Giustolisi, Orazio; Savic, Dragan A.; Laucelli, Daniele

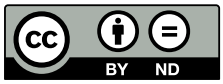
Data Mining for Management and Rehabilitation of Water Systems: The Evolutionary Polynomial Regression Approach

Dresdner Wasserbauliche Mitteilungen

Zur Verfügung gestellt in Kooperation mit/Provided in Cooperation with:
Technische Universität Dresden, Institut für Wasserbau und technische Hydromechanik

Verfügbar unter/Available at: https://hdl.handle.net/20.500.11970/103916

Vorgeschlagene Zitierweise/Suggested citation:
Giustolisi, Orazio; Savic, Dragan A.; Laucelli, Daniele (2004): Data Mining for Management and Rehabilitation of Water Systems: The Evolutionary Polynomial Regression Approach. In: Technische Universität Dresden, Institut für Wasserbau und technische Hydromechanik (Hg.): Risiken bei der Bemessung und Bewirtschaftung von Fließgewässern und Stauanlagen. Dresdner Wasserbauliche Mitteilungen 27. Dresden: Technische Universität Dresden, Institut für Wasserbau und technische Hydromechanik. S. 285-296.



Wasserbauliche Mitteilungen (2004) Heft 27
Institut für Wasserbau und THM der TU Dresden

Data Mining for Management and Rehabilitation of Water Systems: The Evolutionary Polynomial Regression Approach

Orazio Giustolisi, Dragan A. Savic, Daniele Laucelli

Risk-based management and rehabilitation of water distribution systems requires that company asset data are collected and also that a methodology is available to efficiently extract information from data. The process of extracting useful information from data is called knowledge discovery and at its core is data mining. This automated analysis of large or complex datasets is performed to determine significant patterns among data. There are many data mining technologies (Decision Tree, Rule Induction, Statistical analysis, Artificial Neural Networks, etc.), but not all are useful for every type of problem. This paper deals with a novel data mining methodology for pipe burst analysis, which integrates numerical and symbolic regression. This new technique is named Evolutionary Polynomial Regression and uses polynomial structures whose exponents are selected by an evolutionary search, thus providing symbolic expressions.

A case study from UK is presented to illustrate the application of the Evolutionary Polynomial Regression methodology to prediction of main bursts and to identification of the network features influencing them.

Keywords: Water Distribution Systems, Bust Risk Analysis, Data Mining, Modelling, Evolutionary Polynomial Regression.

1 Introduction

Economic and social costs of pipe bursts are on the increase and there is a clear need to predict pipe failure rates for water supply systems. However, we have to accept that processes causing pipe bursts are often difficult to describe using classical mathematical tools. Furthermore, collecting all potentially useful information for a full physical description of such a phenomenon is very expensive, because of the large (often unknown) domain of the problem. The use of classical statistical analysis to describe the influence of particular parameters is also a difficult task because of the complexity of the physical system. The most commonly identified parameters thought to have an influence on pipe bursts are the pipe age, material and diameter. Other parameters, such as soil corrosivity,

meteorological conditions, traffic loading, internal pressure and external stress may also be important, but are difficult to obtain. Considering aforementioned difficulties, water managers have to make long-term decisions about strategic investments into their assets and urgently need a robust and reliable tool to do assess the need for investment.

With improvements in monitoring of water systems, data-driven techniques are becoming more interesting and useful. In particular, data mining is an obvious approach to investigate, as it can deal with a high degree of complexity within the given data, see *Fayyad et al. (1996)*. Data mining is the search for valuable information in large data sets, trying to discover patterns in the data. Data mining techniques can be used to classify data records and to allow for the creation of new hypotheses about the system behaviour. For example, *Savic et al. (2003)* used Classification Tree and Rule Induction algorithms as data mining strategies for water systems management purposes.

This paper deals with a novel data mining methodology for prediction of failures in water mains and development of hypotheses for the possible causes of pipe bursts. This new technique is Evolutionary Polynomial Regression (EPR), which integrates numerical and symbolic regression. The EPR strategy provides symbolic expressions that can be used and manipulated before, during and after the data mining process. Mainly, the selection of the desired parsimony level and the inclusion of known mathematical structure into the research strategy are some of the key features of EPR that are useful for data mining *(Giustolisi and Savic, 2004)*. In this way, human expertise can still play a significant role in analysing the output expressions. These features could (i) play a relevant role in planning and design of monitoring networks, (ii) help to understand what the actual usefulness of each considered factors is and (iii) allow cost savings to be achieved.

2 Data Mining

2.1 Classification of modelling techniques

In the context of this particular work, the final goal of data mining is to create a pipe bursts model to be used in water system management. The problem of building mathematical models of complex systems, such as waterworks, starting from observed data is usually called system identification. Colour coding of mathematical modelling is often used to classify models according to the levels

of prior information requirement, i.e. white box models, black-box models and grey-box models, *(Giustolisi 2003)*. Therefore, it is useful to classify EPR according to this classification (see Figure 1):

- A white-box (WB) model is a system where all necessary information is available, i.e. the model is based on first principles (e.g. physical laws), known variables and known parameters. Because the variables and parameters have physical meaning, they also explain the underlying relationships of the system;

- A black-box (BB) model is a system of which there is no prior information available. These are data-driven or regressive models, whose both the functional form of relationships between variables and the numerical parameters in those functions are not known and need to be estimated (e.g. Artificial Neural Networks);

- Grey-box (GB) models are conceptual models whose mathematical structure could be known through conceptualisation of physical phenomena or through simplification of differential equations describing the phenomena under consideration. These models usually need parameter estimation by means of input-output data analysis, but the range of parameter values is normally known (e.g. Dimensionally Aware Genetic Programming, Classification tree);

In addition to being based on first principles, white-box models have the advantage of being able to describe the underlying relationships of the process being modelled. However, the construction of white-box models can be difficult because the underlying mechanisms may not always be completely known, or because the experimental results obtained in the laboratory environment do not correspond well to the prototype environment. Due to these problems approaches based on data-driven techniques are gaining considerable interest.

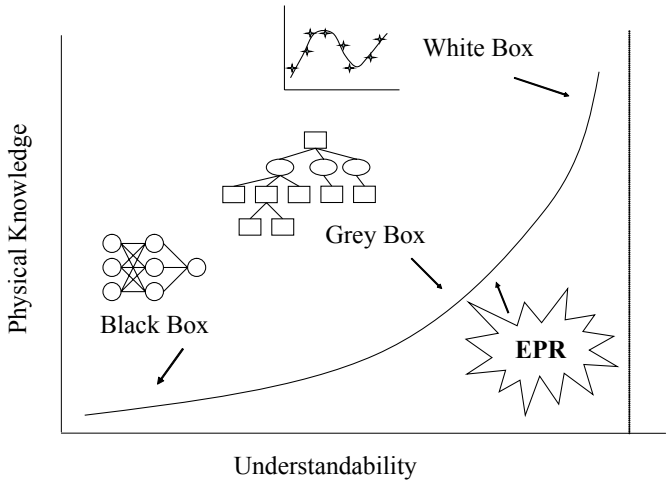


Figure 1: Data-driven modelling techniques

2.2 Some Data Mining Techniques

Artificial Neural Networks (ANNs) are the best known of BB approaches to data mining, being able to discover a model that fits data, see *Haykin (1999)*. They always require all numerical data, while non-numerical data need to be remapped into numerical information. ANNs work very well when modelling complex processes, but they do not explain the underlying relationships of the system nor they allow easy understanding of the influence of each input variable on the output results (because they are unable to produce symbolic representation of the relationships).

Among GB techniques, the C5.0 algorithm, *Quinlan (1993)*, has often been used to understand the main process and parameters involved in pipe bursts. Like many other data-mining tools, C5.0 performs at its best when there is a similar number of cases for each class created. *Savic et al. (2003)* created an additional post-processing tool needed to perform post data-mining rule optimisation, which allows the combination, alteration and visualisation of rules from a variety of sources. The tool allowed inclusion of rules added by a human expert or derived from different machine learning processes.

2.3 A new approach: EPR

EPR is a multipurpose tool, see *Giustolisi and Savic (2004)*, with a high level of flexibility, which also allows interactive development of models (see Figure 2). A user is free to tune each rule of the search according to his/her needs and

available information. Therefore, EPR can be used to describe the relationships among data (i.e. when or which pipe will fail) and then to discover new knowledge about the studied phenomenon (i.e. what factors influence pipe breaks). EPR produces simple symbolic expressions consisting of functions and constant values. According to his/her preferences and expertise, the user can select the level of parsimony and which functions are to be used in EPR. This is important because the result of applying the methodology could be a simple symbolic model with some physical insight that is easy to use by engineers.

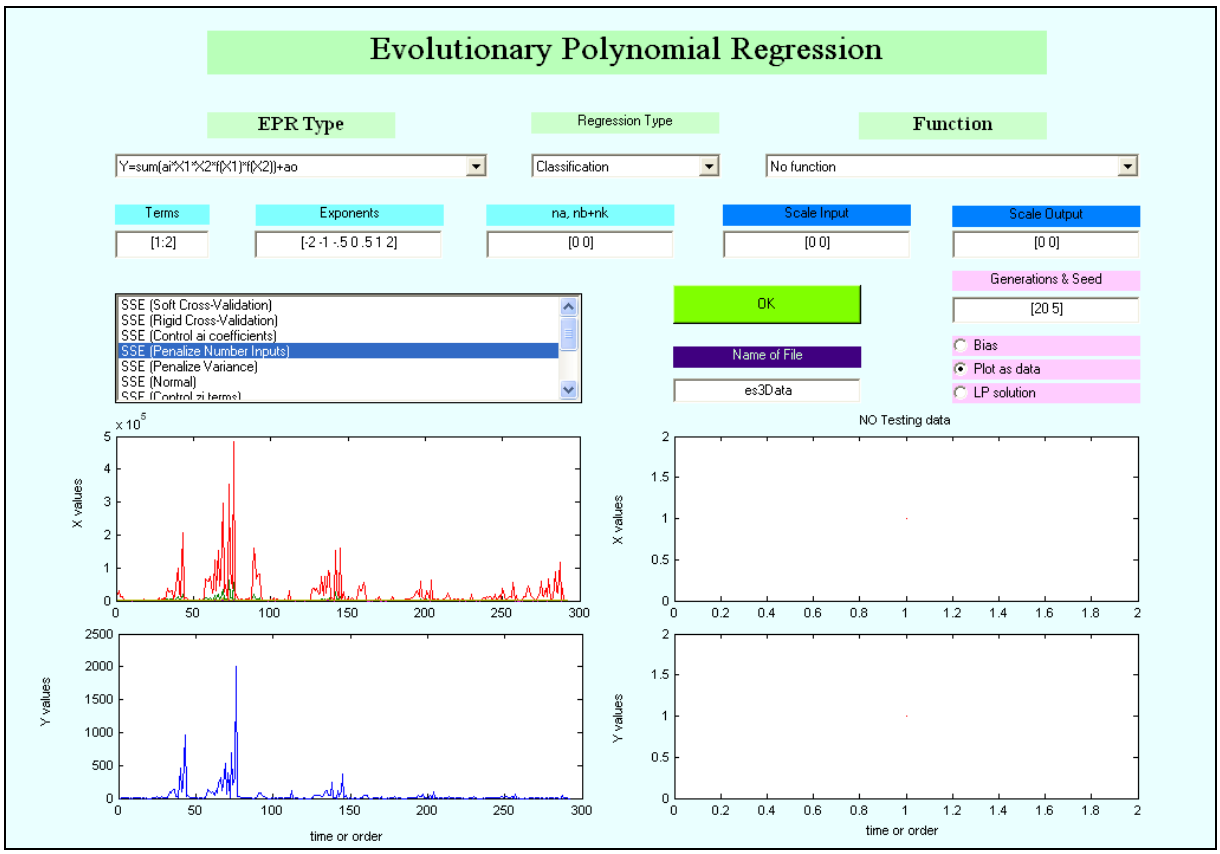


Figure 2: EPR tool user-friendly interface.

2.3.1 Brief overview on EPR

EPR is a data-driven hybrid tool, aimed at searching for polynomial structures. We can represent the general EPR expression as:

$$(1)\quad y = \sum_{j=1}^{m} f\left(\mathbf{X}, a_j\right) + a_0$$

where y is the estimated output of our system; a_j is a constant value; f is a function constructed in the EPR process; $\mathbf{X}$ is the matrix of input variables; and m is the length, i.e. number of terms, of the polynomial expression (excluding bias,

a_0), see *Giustolisi and Savic (2004)*. The general functional substructure represented by $f(\mathbf{X},a_j)$ is constructed in EPR by means of a genetic algorithm. The algorithm simply selects the useful input vectors from **X** to be combined. The structure of $f(\mathbf{X},a_j)$ is set by the user, according to any physical insights. The parameters a_j are computed by the least squares method, while the selection of feasible structures is performed by an evolutionary search.

In this way, starting from field data, clearly understandable expressions involving all the available parameters could be constructed. This also makes possible the identification of the most influencing factors in the data mining process (by analysing the values of polynomial exponents). In our case, EPR is able to identify those pipe features most influencing the burst life of a main.

3 Main Factors for Water Pipe Failures

Traditionally, many researchers and engineers have tried to find a clearly understandable relationship between the most commonly available pipe information and pipe burst rates. The *age* of water pipes undoubtedly plays an important role in determining the life of a main, see *Herbert (1994)*. Nevertheless, age alone is not enough to fully describe the phenomenon. *Kettler and Goulter (1985)* reported a strong correlation between age and an increase in burst rate of asbestos cement pipes. These findings indicate that *material* of water mains should also be considered. Throughout the UK, the most commonly used material is cast iron and a number of studies have investigated the failure rates of cast iron pipes. For example, *Lackington and Burrows (1994)* found bursts mainly occurring in the region of heavily corroded walls and suggested a relining of pipes with cement mortar as a protection against internal corrosion. Results of *Kettler and Goulter (1985)* suggest that in particular joints of water mains deteriorate most quickly. It has been also suggested that pipe *diameter* plays a significant role in that smaller diameter mains are subject to a higher failure rate. *Walski et al. (1986)* reported that smaller diameter pipes were more likely to fail due to circumferential cracks and hole blowouts, indicating a lack of bending strength and a high level of corrosion. Therefore, based on technical literature, *age*, *material* and *diameter* are the main parameters to be included in the analysis.

Further parameters could be the *number of pipes*, *supplied properties*, *main length*, *soil corrosivity*, *meteorological conditions*, *traffic loading*, *internal pressure*, *external stress*, etc.

4 Case Study

In this paper, the EPR approach is used to predict pipe bursts, using recorded data from a UK water distribution system. The same dataset was analysed by *Savic et al. (2003)* by means of a classification approach.

The available data are from one single company. Recorded bursts in the water system are from 1970 to 2000, at a pipe level. Pipe bursts development over time is shown in Figure 3, which, due to a confidentiality agreement, is given dimensionless.

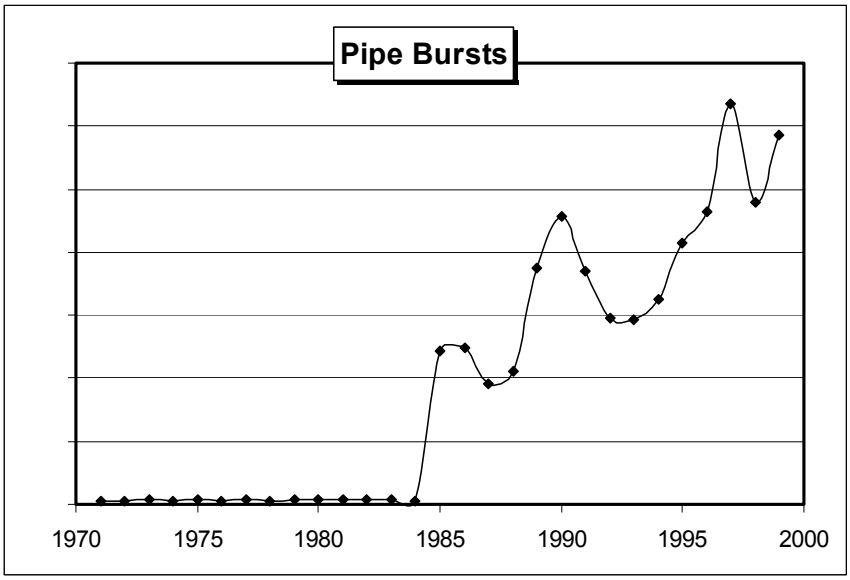


Figure 3: Development of pipe bursts over the last 30 years

Table 1 Available pipe features

Features	Values
Materials	AC, CI, ST, DI, SI, MDPE, HDPE, PVC, GRP, SOL
Year laid	From 1894 to 2000
Diameters	From 32 mm to 914 mm
Lengths	Total 6,207 km
Supplied properties	Total 555,036

Looking at Figure 3, it can be argued that data have been collected systematically during the last 15 years. In our analysis both data describing pipes with bursts and "no burst pipes" were used. It is particularly important to consider information from the entire network in order to get all the statistical correlations between those factors which cause pipe failures. For example, pipes with the same age and material should have the same statistical failure rate, but some of them had no burst recorded during the 15-year period of systematic data collection. For this reason, pipes with no receded bursts have to be included in statistical analysis. Table 1 gives a brief summary of the available information about the water system. The total number of pipes was 92,701 with the total length of about 6,200 km.

4.1 Data pre-processing and input selection

Information recorded for 92,701 pipes in the network related to pipe material, year when it was laid, pipe diameter, length, number of properties supplied by each pipe, district metering zone location, type of pipe, number of bursts, type of burst and failure date.

In order to build the EPR model, the number of bursts, year the pipe was laid, diameter, pipe length and supplied properties were chosen as the most important variables. The information about material was neglected at this stage because it was assumed that a strong correlation exists among the diameter, age of pipe and the material. There are also difficulties in re-mapping the material information into a numerical format. Furthermore, some missing data about the year pipe was laid have been reconstructed on the basis of the correlation between age and material, therefore, the information on pipe material was implicitly used. For these pipes (less than 10% of all mains), missing information on age was estimated as the average of those pipes in the same class of material and diameters. This procedure ensured that statistical properties of the whole set of data did not change significantly. In cases when pipe material was unknown, the average year the pipe was laid was deduced based on diameter. Finally, for all the pipes the *supplied properties density* was computed as a ratio between the number of supplied properties and pipe length. The "*working age*" of a pipe was computed as a difference between the last monitoring year (2001) and year its was laid.

4.2 EPR inputs

After pre-processing, the four inputs (diameter, length, working age and supplied properties) and one target output (number of bursts) were available for all pipes. Before EPR was applied to the dataset, individual records were grouped

using the working age and the range of diameters as classification criteria. Each working age class was 5 years long, such that the pipes that were 1 to 5 years old, 6 to 10 years old, etc., belong to the same class. The working age of the actual class ('Age') was calculated as the average value. For each class the following attributes were calculated: (1) the sum of pipe lengths (LT); (2) the sum of supplied properties (PR); (3) the number of pipes (NP); (4) the average of supplied properties density (Den); and (5) the equivalent diameter. The equivalent diameter is defined as

$$(2)\quad Deq = \frac{\sum_{class} L \cdot D}{LT}$$

where D is the single pipe diameter and L is its length.

The sum of bursts (BR) was assigned to each class and defined as a target output for EPR. Finally, EPR was applied to the aforementioned inputs (Age, LT, NP, Den, PR, Deq) to model the number of bursts for each class, having BR as the target (total bursts in 15 years).

4.3 Results and comparisons

A cost function used within EPR was aimed at penalising the complex structures by limiting the number of involved inputs, as in *Giustolisi et al. (2004)*. At the end of the search phase, EPR selected 4 inputs (out of 6) as significant. The symbolic expression obtained is

$$(3)\quad BR = 6.4811 \cdot 10^{-5} \cdot Age^{2} \cdot LT^{1} \cdot Deq^{-1} \cdot NP^{-1}$$

which has a Coefficient of Determination (CoD) = 0.92 and correctly describes the 95.7 % of all the recorded bursts. Considering Eq. (3) under a knowledge discovery point of view, one can say that the age and pipe length are proportional to the number of pipe bursts. The opposite is true for the equivalent diameter, i.e. smaller number failures is observed on larger diameter pipes, which agrees with findings of *Walski et al. (1986)*. Moreover, in Eq. (3) we found that the average length of the pipes belonging to the same class is influent in determining the total bursts of the relative class, so increasing our knowledge about statistical correlations in our water system.

4.4 How to use the EPR model in water systems management

Looking at Eq. (3) a decision-maker can easily get reliable information about how many failures there will likely be in the water network during a particular time period. For example, if a user needs to know how many failures will likely happen in the next three years in order to develop an asset management strategy, he/she will use Eq. (3) to compute BR. The number of bursts in the next three years could be estimated as the difference between the recorded bursts in 2000 and predicted number of bursts in 2003.

5 Conclusions

An application of the new data mining technique for pipe bursts prediction is described in this paper. This new technique, called Evolutionary Polynomial Regression (EPR), provides symbolic expressions as a result of data analysis. The prediction model was developed and tested on a real-world case. The technique is also used to investigate the causal effects/factors, starting with the most commonly available, such as pipes age; pipe length, number of properties supplied and pipe diameter.

Data mining by EPR produced good, simple and understandable relationships/models that provide the high level of statistical correlation between the variables. The model obtained indicates that the age and pipe lengths are directly proportional to pipe bursts, while the pipe diameter is inversely proportional to the propensity of pipes to fail. These results confirm the same physical insights obtained in similar studies of different pipe networks, *Kettler and Goulter (1985); Walski et al. (1986); Lackington and Burrows (1994)* and *Savic et al. (2003)*.

From an engineering point of view, EPR appears to be a reliable prediction tool, being able to support long-term strategy decisions for asset management for water companies. However, similar to the other data-driven techniques, its usefulness will depend on the diversity and quality of the input data. This is where EPR can help identify which important information is needed and what pipe burst factors have to be properly collected.

References

Fayyad U.M., Piatetsky-Shapiro G., Smyth P., (1996), From Data Mining to Knowledge Discovery: An Overview, in "Advances in Knowledge Discovery and Data Mining", AAAI Press and the MIT Press, chapter 1, pp.1-34.

Giustolisi O. and Savic D.A., (2004), A Symbolic Data-driven Technique Based on Evolutionary Polynomial Regression, Journal of Computing in Civil Engineering, ASCE, in review.

Giustolisi O., "Using Genetic Programming To Determine Chézy Resistance Coefficient In Corrugated Channels", in press, J. of Hydroinformatics, IWA Publishing (2003).

Haykin, S., "Neural Networks: A Comprehensive Foundation 2/e", Prentice-Hall Inc., Upper Saddle River, New Jersey, (1999).

Herbert H., (1994), Technical and economic criteria determining the rehabilitation and/or renewal of drinking water pipelines, Water Supply, 12 (3/4, Zurich), pp.105-118.

Kettler A. J. and Goulter I. C., (1985), An analysis of pipe breakage in urban water distribution networks, Canadian Journal of Civil Engineering, 12 pp.286-293.

Lackington D. W. and Burrows B. L., (1994), Criteria to determine appropriate levels of investment for rehabilitation, Water Supply, 12 (3/4, Zurich), pp.21-32.

Quinlan J.R., (1993), C4.5 Programs for Machine Learning, Morgan Kaufmann Pub., San Mateo, CA.

Savic D.A., Bessler F and Walters G.A., (2003) Comparison of Data Mining Methods to Statistical Approaches For Pipe Burst Risk Analysis, IWA-IAHR joint conference: Pumps, Electromechanical Devices and Systems PEDS 2003.

Walski T. M., Wade R., Sharp W. W., Sjostrom J. W. and Schlessinger, D., (1986), Conducting a pipe break analysis for a large city, AWWA Conference Symposium, pp.387-402.

Authors:

Prof. Orazio Giustolisi

Dept. of Civil and Environmental Engineering
Faculty of Engineering
Technical University of Bari
V.le del Turismo 8 – Paolo VI
74100 Taranto (Italy)

Tel.: ++39 – 099 – 4733214
Fax: ++39 – 099 – 4733229
Email: o.giustolisi@poliba.it

Prof. Dragan A, Savic

Dept. of Engineering
School of Engineering, Computer Science and Mathematics
University of Exeter
Harrison Building, North Park Road
EX4 4QF Exeter (UK)

Tel.: ++44 – 1392 – 263637
Fax: ++44 – 1392 – 217965
Email: d.savic@exeter.ac.uk

Dr. Daniele Laucelli
Dept. of Civil and Environmental Engineering
Technical University of Bari
V.le del Turismo 8 – Paolo VI
74100 Taranto (Italy)

Tel.: ++39 – 099 – 4733210
Fax: ++39 – 099 – 4733230
Email: d.laucelli@poliba.it