Evolutionary Polynomial Regression application for missing data handling in meteo-climatic gauging stations

E. Barca[1], L. Berardi[2,*], D. B. Laucelli[2], G. Passarella[1], O. Giustolisi[2]

[1] *Water Research Institute of the National Research Council, Department of Bari, Viale F. De Blasio, 5 70123 Bari, Italy; emanuele.barca@ba.irsa.cnr.it, giuseppe.passarella@ba.irsa.cnr.it*
[2] *Dept. of Civil Engineering and Architecture, Technical University of Bari, Via E. Orabona 4, 70125 Bari, Italy; luigi.berardi@poliba.it, daniele.laucelli@ poliba.it, orazio.giustolisi@poliba.it*
[*] *Corresponding author*

Abstract. *One of the most often encountered modelling problems is that of handling missing data, i.e. the problem of intermediate data gaps, where data/observations before and after the missing observations are available. The gaps in data represent discontinuities, which can pose difficulties both for model construction and model application phases. Evolutionary Polynomial Regression (EPR-MOGA) is a data-driven hybrid technique, which combines the effectiveness of genetic programming with the numerical regression for developing simple and easily interpretable mathematical model expressions. Evolutionary Polynomial Regression takes advantage of the evolutionary computing approach that allows the construction of several model expressions based on training data and least squares methodology to estimate numerical parameters/coefficients. These models can then be verified on a test set and gaps can be in-filled in test datasets by using one selected model. Because of the pseudo-polynomial formulations achievable by EPR-MOGA, it requires fewer numbers of parameters to be estimated, which in turn requires shorter time series for training. Another advantage of the EPR-MOGA approach is the ability to choose objective functions pertaining accuracy and parsimony. In the present work, an application of EPR-MOGA is shown on some sites belonging to the Apulian meteo-climatic monitoring network.*

Keywords. *Evolutionary Polynomial Regression; Missing Data Handling; Environmental Monitoring Networks.*

1 Introduction

In the framework of missing data handling, the need arose for methodologies powerful in addressing the issue and more intuitive in their estimation mechanism, particularly in dealing with variables having a well-known space-time structure such as rainfall and temperature. In the present work, a first attempt to deal with the missing data issue via Evolutionary Polynomial Regression (EPR-MOGA) is shown. EPR-MOGA is a data-modelling hybrid technique, which combines the effectiveness of genetic programming with numerical regression for developing simple and easily interpretable mathematical model expressions. Features that make EPR-MOGA paradigm potentially useful for such applications stem from the reduced number of parameters to tune, which in turn requires shorter time series for model training, and the possibility of building non-linear relationships among input-output data, thus going beyond the linear hypothesis underpinning classical geostatistical approaches.

1 Materials and Method

1.1 Evolutionary Polynomial Regression

Evolutionary Polynomial Regression (EPR) 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 ([2]). 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-MOGA can be defined as a non-linear global stepwise regression for symbolic data modelling. EPR generalizes the original stepwise regression of [1, 3] 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 reported in Eq. (1):

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

where m is the number of pseudo-polynomial terms, a_j are numerical parameters (coefficients) to be estimated, $\boldsymbol{X}_i$ are candidate explanatory variables, $\mathbf{ES}(j,z)$ (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). Model parameters are computed from data by solving a linear inverse problem in order to guarantee a two-ways (i.e., unique) relationship between each model structure and its parameters ([2]). From a regressive standpoint, EPR may produce a non-linear mapping of data (like that achievable by Artificial Neural Networks although with few constants. 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. Such a flexible coding of mathematical expressions permits to explore the space of the models as the combinatorial space of exponents in Eq. (1). 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, [4]) 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). Due to these features, 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 is available as an add-in function for Excel (Microsoft-Office®) at `www.hydroinformatics.it`.

1.2 Study area, monitoring network and rainfall time series

The proposed method has been applied to the monthly total rainfall time series originating from 81 stations irregularly positioned within the Apulia Region (South-Eastern Italy) (Figure 1). All gauging stations belong to the meteorological monitoring network of the Hydrographic Services of Land Protection Department of the Apulia Region. The time series range from January 1931 to December, 2010. The elevation of each station ranges from 2.00 m a.s.l., (Manfredonia station) to 954.00 m a.s.l. (Pescopagano station) and the average distance between the monitoring stations is around 120 km with a standard deviation of 26 km.

2 Results and Discussion

2.1 Using EPR in missing data reconstruction

In this paper, EPR is used to infill artificially created gaps in rainfall monthly data for the measurement gauge of Canosa ($P^{Can}(t)$), using the observed rainfall monthly data of Cerignola(P^{Cer}) and Andria (P^{And}) gauges. Available monthly rainfall data cover the period from January 1926 to December 2004, without gaps in data for the three rainfall gauges. Data until December 1990 were used as training data to develop EPR-MOGA models. Data from January 1991 to December 2004 where used as test data assuming hypothetic randomly distributed gaps in data, as reported in Table 1. In the case study, the available data has been considered as time series. Thus the inputs used for the estimation of PCan(t) include rainfall monthly data up to 4 month before the time t (e.g., t-1, t-2, t-3 and t-4) (i.e. P^{Cer}, P^{And} and P^{Can}). The set of candidate exponents is [-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2], the maximum number of allowed polynomial terms is 3 and the presence of a bias term is admitted.

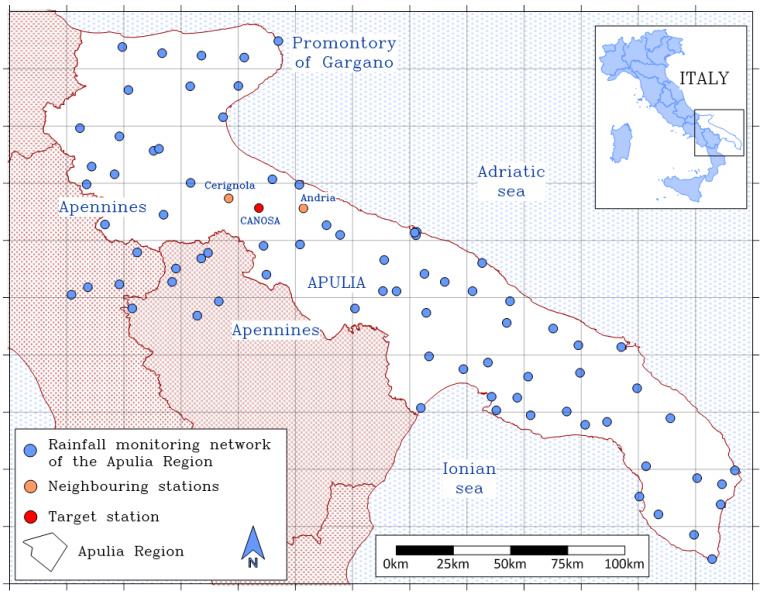

Figure 1. Study area and monitoring sites.

Length of the gap in months	Number of gaps
1	20
2	1
3	1

Table 1. Artificial gaps in the test set.

The EPR-MOGA searching procedure returned 8 models of different complexity, in which the simplest one requires the presence of the only $P^{Cer}(t)$ rainfall monthly value (with a CoD = 0.76), thus indicating that this is the most important input to estimate $P^{Can}(t)$. With the increasing of EPR models accuracy (maximum CoD = 0.82) the complexity also increase, and more data has been selected by the procedure. As a trade-off between accuracy and complexity the following model has been chosen for the present case study:

$$P^{Can}(t) = 0.91\sqrt{P^{And}(t)P^{Cer}(t)} + 4.135$$

This model is featured by a CoD = 0.813, showing the presence of the only rainfall monthly values $P^{Cer}(t)$ and $P^{And}(t)$. Note that rainfall data from previous months are selected only for the most complex models, whose increased number of terms and input variables does not result into an increased accuracy. Accordingly, they are not considered really influent for the estimation of $P^{Can}(t)$. Figure 2 shows the comparison among real and estimated values, while Table 2 shows some error indicators.

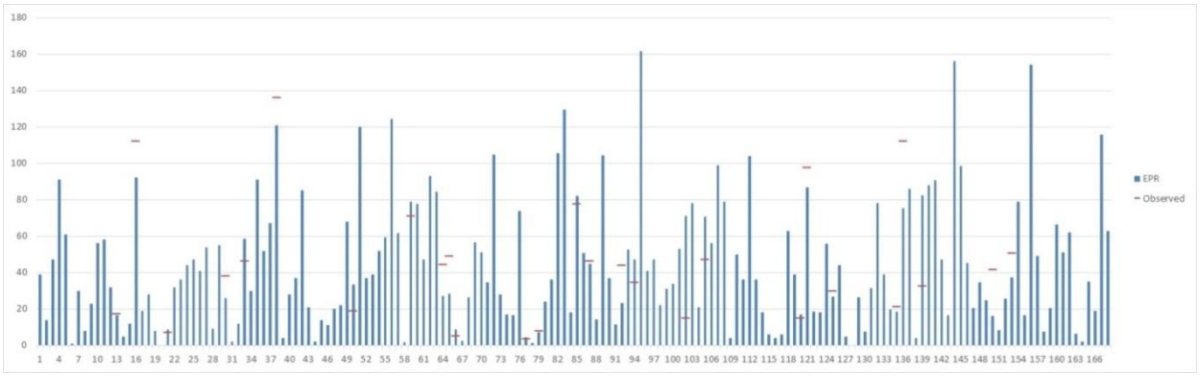

Figure 2. Comparison among real and estimated values of $P^{Can}(t)$.

Maximum Error	56.00 mm
Minimum Error	0.27 mm
Average Error	14.38 mm
Standard deviation of errors	14.44 mm

Table 2. Statistics of the fictitious gaps in the test set.

1.1 Possible future applications of EPR

In the present study, EPR-MOGA has hinted at possible application for identification of multiple correlations among rainfall gauges on a relatively wide territory. Having more time series from a range of gauge station can allow an analysis on how monthly rainfall are distributed and correlated in a (possibly) wide period of time, without excessive data-(pre)processing, but just considering the produced EPR-MOGA models and selected inputs. Considering the analysis of the single gauge station, the availability of different climatic variables (e.g., temperatures, day-night temperature range, wind speed, etc.) can allow the possible correlation with rainfall, aggregated both at monthly and daily scale, eventually allowing missing data reconstruction without resorting to other gauge station records. Furthermore, the obtained result is interesting because proposes a dependence type between the monitoring target station (Canosa) and the neighbouring stations (Andria and Cerignola), different from the linear one commonly used. This approach actually suggests a different paradigm with respect the geostatistical one. In addition, more complex models using an increasing number of neighbouring stations need to be investigated possibly resulting into more reliable filling of missing data.

References

[1] Draper, N.R., Smith, H. (1998) *Applied Regression Analysis*. John Wiley & Sons, New York, 1998
[2] Giustolisi, O., Savic, D.A. (2006).A symbolic data-driven technique based on evolutionary polynomial regression. *J. Hydroinformatics*, **8**, 207-222.
[3] Giustolisi, O., Savic, D.A. (2009).Advances in data-driven analyses and modelling using EPR-MOGA. *J. Hydroinformatics*, **11**, 225-236.
[4] Laucelli, D., Giustolisi, O. (2011) Scour depth modelling by a multi-objective evolutionary paradigm. *Environmental Modelling & Software*, **26**, 498-509.