Eureqa – Software Review

Renáta Dubčáková
Faculty of Safety Engineering, VŠB – Technical University of Ostrava,
Ostrava, the Czech Republic

1. Introduction

The Eureqa software, sometimes called the robot scientist [4][5], was developed at the Computational Synthesis Lab at Cornell University by Dr. Hod Lipson [1]. It uses symbolic regression for detecting equations and hidden mathematical relationships in raw data. The Eureqa algorithm is similar to Lipson's algorithm for self-contemplating robots that are able to repair themselves [5].

I have been using Eureqa for the testing of radon detection devices and protocols. While I focus on this application below I also comment on the usefulness of Eureqa for genetic programming applications more generally.

My aim is to analyse the uncertainty in the short term integrated measurement of radon by RM-1 electret detectors [2]. See Figure 1. This is to verify the manufacturer's method [2] for evaluating their detectors' measurements. This work is needed to help form strategies to protect against natural radioactive gas radon in buildings and lung cancer. This is particularly important in the Czech Republic where the average domestic radon concentration is one of the highest of Europe [3].

In detail, I was trying to map the relationship between real measured values and values calculated according to manufacturer's manual [2]. I first used the statistical tool Statgraphics Plus [6] to do multiple linear regression. Traditional linear and nonlinear regression methods can fit parameters of a given model. However, in my case, the model is not known.

I heard about the Eureqa software on the Czech Radio Leonardo in December 2009 [8]. They said that using real datasets Eureqa had found, among others natural laws, Newton's second law of motion and also the law of conservation of momentum. I decided to try Eureqa. I was very positively surprised: I had not expected that there would be a general user friendly software available to non-GP-experts.

2. Application of the Eureqa software on the radon measurement evaluation

Fig. 1 Electret ion chamber detectors (200 ml volume)

Radon measurements were taken in a family house in the Czech Republic over five weeks. We obtained five datasets of 26 measured values. Due to measurement drift and other forms of noise, Electret detectors are used in pairs. The first idea of my analysis was to verify the significance of the smaller value of each pair and then try to simplify the original method of evaluation (as given in the manual pages). So, all possible combinations of two

measured values were created and the estimations of the true value were calculated according to the manual pages [2]. In summary, 1625 combinations were tested [9].
I easily entered this data into Eureqa using its spreadsheet like interface. My data were represented as two columns of independent measurements and a column of data calculated according to the manual. Eureqa was told that calculated data were the independent values.

Next Eureqa requires the function set to be chosen. (Eureqa calls this Pick Modelling.) I chose addition, subtraction, multiplication and division, together with both Gaussian and exponential functions. Eureqa supports several fitness functions. They include for example the mean absolute error (MAE), correlation coefficient, and maximum error. I used MAE.

Eureqa also supports parallel operation over multiple computers. This was quite successfully and in the end I used two computers for my experiments. Figure 2 shows the user interface used to start and stop Eureqa.

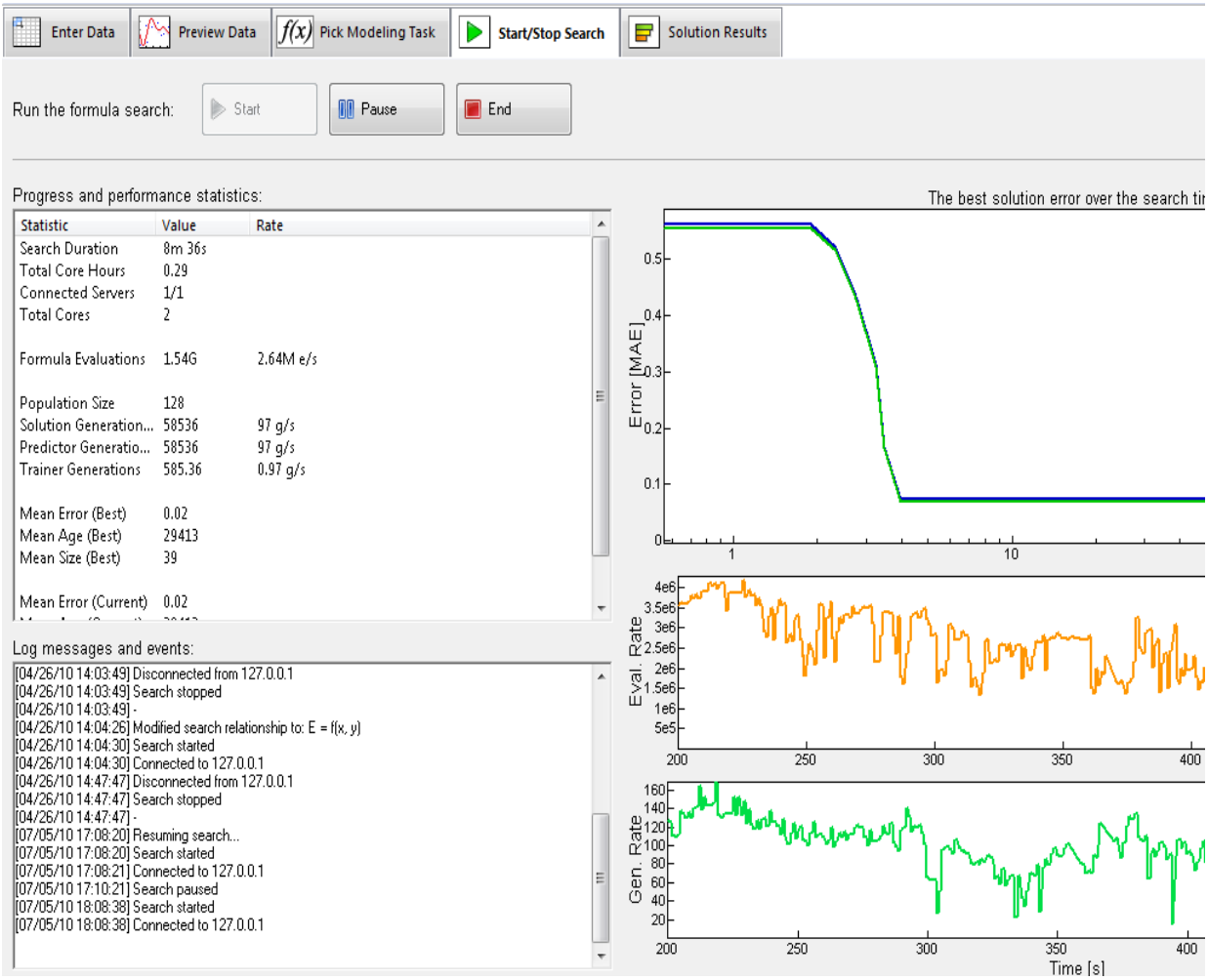

Fig. 2 Progress of model searching by Eureqa

3. Eureqa results and comparison with multiple linear regression model

As well presenting graphs of the population as it evolves, Eureqa can display the new equations as it find them.

Eureqa splits automatically training and validation dataset. But you can set the percentage of the training or validation data of measurements. The user can split dataset by one of the following ways:

- conjunctive: e.g. first 95% of data are used as training and last 50% are used as validation (Fig. 3). It means that last (remaining) 5% are used only as validation and 45% of dataset are used both as training and as validation.

- disjunctive where the user ensures that training and validation sets are kept separate: e.g. first 95% into training and last (remaining) 5% into validation.

For experiments, I used the conjunctive split of the data set. I chose also the randomly shuffle before splitting the data set (Fig. 3).

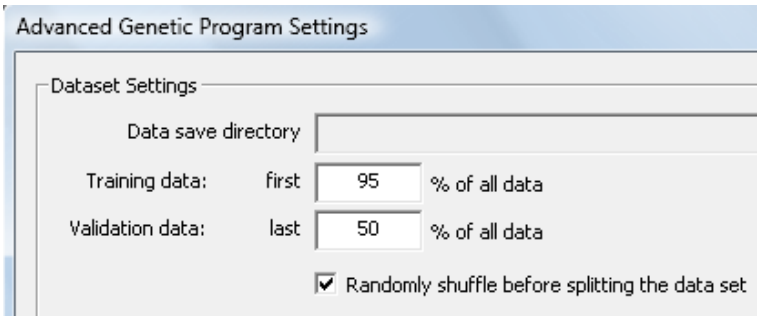

Fig. 3 Properties of training and validation data setting

Figure 4 shows an example of model solutions and their properties, including their score on training and validation dataset and statistical properties.

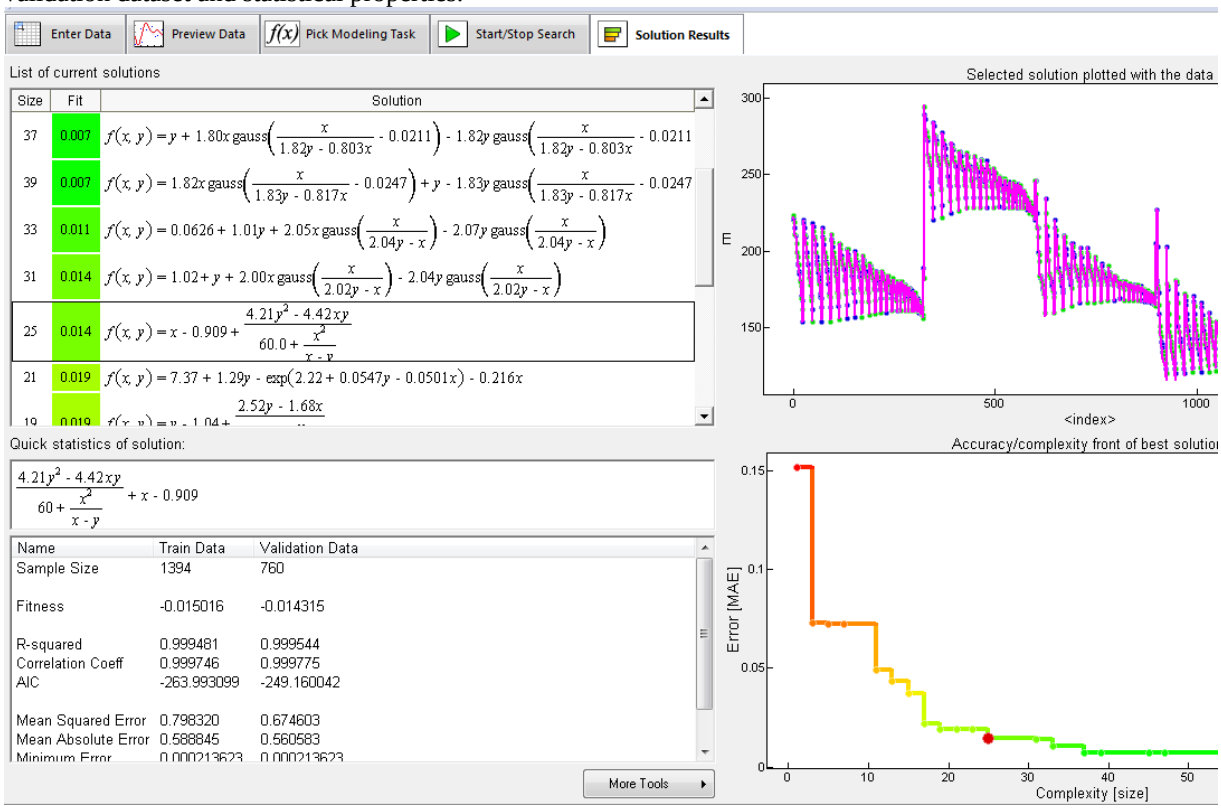

Fig. 4 Eureqa "Solution Results" sheet

The Eureqa software was able to find various possible models. But some models are not defined for special values. For example, the regression model

$$f(x,y) = (4.20908*y*y - 4.41995*x*y)/(60.0237 + x*x/(x - y)) + x - 0.909361$$

is not defined for the value y = x, because of a division by zero error.

Eureqa produced similar but not identical models in other indepenent runs. It is associated with properties of training and validation data settings (ticking the setting "Randomly shuffle before splitting the data set"). The final linear model was chosen by hand.

It was necessary to use Eureqa because it found also several alternative models which has the similar behaviour to original model. The maximum relative erorr of some models were less than ±1%. They use also the same function set as manufacturer. For example:

f(x, y) = y + 1.1171*x^2*gauss(1.1171*x^2/(1.07247*y^2)/y - 1.13703*y*gauss(1.1171*x^2/(1.07247*y^2)),

where gauss(x) = exp(-x^2), R-squared = 0.99986, MAE = 0.37.

On the other hand the linear model is easier to use and the model erorr is negligible compared to the uncertainty of levels of indoor radon concentration.

The finally selected model was chosen under the highest value of R-Squared and minimal mean absolute error (MAE) with respect to simplicity and numerical stability of the model. In the end, we selected a linear regression model.
For comparison purposes, the same dataset was also analyzed using multiple linear regression (using Statgraphic Plus). The two models are very similar and are not statistically different, see the Table 1. P-value of the regression constant -0.466649 was equal to 0.03, which is higher than 0.01. It indicates that the regression constant of the S(x,y) model is not statistically significant at the 99% confidence limit.

Table 1 Comparing Eureqa model vs. multiple linear regression model

Using the software:	Selected models	R-Squared	MAE
Statgraphics Plus	S(x,y) = -0.466649+0.0099157*x+1.02251*y	0.9927	2.83
Eureqa	E(x,y) = 0.0147201*x+1.01696*y	0.9928	2.86

4. Conclusions: How well Eureqa works

Eureqa software provides user friendly interface suitable also for non-expert GP users. Results of the symbolic regression are presented also in intuitive way. Moreover, the software can be freely down loaded. The Eureqa documentation guides users through all the steps in their data analyses. Each step is represented in Eureqa by one its graphical interface sheets.

This is why Eureqa continues to play a big role in my research.

The user guide offers only basic information. For example, the exact mathematical definitions of some statistical characteristics are missing. On the other hand, tips and examples on the blog are very useful. Moreover, the authors of the Eureqa software were ready to answer user questions.

The Eureqa software is not like standard statistical tools because of its symbolic regression engine. Eureqa verifies its models on automatically chosen validation data. Parameters of the model verification are given in the result sheet. The user can choose the amount of training or validation data or use the default settings. It seems that Eureqa uses a robust approach for creating training and validation datasets, which may be non-standard however the model found by Eureqa has already been successfully tested on additional datasets. For example datasets provided by the National Radiation Protection Institute in Prague. By repeating the calculation I found several alternative models with very small statistical errors.

Eureqa helped me to find two models: a very accurate nonlinear model with similar behaviour to the original manufacturer model and an alternative linear model which is easier to use. The nonlinear model chosen by Eureqa is the one with the best fit and the linear model was chosen by me because of its simplicity. The nonlinear model contains similar functions to those in the original manufacturer‘s model and yet incredibly the maximum relative error is less than 1%.

The linear model founded by Eureqa was also compared with a linear model founded by the Statgraphic Plus software. Unlike Statgraphics, it seems that Eureqa cuts out statistically insignificant variables. The Eureqa linear model was helpful for simplification of uncertainty modelling. Moreover, the manufacturer of the analysed

electret device is interested in the alternative model found by Eureqa. The detailed evaluation and testing of the alternative model will be the subject of the future research.

ACKNOWLEDGEMENT
I would like thank Pavel Praks and William B. Langdon for their useful comments on an earlier draft of this review. This research has been supported by the Ministry of Education, Youth and Sport of the Czech Republic, Project No. 1M06047.

5. References

[1] Schmidt M., Lipson H., Distilling Free-Form Natural Laws from Experimental Data, Science, Vol. 324, No. 5923, 2009, pp. 81 - 85.
[2] Manual of integrated electret ion chamber system RM-1, Dr. Froňka NUCLEAR TECHNOLOGY. Prague, April 2003.
[3] Dubois G. An overview of radon surveys in Europe. Luxembourg, Office for Official Publications of the European Communities, 2005
[4] Edwards, Lin: Eureqa, the Robot Scientist. Available online: http://www.physorg.com/news179394947.html
[5] Brandon Keim: Download Your Own Robot Scientist. Available online: http://www.wired.com/wiredscience/2009/12/download-robot-scientist/
[6] STATGRAPHICS Plus Version 5.0 Professional Edition. Statistical Graphics Corp. 1994 – 2000.
[7] ISO/IEC Guide 98-3:2008 Guide to the Expression of Uncertainty in Measurement.
[8] Vachtl, P.: Download the software scientist (Stáhněte si softwarového vědce). Czech Radio Leonardo. Available online: http://www.rozhlas.cz/leonardo/technologie/_zprava/674688 [cited: 30th December 2009].
[9] Dubčáková, R., Praks, P.: Statistical verification of the method of evaluation the short term radon integrated measurements. In: 16th International Conference on Soft Computing, MENDEL 2010. Radomil Matoušek, (Ed.). Brno: University of Technology, 2010, Brno, Czech Republic.