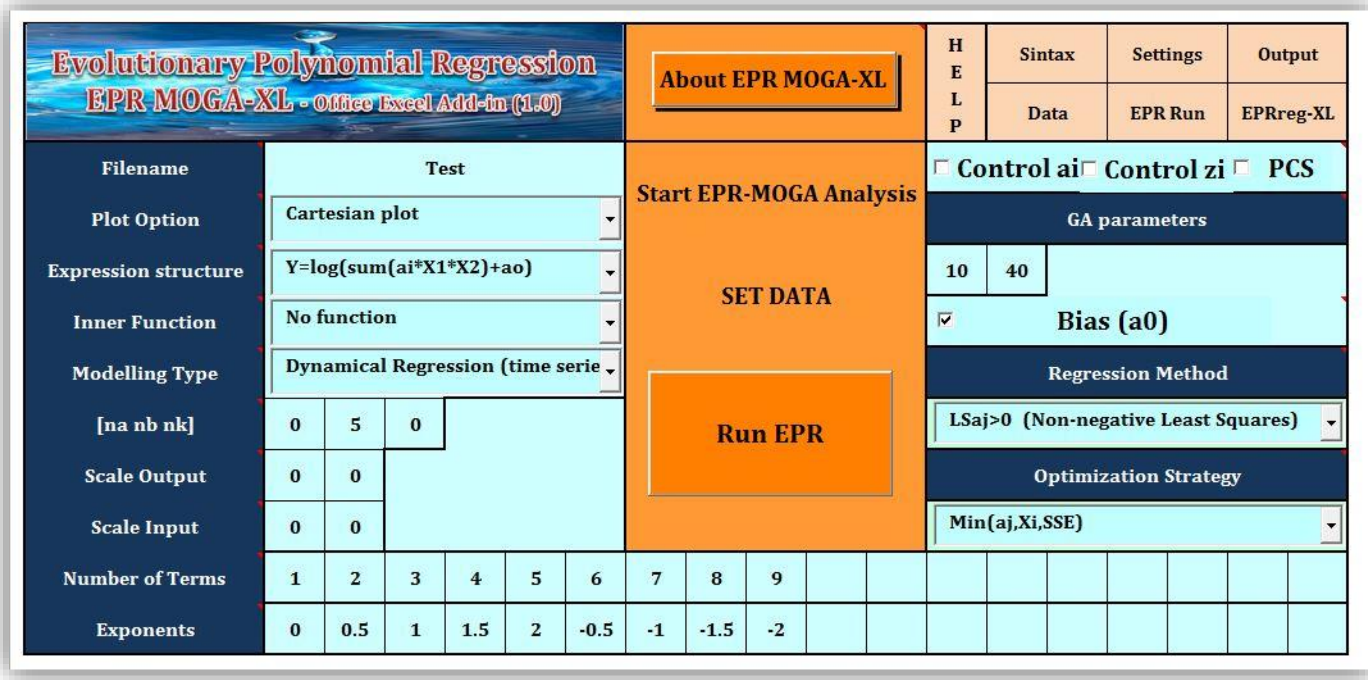
EPR-MOGA-XL

The availability of large and detailed databases together with the increased computational capabilities has motivated researchers to propose innovative techniques and methodologies to mine information from data. The data-driven modelling have been increasingly used for modelling natural and artificial systems.



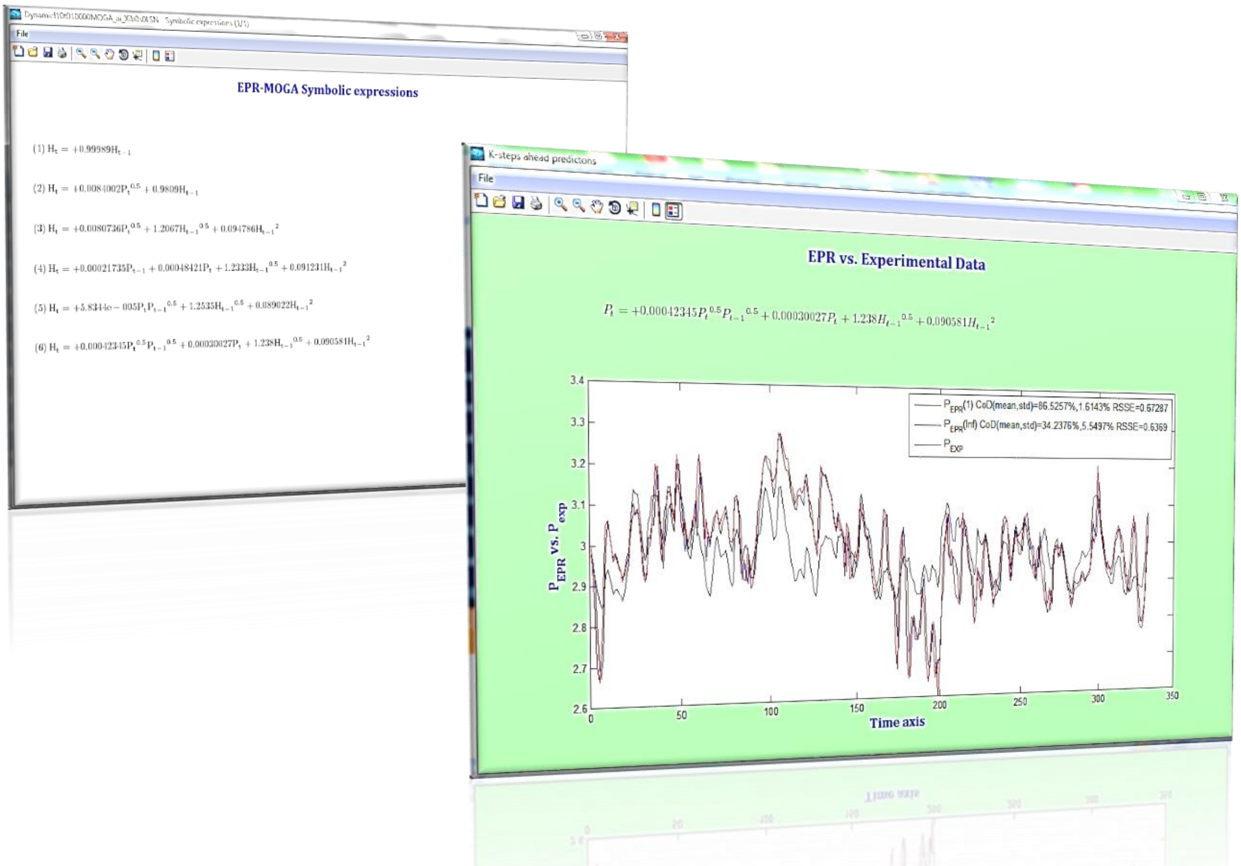
Evolutionary Polynomial Regression is a data-driven hybrid technique, which combines the effectiveness of genetic programming with numerical regression for developing simple and easily interpretable mathematical model expressions.

The EPR approach overcomes some drawbacks of other modelling approaches, such as physically based models and black-box data-driven models. The former can be difficult to be constructed due to the underlying mechanisms that may be not always fully understood, or to the need of many data, sometimes difficult to be measured on field. The latter, as for example artificial neural networks, are very effective in reproducing whatever database related to some observed phenomenon, but bring with them some overwhelming problems, like the model structure identification, the over-fitting to training data, and the inability to exploit physical insight about the phenomenon at stake. The EPR can overcome these problems by means of an explicit model expression for the system under observation. EPR generalizes the original stepwise regression by considering non-linear model structures (i.e., pseudo-polynomials) although they are linear with respect to regression parameters.

One of the general model structures that EPR-MOGA can manage is:

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

where m is the number of pseudo-polynomial terms, a_j are numerical parameters (coefficients) to be estimated, $\mathbf{X}_i$ are candidate explanatory variables, **ES**(j,z) (with z = 1,..., 2k) are the exponents selected from a user-defined set of candidate values (which should include 0), f is a user-selected function (it can be also "no function" resulting into terms obtained by combining input variables).



Such a flexible coding of mathematical expressions permits to explore the space of the models as the combinatorial space of exponents. Model search is cast as the solution of a multi–objective optimization problem where fitting to observations (i.e. model accuracy) is maximized while minimizing the complexity of resulting model expressions. Such search exploits the OPTIMized Multi-Objective Genetic Algorithm (OPTIMOGA) and give rise to a Pareto set of model expressions whose increasing complexity in terms of input variables (i.e. with non-null exponent) and/or number of additional terms, is justified only against an increased fitting performance (i.e. Coefficient of Determination). These features, in turn, help avoiding over-fitting to training data thus improving generalization of resulting models. Furthermore, due to the search for model structure, EPR does not require a prior rigid selection of mathematical expressions and number of parameters.

EPR-MOGA allows to select among optimal models according to the need of the user (e.g. selected model as a trade-off between accuracy and complexity). Additionally, the models can be selected according to the available physical insight about the problem at stake (e.g., recognizing the presence of some known relationship into the explicit formulation of EPR model); conversely, EPR-MOGA can help in discovery some new relationships coming out from the observed data

The EPR-MOGA-XL tool permits to run EPR-MOGA as add-in functions in MS-Excel.