely those of the testing data, sustaining

Table 8
Comparison of the sensitivity analysis of creep index models against the mean model parametric response: R^2 and RMSE values of models generated by the Monte Carlo simulations and optimal MODEGA-SD, MODEGA-SR1, MODEGA-SR2 and SOGA-SR models.

Inputs	Metrics	Monte Carlo simulations				EPR models			
		Max.	Min.	Mean.	Std. Dev.	1[a]	2[b]	3[c]	4[d]
CI	R^2	**0.989**	0.010	0.681	0.365	0.989	< 0	0.760	< 0
	RMSE	0.282	**0.006**	0.042	0.054	0.006	0.125	0.030	0.148
LL	R^2	**0.973**	0.010	0.288	0.358	0.810	< 0	< 0	< 0
	RMSE	0.207	**0.004**	0.043	0.047	0.011	0.128	0.054	0.160
I_p	R^2	**0.999**	0.006	0.925	0.189	0.999	0.775	0.917	0.794
	RMSE	0.642	**0.007**	0.057	0.082	0.007	0.142	0.086	0.136
e	R^2	**0.999**	0.586	0.969	0.073	0.999	0.860	0.979	0.753
	RMSE	0.241	**0.010**	0.047	0.047	0.012	0.139	0.054	0.185

[a]MODEGA-SD.
[b]MODEGA-SR1.
[c]MODEGA-SR2.
[d]SODE-SR.

the robustness of our methodology. Our novel EPR toolbox enabled us to examine the physical meaning of each explanatory variable in the model, including the underlying uncertainty involved in the sensitivity analysis. This provides a relatively simple way to test the generalization ability of the optimum EPR model structure and inspires confidence in our findings.

Despite our efforts to address strategies that explicitly account for sensitivity analysis during model structure selection of the evolutionary polynomial regression framework, there are indeed several implications and opportunities that go beyond the case studies of this work. The framework proposed here, while focused on two engineering problems, could also be applied to a wide variety of complex systems. More substantial research may be needed to improve the statistical metrics used to quantify model adequacy. Moreover, without sufficiently large calibration data, the application of a model to out-of-sample prediction, that is, for data outside the domain spanned by the observations, is particularly challenging. Therefore, additional data collection would help to develop more realistic models with the proposed sensitivity-driven approach. To further support the robustness of our methodology, cross-validation approaches should be used to evaluate the model performance when the EPR is trained on different data. Finally, the sensitivity of algorithmic parameters to the performance of EPR methods has been overlooked, and thus more focused studies on how algorithmic parameters affect model selection may be worthwhile.

CRediT authorship contribution statement

Ruan G.S. Gomes: Investigation, Methodology, Database, Formal analysis, Software, Writing – original draft. **Guilherme J.C. Gomes:** Conceptualization, Formal analysis, Software, Validation, Writing – review & editing. **Jasper A. Vrugt:** Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors greatly acknowledge financial support from the Brazilian National Council for Scientific and Technological Development, CNPq. The MATLAB code of the EPR toolbox developed herein is available upon request from the first author. This includes access to different data, codes and scripts of this paper. The quality of this paper has been greatly enhanced by the constructive comments of three anonymous reviewers and associated editor handling this manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.engappai.2022.105421.

References

Ahangar-Asr, A., Faramarzi, A., Javadi, A., 2010. A new approach for prediction of the stability of soil and rock slopes. Eng. Comput. 27 (7), 878–893. http://dx.doi.org/10.1108/02644401011073700.

Ahangar-Asr, A., Faramarzi, A., Javadi, A., Giustolisi, O., 2011b. Modelling mechanical behaviour of rubber concrete using evolutionary polynomial regression. Eng. Comput. 28 (4), 492–507. http://dx.doi.org/10.1108/02644401111131902.

Ahangar-Asr, A., Faramarzi, A., Mottaghifard, N., Javadi, A., 2011a. Modeling of permeability and compaction characteristics of soils using evolutionary polynomial regression. Comput. Geosci. 37 (11), 1860–1869. http://dx.doi.org/10.1016/j.cageo.2011.04.015.

Ahangar-Asr, A., Johari, A., Javadi, A., 2012. An evolutionary approach to modelling the soil–water characteristic curve in unsaturated soils. Comput. Geosci. 43, 25–33. http://dx.doi.org/10.1016/j.cageo.2012.02.021.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19 (6), 716–723. http://dx.doi.org/10.1109/TAC.1974.1100705.

Alani, A.M., Faramarzi, A., 2014. An evolutionary approach to modelling concrete degradation due to sulphuric acid attack. Appl. Soft Comput. 24, 985–993. http://dx.doi.org/10.1016/j.asoc.2014.08.044.

Alzabeebee, S., 2020. Application of EPR-MOGA in computing the liquefaction-induced settlement of a building subjected to seismic shake. Eng. Comput. http://dx.doi.org/10.1007/s00366-020-01159-9.

Balf, M.R., Noori, R., Berndtsson, R., Ghiasi, A.G.B., 2018. Evolutionary polynomial regression approach to predict longitudinal dispersion coefficient in rivers. J. Water Supply 67 (5), 447–457. http://dx.doi.org/10.2166/aqua.2018.021.

Berardi, L., Giustolisi, O., Kapelan, Z., Savic, D.A., 2008. Development of pipe deterioration models for water distribution systems using EPR. J. Hydroinform. 10 (2), 113–126. http://dx.doi.org/10.2166/hydro.2008.012.

Bruno, D.E., Barca, E., Goncalves, R.M., de Araujo Queiroz, H.A., Berardi, L., Passarella, G., 2018. Linear and evolutionary polynomial regression models to forecast coastal dynamics: Comparison and reliability assessment. Geomorphology 300, 128–140. http://dx.doi.org/10.1016/j.geomorph.2017.10.012.

Costa, V., Fernandes, W., Starick, A., 2020. Identifying regional models for flow duration curves with evolutionary polynomial regression: Application for intermittent streams. J. Hydrol. Eng. 25 (1), 04019059. http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0001873.

Creaco, E., Berardi, L., Sun, S., Giustolisi, O., Savic, D., 2016. Selection of relevant input variables in storm water quality modeling by multiobjective evolutionary polynomial regression paradigm. Water Resour. Res. 52 (4), 2403–2419. http://dx.doi.org/10.1002/2015WR017971.

Cunha, A., Nasser, R., Sampaio, R., Lopes, H., Breitman, K., 2014. Uncertainty quantification through the Monte Carlo method in a cloud computing setting. Comput. Phys. Comm. 185 (5), 1355–1363. http://dx.doi.org/10.1016/j.cpc.2014.01.006.

Dao, D.V., Adeli, H., Ly, H.-B., Le, L.M., Le, V.M., Le, T.-T., Pham, B.T., 2020. A sensitivity and robustness analysis of GPR and ANN for high-performance concrete compressive strength prediction using a Monte Carlo simulation. Sustainability 12 (3), http://dx.doi.org/10.3390/su12030830.

Doglioni, A., Mancarella, D., Simeone, V., Giustolisi, O., 2010. Inferring groundwater system dynamics from hydrological time-series data. Hydrol. Sci. J. 55 (4), 593–608. http://dx.doi.org/10.1080/02626661003747556.

Doglioni, A., Simeone, V., 2021. Data-driven modelling of water table oscillations for a porous aquifer occasionally flowing under pressure. Geosciences 11, http://dx.doi.org/10.3390/geosciences11070282.

El-Baroudy, I., Elshorbagy, A., Carey, S.K., Giustolisi, O., Savic, D., 2010. Comparison of three data-driven techniques in modelling the evapotranspiration process. J. Hydroinform. 12 (4), 365–379. http://dx.doi.org/10.2166/hydro.2010.029.

Faramarzi, A., Javadi, A.A., Alani, A.M., 2012. EPR-based material modelling of soils considering volume changes. Comput. Geosci. 48, 73–85. http://dx.doi.org/10.1016/j.cageo.2012.05.015.

Fiore, A., Berardi, L., Marano, G.C., 2012. Predicting torsional strength of RC beams by using evolutionary polynomial regression. Adv. Eng. Softw. 47 (1), 178–187. http://dx.doi.org/10.1016/j.advengsoft.2011.11.001.

Fiore, A., Quaranta, G., Marano, G.C., Monti, G., 2016. Evolutionary polynomial regression-based statistical determination of the shear capacity equation for reinforced concrete beams without stirrups. J. Comput. Civ. Eng. 30 (1), 04014111. http://dx.doi.org/10.1061/(ASCE)CP.1943-5487.0000450.

Giustolisi, O., Doglioni, A., Savic, D.A., di Pierro, F., 2008. An evolutionary multiobjective strategy for the effective management of groundwater resources. Water Resour. Res. 44 (1), http://dx.doi.org/10.1029/2006WR005359.

Giustolisi, O., Doglioni, A., Savic, D., Webb, B., 2007. A multi-model approach to analysis of environmental phenomena. Environ. Model. Softw. 22 (5), 674–682. http://dx.doi.org/10.1016/j.envsoft.2005.12.026.

Giustolisi, O., Laucelli, D., 2005. Improving generalization of artificial neural networks in rainfall–runoff modelling. Hydrol. Sci. J. 50 (3), null–457. http://dx.doi.org/10.1623/hysj.50.3.439.65025.

Giustolisi, O., Savic, D.A., 2006. A symbolic data-driven technique based on evolutionary polynomial regression. J. Hydroinform. 8 (3), 207–222. http://dx.doi.org/10.2166/hydro.2006.020b.

Giustolisi, O., Savic, D.A., 2009. Advances in data-driven analyses and modelling using EPR-MOGA. J. Hydroinform. 11 (3–4), 225–236. http://dx.doi.org/10.2166/hydro.2009.017.

Gomes, G.J.C., Gomes, R.G.S., Vargas Jr., E.A., 2021a. A dual search - based EPR with self - adaptive offspring creation and compromise programming model selection. Eng. Comput. http://dx.doi.org/10.1007/s00366-021-01313-x.

Gomes, G.J.C., Magalhães, A.J., Rocha, F.L.L., Fonseca, A., 2021b. A sustainability-oriented framework for the application of industrial byproducts to the base layers of low-volume roads. J. Cleaner Prod. 295, 126440. http://dx.doi.org/10.1016/j.jclepro.2021.126440.

Hurvich, C.M., Tsai, C., 1989. Regression and time series model selection in small samples. Biometrika 76 (2), 297–307.

Javadi, A., Ahangar-Asr, A., Johari, A., Fararmazi, A., Toll, D., 2012. Modelling stress–strain and volume change behaviour of unsaturated soils using an evolutionary based data mining technique, an incremental approach.. Eng. Appl. Artif. Intell. 25, 926–933. http://dx.doi.org/10.1016/j.engappai.2012.03.006.

Jin, Y.-F., Yin, Z.-Y., 2020. An intelligent multi-objective EPR technique with multi-step model selection for correlations of soil properties. Acta Geotech. 15, 2053–2073. http://dx.doi.org/10.1007/s11440-020-00929-5.

Jin, Y.-F., Yin, Z.-Y., Zhou, W.-H., Shao, J.-F., 2019a. Bayesian model selection for sand with generalization ability evaluation. Int. J. Numer. Anal. Methods Geomech. 43 (14), 2305–2327. http://dx.doi.org/10.1002/nag.2979.

Jin, Y.-F., Yin, Z.-Y., Zhou, W.-H., Yin, J.-H., Shao, J.-F., 2019b. A single-objective EPR based model for creep index of soft clays considering L_2 regularization. Eng. Geol. 248, 242–255. http://dx.doi.org/10.1016/j.enggeo.2018.12.006.

Karim, M.R., Lo, S.-C.R., 2020. Non-linearity of creep coefficient. Geotech. Res. 7, 90–95. http://dx.doi.org/10.1680/jgere.19.00018.

Laucelli, D., Giustolisi, O., 2011. Scour depth modelling by a multi-objective evolutionary paradigm. Environ. Model. Softw. 26 (4), 498–509. http://dx.doi.org/10.1016/j.envsoft.2010.10.013.

Marasco, S., Cimellaro, G.P., 2021. A new evolutionary polynomial regression technique to assess the fundamental periods of irregular buildings. Earthq. Eng. Struct. Dyn. 50 (8), 2195–2211. http://dx.doi.org/10.1002/eqe.3441.

Marasco, S., Fiore, A., Greco, R., Cimellaro, G.P., Marano, G.C., 2021. Evolutionary polynomial regression algorithm enhanced with a robust formulation: Application to shear strength prediction of RC beams without stirrups. J. Comput. Civ. Eng. 35 (6), 04021017. http://dx.doi.org/10.1061/(ASCE)CP.1943-5487.0000985.

Montes, C., Berardi, L., Kapelan, Z., Saldarriaga, J., 2020. Predicting bedload sediment transport of non-cohesive material in sewer pipes using evolutionary polynomial regression – multi-objective genetic algorithm strategy. Urban Water J. 17 (2), 154–162. http://dx.doi.org/10.1080/1573062X.2020.1748210.

Mujtaba, H., Farooq, K., Sivakugan, N., Das, B.M., 2013. Correlation between gradational parameters and compaction characteristics of sandy soils. Int. J. Geotech. Eng. 7 (4), 395–401. http://dx.doi.org/10.1179/1938636213Z.00000000045.

Naserim, A., Jamei, M., Ahmadianfar, I., Behbahani, M., 2020. Nanofluids thermal conductivity prediction applying a novel hybrid data-driven model validated using Monte Carlo-based sensitivity analysis. Eng. Comput. http://dx.doi.org/10.1007/s00366-020-01163-z.

Omar, M., Shanableh, A., Mughieda, O., Arab, M., Zeiada, W., Al-Ruzouq, R., 2018. Advanced mathematical models and their comparison to predict compaction properties of fine-grained soils from various physical properties. Soils Found. 58 (6), 1383–1399. http://dx.doi.org/10.1016/j.sandf.2018.08.004.

Oparaji, U., Sheu, R.-J., Bankhead, M., Austin, J., Patelli, E., 2017. Robust artificial neural network for reliability and sensitivity analyses of complex non-linear systems. Neural Netw. 96, 80–90. http://dx.doi.org/10.1016/j.neunet.2017.09.003.

Pham, B.T., Nguyen, M.D., Dao, D.V., Prakash, I., Ly, H.-B., Le, T.-T., Ho, L.S., Nguyen, K.T., Ngo, T.Q., Hoang, V., Son, L.H., Ngo, H.T.T., Tran, H.T., Do, N.M., Van Le, H., Ho, H.L., Tien Bui, D., 2019. Development of artificial intelligence models for the prediction of compression coefficient of soil: An application of Monte Carlo sensitivity analysis. Sci. Total Environ. 679, 172–184. http://dx.doi.org/10.1016/j.scitotenv.2019.05.061.

Reddy, B., Grupta, A., 2008. Influence of sand grading on the characteristics of mortars and soil–cement block masonry. Constr. Build. Mater. 22, 1614–1623. http://dx.doi.org/10.1016/j.conbuildmat.2007.06.014.

Rezania, M., Faramarzi, A., Javadi, A.A., 2011. An evolutionary based approach for assessment of earthquake-induced soil liquefaction and lateral displacement. Eng. Appl. Artif. Intell. 24 (1), 142–153. http://dx.doi.org/10.1016/j.engappai.2010.09.010.

Rezania, M., Javadi, A., Giustolisi, O., 2008. An evolutionary-based data mining technique for assessment of civil engineering systems. Eng. Comput. 25 (6), 500–517. http://dx.doi.org/10.1108/02644400810891526.

Rezania, M., Javadi, A.A., Giustolisi, O., 2010. Evaluation of liquefaction potential based on CPT results using evolutionary polynomial regression. Comput. Geotech. 37 (1), 82–92. http://dx.doi.org/10.1016/j.compgeo.2009.07.006.

Savic, D., Giustolisi, O., Laucelli, D., 2009. Asset deterioration analysis using multi-utility data and multi-objective data mining. J. Hydroinform. 11 (3–4), 211–224. http://dx.doi.org/10.2166/hydro.2009.019.

Shahin, M.A., 2015. Use of evolutionary computing for modelling some complex problems in geotechnical engineering. Geomech. Geoeng. 10 (2), 109–125. http://dx.doi.org/10.1080/17486025.2014.921333.

Shahin, M.A., 2016. State-of-the-art review of some artificial intelligence applications in pile foundations. Geosci. Front. 7 (1), 33–44. http://dx.doi.org/10.1016/j.gsf.2014.10.002.

Shahnazari, H., Tutunchian, M.A., Rezvani, R., Valizadeh, F., 2013. Evolutionary-based approaches for determining the deviatoric stress of calcareous sands. Comput. Geosci. 50, 84–94. http://dx.doi.org/10.1016/j.cageo.2012.07.006.

Sridharan, A., Nagaraj, H.B., 2005. Plastic limit and compaction characteristics of fine-grained soils. Proc. Inst. Civ. Eng. - Ground Improvement 9 (1), 17–22. http://dx.doi.org/10.1680/grim.2005.9.1.17.

Tian, W., Song, J., Li, Z., de Wilde, P., 2014. Bootstrap techniques for sensitivity analysis and model selection in building thermal performance analysis. Appl. Energy 135, 320–328. http://dx.doi.org/10.1016/j.apenergy.2014.0.

Vrugt, J., 2016. Markov chain Monte Carlo simulation using the DREAM software package: theory, concepts, and MATLAB implementation. Environ. Model. Softw. 75, 273–316. http://dx.doi.org/10.1016/j.envsoft.2015.08.013.

Vrugt, J.A., Robinson, B.A., 2007. Improved evolutionary optimization from genetically adaptive multimethod search. Proc. Natl. Acad. Sci. 104 (3), 708–711. http://dx.doi.org/10.1073/pnas.0610471104.

Vrugt, J.A., Robinson, B.A., Hyman, J.M., 2009. Self-adaptive multimethod search for global optimization in real-parameter spaces. IEEE Trans. Evol. Comput. 13 (2), 243–259. http://dx.doi.org/10.1109/TEVC.2008.924428.

Yapo, P.O., Gupta, H.V., Sorooshian, S., 1996. Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data. J. Hydrol. 181 (1), 23–48. http://dx.doi.org/10.1016/0022-1694(95)02918-4.

Yin, Z.-Y., Karstunen, M., Chang, C.S., Koskinen, M., Lojander, M., 2011. Modeling time-dependent behavior of soft sensitive clay. J. Geotech. Geoenviron. Eng. 137 (11), 1103–1113. http://dx.doi.org/10.1061/(ASCE)GT.1943-5606.0000527.

Zhu, Q.-Y., Yin, Z.-Y., Hicher, P.-Y., Shen, S.-L., 2016. Nonlinearity of one-dimensional creep characteristics of soft clays. Acta Geotech. 11, 887–900. http://dx.doi.org/10.1007/s11440-015-0411-y.