Genetic Programming within a Framework of Computer-Aided Discovery of Scientific Knowledge

Maarten Keijzer
DHI Water & Environment
Hørsholm, Denmark
mak@dhi.dk

Vladan Babovic
DHI Water & Environment
Hørsholm, Denmark
vmb@dhi.dk

Abstract

Present day instrumentation networks already provide immense quantities of data, very little of which provides any insights into the basic physical phenomena that are occurring in the measured medium. In order to fully exploit the information contained in the data, dimensionally aware GP has been developed. The present paper presents two application studies of dimensionally aware GP on difficult hydraulic data sets.

1 INTRODUCTION

In making the most of a set of experimental data it is generally desirable to express the relation between the variables in the symbolic form of an equation. In view of the necessarily approximate nature of the functional relation, such an equation is described as 'empirical'. No particular stigma should be attached to the name since many ultimately recognised chemical, physical and biological laws have started out as empirical equations.

Sciences devote particular attention to the development of a physical symbol system, such as a scheme of notation in mathematics, together with the evolution of more refined representations of physical and conceptual processes in the form of equations in the corresponding symbols. Each equation can be regarded as a collection of signs, which constitutes a *model* of an object, process or event. Data, on the other hand, remain as 'mere' data just to the extent that they remain a collection of signs that does not serve as a model. From this point of view, the evolution of an equation within a physical symbol system as a means of better conveying the 'meaning' or 'semantic content' that is encapsulated in the data, corresponds to the evolution of a model. Evidently the 'information content' is very little changed, or even unchanged, but the 'meaning value' is commonly increased immensely. Since it is just this increase in 'meaning value 'that justifies the activity of substituting equations for data, there is a natural interest in processes for further promoting such means.

1.1 COMPUTER-SUPPORTED SCIENTIFIC KNOWLEDGE DISCOVERY

Means for data collection and distribution have never been so advanced as they are today. While advances in data storage and retrieval continue at an extraordinary rate, the same cannot be asserted about advances in information and knowledge extraction from data. Without such developments, however, we risk missing most of what the data have to offer. Disregarding the data simply because we do not know how to analyze it would be a real waste. This is particularly pronounced in scientific endeavours, where data represent carefully collected observations about particular phenomena that are under study.

However, analyzing the data *alone* is not the entire story. At least not in scientific domains! Scientific theories encourage the acquisition of new data and these data in turn lead to the generation of new theories. Traditionally, the emphasis is on a theory, which demands that appropriate data be obtained through observation or experiment. In such an approach, the discovery process is what we may refer to as *theory-driven*. Especially when a theory is expressed in mathematical form, theory-driven discovery may make extensive use of strong methods associated with mathematics and with the subject matter of the theory itself. The converse view takes a body of data as its starting point and searches for a set of generalisations, or a theory, to describe the data parsimoniously or even to explain it. Usually such a theory takes the form of a precise mathematical statement of the relations existing among the data. This is the *data-driven* discovery process. However, there is an enormous amount of knowledge and understanding of physical processes that should not just be thrown away. Therefore, we strongly believe that the most appropriate way forward is to combine the best of the two approaches: theory-

driven, understanding-rich with data-driven discovery processes.

The process of scientific discovery has long been viewed as the pinnacle of creative thought. Thus, to many people, including some scientists themselves it seems an unlikely candidate for automation by a computer (Langley, 1998). However, over the past two decades researchers in AI have repeatedly questioned this attitude. The present paper is a modest attempt to describe the use of GP within a scientific discovery framework.

1.2 MODEL INDUCTION

One particular mode of data mining is that of model induction. Inferring models from data is an activity of deducing a closed-form explanation based on observations. These observations, however, always represent (and in principle only) a limited source of information. The question emerges how this, a limited flow of information from a physical system to the observer, can result in the formation of a model that is complete in a sense that it can account for the entire range of phenomena encountered within the physical system in question — and to even describe the data that are outside the range of previously encountered observations. The confidence in model performance can not be based on data alone, but might be achieved by grounding models in the domain so that appropriate semantic content is obtainable. This should be the ultimate goal of knowledge discovery.

Thus, a model induction algorithms that produce models amenable to interpretation next to the ability to fit data is needed. Clearly, every model has its own syntax. The question is whether such syntax can capture the semantics of the system it attempts to model. Certain classes of model syntax may be inappropriate as a representation of a physical system. One may choose the model whose representation is complete, in the sense that a sufficiently large model can capture the data's properties to a degree of error that decreases with an increase in the model size. Thus, one may decide to expand Taylor or Fourier series to a degree that will decrease the error to a certain, arbitrarily given degree. However, completeness of representation is not the issue. The issue is in providing an adequate representation amenable to interpretation.

1.3 THE ROLE OF A SCIENTISTS IN A COMPUTATIONAL DISCOVERY PROCESS

The term computational discovery appears to imply a fully automated process. Indeed, most of the research in AI and specifically in GP may suggests so, simply because it is the automation process that is at the center of attention. However, the most appropriate use of model induction algorithms is the one in which scientists and domain specialists play active role. Langley (1998) summarises the major ways in which scientists can influence the behaviour of discovery systems as:

- *problem formulation*: the discovery problem must be formed so that is can be solved using an induction algorithm. This phase covers problem definition and related choice of dependent and independent variables
- *effective representation*: the background knowledge about the domain in terms of initial theory or previous results can be incorporated through appropriate *representational engineering*.
- *data manipulation*: collected data may be sparse, incomplete, noisy or include outliers. Consequently, data often need to be manipulated in order to improve the results through computational discovery.
- *algorithm manipulation*: setting of the inductive system's parameters
- *postprocessing*: transformation of the inductive system's output into a form which is meaningful to scientific community.

2 DIMENSIONALLY-AWARE GP

In building empirical equations for a physical phenomenon based on data alone, units of measurement form a principal tool that help in interpreting these equations. One standard approach in avoiding potential conflicts with incorrect dimensionality of induced formulations is to use dimensionless values (well known examples are the Mach number and the Reynolds number). This is the 'standard scientific practice'. Units of measurements are effectively eliminated through the introduction of dimensionless ratios. Once the dimensionless numbers are used instead of the original dimensional values the problem of dimensional correctness is conveniently avoided, as all analysed quantities are dimension-free and can be used by knowledge-free induction tools such as regression, neural networks and genetic programming. Dimensionless numbers themselves can be proposed by introducing ratios that seem to make sense, or by the more systematic method of applying Buckingham's Pi-theorem. It is also argued that dimensionless ratios collapse the original search space, making it more compact, thus resulting in a more effective behaviour of algorithms that fit models to the data. At the same time, the information contained in the units of measurements is ignored entirely, effectively violating the basic premise of dimensional analysis.

The resulting equations can then be tested with statistical methods to examine their ability to predict the phenomenon on unseen data. Although physical laws are preferably stated in dimensionless form (Ellis, 1965) an empirically found relationship stated in the problem's units can aid interpretation and subsequently can lead to a better understanding of the process in question.

Dimensionally aware genetic programming (Keijzer & Babovic, 1999) differs from the approach sketched above in that the raw observations are used together with their units of measurement. The system of units of measurement can be viewed as a typing scheme and as such can be used in some form of typed genetic programming. One candidate for this is a strongly typed approach (Montana 1995, Clack & Yu 1997), where the population is initialized with correctly typed equations only and this correctness is maintained during the run. In the case of ill-posed problems or problems where the measured data gives an incomplete picture of the entire problem, a strongly typed approach suffers from the fact that it cannot propose equations that are *more-or-less* correct. Although the object of search is a correctly typed equation in terms of the dimensions, at any time there is a balance between the accuracy of the formulation and the dimensional correctness. When these two objectives for a given problem are contradictory, an important indication that the problem is ill-posed can be given. An example of such a situation can be found below in section 3.2, where an empirical equation was discovered that was not stated in the desired units but which was amenable for subsequent analysis.

The dimensionally aware approach proposes what can be called a *weakly typed* or *implicit casting* approach. Dimensional correctness is not enforced, but promoted. An extra objective for selection, goodness-of-dimension, is introduced that is used in addition to a goodness-of-fit objective. These two objectives are used in a multi-objective optimization routine using the concepts of dominance and Pareto optimality. Goodness-of-dimension is measured by calculating how many constants with appropriate units should be introduced to render an equation dimensionally correct. The fewer are needed, the better the equation's goodness-of-dimension.

In contrast with strongly typed approaches where the burden of typing is implemented in the language itself (most notably in the initialization, crossover and mutation routines), this weakly typed, or casting approach puts the burden of typing in the selection component of the algorithm. The search space is subsequently not reduced, but transformed: selection pressure is added towards correctly typed formulations, but it is not enforced so that all proposed equations are correctly typed.

In the perspective of the bias/variance trade-off when applying genetic programming to a regression-like problem (Keijzer & Babovic 2000), the weakly typed approach introduces less bias in the search than a strongly typed approach and will subsequently imply a larger variance in the resulting formulae. The larger variance is helpful in a process of discovery as it will produce competing equations of varying competence. It is then the task of the user to reduce this variance by employing background knowledge (*representational engineering*). It is our view that in a process of scientific discovery it is more helpful to allow the user to introduce background knowledge when confronted with hypotheses about the problem, rather than insisting in reducing the search space even before it is clear how much information is actually contained in the experimental data. The dimensionally aware approach attempts at introducing enough bias to get useful results, yet without sacrificing general applicability.

The result of a single run of such unit typed genetic programming is a number of equations — a so-called Pareto front of non-dominated solutions — that balance dimensional correctness (goodness-of-dimension) with goodness-of-fit. The role of the user is then to choose the most suitable formulation to further analyze the proposed relationships (*postprocessing*). When the problem is well-posed, the user can proceed by choosing the dimensionally correct formulation, yet when not all data is present the difference in goodness-of-fit between correct formulations and slightly incorrect ones might lead to the selection of an incorrect formulation. The user can exploit background knowledge or implement some belief about the problem domain. The final step lies in examining the selected equation(s) in order to interpret them. Here the user can relate elements of the equation to the actual processes that are under investigation. When a reasonable explanation for the apparent goodness-of-fit of such an equation is produced, the user's belief in the correctness of the equation is enhanced. The equation then no longer functions as a black box for making accurate predictions but as a genuine empirical equation that can be used with more confidence than mere statistical security. The equation and corresponding interpretation is amenable to review by experts and peers. The interpretation step is exceedingly difficult using dimensionless ratios alone. The sections below will give a few examples of this new method of induction of empirical equations.

3 CASE STUDIES

In the sequel two case studies of knowledge discovery using genetic programming are presented. Both of these cases present results in which GP offers results superior to those proposed by human experts.

3.1 CONCENTRATION OF SUSPENDED SEDIMENT NEAR BED

To test the performance of GP within a framework of scientific knowledge discovery, experimental flume data utilized by Zyserman and Fredsøe (1994) were analysed. The experimental data consisted of total, steady state sediment load for a range of discharges, bed slopes and water depths. Zyserman and Fredsøe used the Engelund-Fredsøe and Einstein formulation to calculate the bed concentration of suspended sediment c_b and used these values in conjunction with hydraulic parameters to perform system identification and formulate the expression for bed concentration of suspended sediment c_b. The hydraulic conditions were represented by Shields parameter θ, defined as:

$$\theta = \frac{u_f}{(s-1)gd} \text{ and } \theta' = \frac{u_{f'}}{(s-1)gd} \quad (1) \text{ and } (2)$$

where:

u_f -shear velocity $=(gDI)^{0.5}$

s -relative density of sediment

d_{50} -median grain diameter

D -average water depth

I -water surface slope

u_f' -shear velocity related to skin friction $=(gD'I)^{0.5}$

D'-boundary layer thickness defined through:

$$\frac{v}{u_{f'}} = 6 + 2.5\ln\left(\frac{D'}{k_N}\right) \quad (3)$$

v -mean flow velocity

w_s -settling velocity of suspended sediment

k_N -bed roughness $=2.5d_{50}$

An interesting observation is that all 'directly measurable' quantities do not correlate as well with the concentration of sediment c_b as the derived dimensionless quantities θ and θ'. For example, the correlation coefficient between c_b and u_f' amounts to 0.784 and between c_b and u_f to 0.628. At the same time, correlation between c_b and θ' amounts to 0.894 and between c_b and θ to *0.711*. Bearing such strong correlations in mind, it is a little surprise that, after dimensional analysis, Zyserman and Fredsøe (1994) formulated the following expression:

$$c_b = \frac{0.331(\theta' - 0.045)^{1.75}}{1 + \frac{0.331}{0.46}(\theta' - 0.045)^{1.75}} \quad (4)$$

so that c_b is a function only of θ'. Comparative analysis with some other and more complex expressions involving more variables presented in their 1994 paper, has shown that formula (4) is of comparable, if not higher accuracy.

3.1.1 Results Based on Standard Genetic Programming

Firstly, a standard genetic programming environment was set-up in such a way as to comprehend all corresponding parameters based on both directly observed the derived quantities. The evolutionary process resulted in a number of expressions, of which only the best performing is presented. The best performing expression can be written in an ordinary notation as:

$$c_b = \frac{0.31\left(\theta' - \frac{\theta}{w_s}\right)}{0.65\left(\theta' - 0.403\sqrt{d_{50}}\right)^{1.66} + 1.11} \quad (5)$$

Statistical measures of accuracy for the equation (5) are given in Table 1. The degree of accuracy is rather high, and it offers an improvement over the human-proposed equation (4). However, there is an immediate question of interpretability of such an equation. The formula above is dimensionally incorrect, and it is rather difficult to interpret it in physical terms. This is an example of pure fitting.

3.1.2 Results Based on Dimensionally Aware Genetic Programming

By way of comparison, a dimensionally aware genetic programming environment was set-up to comprehend all measured data and *not* the corresponding dimensionless parameters based on the measurements. The purpose for conducting such experiment was to test whether such a GP setup is capable of creating a dimensionally correct and still accurate formulation. Since the pre-processing of raw observations (formation of dimensionless θ and θ') was not employed here, it can be argued that GP was confronted with a problem of trying to formulate a solution from first principles. The evolutionary processes resulted in a number of expressions, of which only the most interesting one is presented here:

$$c_b = 1.12\,10^{-5}\,\frac{(u_f' - w_s)\left(1 + 100\frac{u_f' w_s}{gd_{50}}\right)}{u_f + u_f'} \quad (6)$$

The degree of accuracy of the induced expression is quite satisfactory. A statistical measure of conformity, such as the coefficient of determination, gives a value of 0.82. This provides an improvement over the value of 0.81 based on the Zyserman-Fredsøe relationship (Eq.5). At the same time, the formula is dimensionally correct, it uses the most relevant physical properties in the relevant

context. For example, the dimensionless term $\frac{u_f' w_s}{g d_{50}}$ is effectively a ratio of shear and gravitational forces. Shear forces are represented by u_f', 'responsible' for elevating sediment particles into the stream, while the gravitational term $\frac{g d_{50}}{w_s}$ is 'responsible' for settling the particles. The remaining group $\frac{(u_f' - w_s)}{u_f + u_f'}$ is a ratio of resultant energy near the bed and of the total available energy in the flow transporting the particles.

Table 1 Statistical summary for expressions (4), (5) and (6) - where: r denotes the correlation coefficient, R^2 Pearson's product moment correlation squared, RMS Root Mean Squared Error, NRMS Normalised Root Mean Squared Error RMS is normalised by the standard deviation of the desired outcome.

	r	R^2	*RMS*	*NRMS*
(4)	0.89	0.81	0.049	44.46 %
(5)	0.90	0.82	0.048	42.88 %
(6)	0.90	0.82	0.048	43.74%

Formula (6) offers a marginal improvement regarding accuracy over the formula induced through standard scientific practice. However, the simple fact that this formula was induced through automatic means based on raw data and that it provides a competing view on the importance of the processes occurring in this phenomenon is very exciting indeed. It may even be argued that expression (6) can be more easily interpreted than the Zyserman-Fredsoe expression (4).

3.2 ADDITIONAL RESISTANCE TO FLOW INDUCED BY FLEXIBLE VEGETATION

Based on the bitter experiences of recent floods in Europe and the USA, many pressure groups have promoted the restoration of natural wetlands that would act as natural 'sponges' capable of absorbing excess water and thus reducing flooding risks. The wetland restoration projects favour the growth of reeds and other similar vegetation within a river basin. The presence of vegetation influences the flow conditions, and in particular the bed resistance, to a large degree. However, the influence of the rigid and flexible vegetation on flow conditions is not understood well enough.

Recently, a numerical model has been developed with the intention of deepening the understanding of the underlying processes (Kutija & Hong 1996). This model is a one-dimensional vertical model based on the equations of conservation of momentum in the horizontal direction. This numerical model is employed here as an experimental apparatus in the sense that this, fully deterministic (even if highly parametrised), model is used as a source of data that are the further processed by two apparently different methodologies in order to induce a more compact model of the additional bed resistance caused by vegetation.

3.2.1 Data

The Kutija-Hong model, used as a generator of data, was in effect used as a truthful representation of a physical reality, while providing the conveniences of fast calculation and an ability to produce results with any degree of scale refinement. In this way, the numerical model not only replaced physical scale modelling facilities within this exploratory environment, but also introduced several intrinsic advantages over scale models. It is well known that so-called *roughness scaling* is one of the principal difficulties in the development of physical models. Since the roughness is the primary phenomenon in question here, the issue of its physical correctness remained critical. The 'realism' of the complete numerical model was reasonably well proven against experimental data in the case of stiff (non-flexible) vegetation (Kutija & Hong 1996). As the first attempt towards the development of a model of additional roughness, only the effects of non-flexible reeds with high stiffness were simulated. Altogether, some 4,800 items of training data were generated. The training data consisted, in the first instance, of dimensional numbers formed from:

- water depth h_w, varied in [2.5 – 4.0]
- reed height h_r, varied in [0.25 – 2.25]
- reed diameter d, varied in [0.001 – 0.004]
- number of individual reed shoots per square meter m, varied in [50 – 350]
- a numerical parameter p related to the eddy-viscosity approximation and its further relation to the vegetated layer height, which varied in [0.4 – 1.0].

The target variable is Chezy's *roughness* coefficient C. This coefficient is stated in the derived units of square root of length over time. The awkwardness of the units of C suggests that it is chosen in such a way that the overall dimension of a more encompassing model will match. As such, the physical meaning (grounding) of this calibration coefficient C is questionable.

3.2.2 Results Based on Standard Genetic Programming

The following two sections are based on Babovic (1996) and Babovic & Keijzer (1999). The results of

Kutija-Hong simulations were presented as *C*: the Chezy number corresponding to the flow conditions with developed vegetation.

Dimensional values

In the first attempt, Babovic (1996) used standard symbolic regression to approximate the data in their original, dimensional form, resulting in the following:

```
( / (exp (-d hr ) (-(rlog (/ (rlog (/ m
hw )) (sqrt (sqrt (/ d d))))) (* (/ hr
(sqrt (sqrt (/ d (sqrt (sqrt (exp (-d hr
))))))))(* (* (rlog (/ (rlog (/ m hr ))
(exp (-d hr )))) (* (-(-d hr ) hr ) -
0.00410)) p))))
```

The shear complexity of the formulation almost immediately eliminates it from a knowledge induction framework.

In order to improve the interpretability, Babovic & Keijzer (1999) employed a more advanced version of GP with the best performing formula being:

$$C_{new} = \frac{55.93 h_w}{h_w + d + 3dm + 3h_r - p} - 11.39 \qquad (7)$$

There is some dispersion or the higher values of C_{new} but otherwise the equation exhibits good accuracy (see Table 2). The scatter plot for Equation (7) is depicted in Figure 1. However, it has to be emphasised that Equation (7) is not dimensionally correct. This shortcoming can be corrected by introducing an auxiliary constant with dimension of length and magnitude of 1.0 that should be multiplied with dimensionless *p* to correct dimensions. At the same time, both constants 55.93 and 11.39 should be assigned dimensions of the Chezy number [$m^{½}$ s^{-1}] to make this formula dimensionally correct. As indicated earlier, such behaviour is not surprising when applying standard instances of GP. Satisfactory goodness-of-fit may be obtained, but the semantics of the generated expressions cannot be warranted.

Dimensionless values

In the dimensionless case the results of Kutija-Hong simulations were presented as a dimensionless ratio η of an original Chezy number, that corresponding to an absence of vegetation, and a new Chezy number, that corresponding to developed vegetation. This ratio η can be conveniently incorporated in the Chezy formula for velocity under steady flow conditions:

$$u = \eta\, C\sqrt{Ri} \qquad (8)$$

For example, for η=0,the resistance to flow becomes infinitely large, thus stopping the water flow, which is physically unlikely situation. The smallest values of η experienced within Kutija-Hong numerical model were η = 0.1. For η = 1, the influence of vegetation on the roughness amounts to zero. Another set of model induction experiments has been performed, but in this case a collection of dimensionless numbers has been used. The dimensionless ratios introduced were defined as follows: $h_{rel} = \frac{h_w}{h_r}$, $w_d = \frac{h_w}{d}$, $r_d = \frac{h_r}{d}$, $h_{hwhrd} = \frac{h_w - h_r}{md}$.

In addition to these, parameter *p* and *m* were used without any changes. The best performing expression found is:

$$\eta = \left[2\left(\frac{h_w}{h_r}\right)^{0,75} + \sqrt{\frac{h_w - h_r}{md}} \right] 0{,}06 - 0{,}24 \qquad (9)$$

The performance statistics are presented in Table 2. The first interesting observation is slightly counter-intuitive: the accuracy of induced formulation in a dimensionless case is not as good as the accuracy of dimensional formulation. Even if the accuracy of η would be acceptable, it is the behaviour of Chezy's *C* (calculated as $C_{new} = \eta\, C_{org}$) that is undesirable in this case (see the upper two graphs in Figure 1). Thus, the acclaimed compression of the search space through the use of dimensionless values obviously incorporates several hidden risks that need to be handled with considerable care.

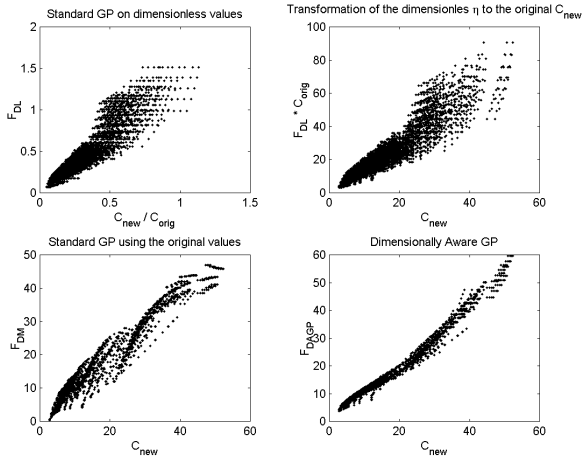

Figure 1 Scatter plots for expressions (7), (9) and (10) on the test set. The upper two graphs depict behaviour of an induced relationship in dimensionless case (denoted as F_{DL}). The graph in the upper left corner is a scatter plot for dimensionless η, whereas the graph in the upper right corner is a plot for the corresponding Chezy coefficient. The graph in the left right corner depicts performance in the dimensional case (F_{DM}). Finally the graph in the lower right corner depicts the results for the dimensionally aware GP (F_{DAGP})

3.2.3 Results based on dimensionally aware genetic programming

Again, a dimensionally aware genetic programming environment was set up to comprehend all measured data and *not* the corresponding dimensionless parameters based on the measurements. As in the case of concentration of suspended sediment near bed, the pre-processing of raw observations was *not* employed here. The purpose or conducting such an experiment was to test whether such a dimensionally aware GP setup is capable of creating a dimensionally correct and still accurate formulation. The evolutionary processes resulted in a number of expressions, of which only the most interesting one is presented here:

$$C_{new} = \frac{g\left[\frac{h_w - h_r}{dm}\, p\right]^{1/4}}{h_r^{1/2}} \qquad (10)$$

This formulation is statistically superior to other formulations found using the different techniques.

Table 2 Statistical summary for expressions (7), (9) and (10) on an independent test set. The large difference in RMS between equation (9) and equations (7) and (10) comes from the calculation of model performance with respect to the transformed target η.

	r	*R2*	*RMS*	*NRMS*
(7)	0.96	0.92	2.880	27.89%
(9)	0.94	0.89	0.076	37.47%
(10)	0.98	0.97	1.800	17.44%

At the same time the formula is dimensionally consistent, it uses some of the most relevant physical properties in the relevant context. For example, the dimensionless term $\frac{h_w - h_r}{dm}\, p$ describes a ratio between the effectively available cross-section $(h_w\text{-}h_r)\ p$ and a part of the cross-section that is blocked by the plants per unit width of the channel. The remaining group $\frac{g}{h_r^{1/2}}$ represents a ratio of gravity forces and flow resistance 'force' expressed through the reed height.

In this example, evolution produced a dimensionally consistent, meaning-rich formulation that is very accurate. It did so without employing assumptions (other than units of measurement); the process operated only on raw observations. Still, Equation (10) is not dimensionally correct; it does not produce the derived units for the Chezy coefficient. This may originate in at least two causes:

1. Incomplete data: in the present data set neither time nor elasticity components were provided (the authors supplied g=9.81 m/s^2). The next iteration in this direction must resolve this deficiency in one way or another.
2. The problem may simply be ill posed: the authors attempted to model Chezy's *C* despite their reservation about its grounding (however, without any reservations about its usefulness).

4 GP AS AN AID IN THE DISCOVERY OF SCIENTIFIC KNOWLEDGE

When using dimensional correctness as a as a second objective in searching for accurate equation, the user of this system will be confronted not only with a single best solution, but generally with a *front of non-dominated* solutions balancing goodness-of-fit with dimensional correctness. Taken further into account that genetic programming is a randomized algorithm and common practice dictates that multiple runs are needed to obtain good results, the question then remains *what to do with all these proposed formulae*? For the examples presented in this paper, the authors made the choices: in the sediment transportation problem several runs were performed, each leading to different dimensionally correct and incorrect equations. The solution (6) presented here was selected as it was the most accurate equation among dimensionally correct equations. Further analysis then revealed the underlying structure of the equation. In the problem involving roughness coefficient, none of the runs produced a dimensionally correct equation that had an accuracy comparable to the incorrect equations. It was then *judged* that the most accurate formulation that had second best dimension correctness should be further analyzed. It turned out that this equation was *internally consistent* although it did not produce a quantity in the desired dimensions of roughness (10). The fact that adequate dimensionless expressions could not be found furthermore gave an indication that data were missing from the problem.

It is our view that the process of generating hypothesis about the data and subsequently *judging* and *analyzing* a set of such hypothesis reveals the true strength of this approach. Scientific discovery is not, and perhaps should never be, a fully automated process where the machine generates solutions that are accepted at face value. Physical interpretation of the proposed equations is needed! It is our firm belief that dimensionally aware GP can be best used as a generator of novel formulations, balancing important properties such as goodness-of-fit, dimensional correctness and parsimony. Domain experts

should then be exposed to a completely new set of formulations, off the beaten track, yet within the domain of physical validity.

5 CONCLUSIONS

Traditionally, dimensionless numbers are used as the dominant vehicle in interpretation and modelling of experimental values. Such a choice is natural as this alternative conveniently avoids the issues related to dimensional analysis and its correctness. It is also believed that dimensional numbers collapse the search space and that resulting formulations are more compact. This paper demonstrated that it can be advantageous to use data together with its dimensions. The knowledge discovery software system uses this information to guide the search for an accurate and physically sound formulation.

The authors maintain that the approach presented in this paper is very useful for the purposes of model induction. The dimensionally aware approach is open-ended in that it does not strictly adhere to the dimensional analysis framework. The authors will go even further to claim that the dimensionally aware approach is much more useful than a strict use of dimensional analysis to create and use only dimensionless ratios. At the same time, the authors remind the reader that the object of the presented exercise is to find an *empirical equation* based on data. The present work cannot be characterised as a search for a universal law (though it might help). Being able to use units of measurements (either through the dimensionally aware or through strong adherence to dimensional analysis) provides an opportunity to truly *mine the knowledge from the data*, to learn more from data and other associated information. The ultimate objective is to build models that can be interpreted by the domain experts. Once a model is interpreted, it can be used with more than just statistical confidence. It is only in this way that one can take full advantage of knowledge discovery and advance our understanding of physical processes.

As the examples in the previous sections show, genetic programming can also contribute to creation of novel knowledge. The obvious corollary of the discussion above is that main intention should be to use genetic programming as an aid to scientists, rather than their replacement. Clearly, we are only beginning to develop effective ways of combining the strengths of human cognition with those of computational discovery systems. However, it is fairly easy to predict a more widespread use of genetic programming in the process of scientific discovery.

Acknowledgments

The authors gratefully acknowledge Dr Vedrana Kutija for the use of roughness data. This work was in part funded by the Danish Technical Research Council (STVF) under the Talent Project N° 9800463 entitled "Data to Knowledge — D2K". Their support is greatly appreciated. For more information on the project, visit http://www.d2k.dk

References

Babovic, V., 1996, *Emergence, Evolution, Intelligence; Hydroinformatics*, Balkema, Rotterdam

Babovic, V., and Keijzer, M., 1999, Computer supported knowledge discovery — A case study in flow resistance induced by vegetation, *Proceedings of the XXVI Congress of International Association for Hydraulic Research, Graz 1999*

Babovic, V., and Keijzer, M., 2000., *Genetic programming as a model induction engine*, Journal of Hydroinformatics, Vol. 2, No. 1, pp. 35 – 61

Clack, C., and Yu, T., 1997, Performance enhanced Genetic Programming, *Proceedings of the sixth conference on Evolutionary Programming*, Indianapolis

Ellis, B. D., 1965, *Basic Concepts of Measurement* Cambridge University Press

Keijzer, M., and Babovic, V., 1999, Dimensionally aware genetic programming, in Banzhaf, W. et. al. (eds.)., *Proceedings of GECCO-99*, Morgan Kaufmann.

Keijzer, M., and Babovic, V., 2000, Genetic Programming, Ensemble Methods and the Bias/Variance Tradeoff, in *Proceedings of EuroGP 2000*, Springer

Koza J R., 1992., *Genetic Programming: On the Programming of Computers by Natural Selection*, MIT Press, Cambridge, MA

Kutija, V., and Hong, H.T.M., 1996., A numerical model for addressing the additional resistance to flow introduced by flexible vegetation, *Journal of Hydraulics Research*, Vol.34, No. 1, pp. 99-114

Langley, P., 1998., The computer-aided discovery of scientific knowledge, in *Proceedings of the First International Conference on Discovery Science*, Fukuoka, Japan

Montana, D., 1995, Strongly typed genetic programming, *Evolutionary Computation 3 (1995),* no.2, pp. 199-230

Zyserman, J.A., and Fredsøe, J., (1994), Data analysis of bed concentration of suspended sediment, *Journal of Hydraulic Engineering*, ASCE, 120, No.9, pp.1021-1042