Computer-Aided Civil and Infrastructure Engineering **15** (2000) 383–389

SHORT CONTRIBUTION

Data Mining and Knowledge Discovery in Sediment Transport

Vladan Babovic

Danish Hydraulic Institute, Hørsholm, Denmark

Abstract: *The means for data collection have never been as advanced as they are today. Moreover, the numerical models we use today have never been so advanced. Feeding and calibrating models against collected measurements, however, represents only a one-way flow: from measurements to the model. The observations of the system can be analyzed further in the search for the information they encode. Such automated search for models accurately describing data constitutes a new direction that can be identified as that of* data mining. *It can be expected that in the years to come we shall concentrate our efforts more and more on the analysis of the data we acquire from natural or artificial sources and that we shall* mine for knowledge from the data so acquired.

Data mining and knowledge discovery aim at providing tools to facilitate the conversion of data into a number of forms, such as equations, that provide a better understanding of the process generating or producing these data. These new models combined with the already available understanding of the physical processes—the theory—result in an improved understanding and novel formulations of physical laws and improved predictive capability.

This article describes the data mining process in general, as well as an application of a data mining technique in the domain of sediment transport. Data related to the concentration of suspended sediment near a bed are analyzed by the means of genetic programming. Machine-induced relationships are compared against formulations proposed by human experts and are discussed in terms of accuracy and physical interpretability.

1 INTRODUCTION

The formation of modern science took place approximately in the period between the late fifteenth century and the late eighteenth century. The new foundations were based on the utilization of the concept of a *physical experiment* and the application of a *mathematical apparatus* in order to describe these experiments. The works of Brahe, Kepler, Newton, Leibniz, Euler, and Lagrange clearly personify such an approach. Prior to these developments, scientific work consisted primarily of collecting the observables or recording the *readings of the book of nature itself* only.

This novel scientific approach was principally characterized by two stages: a first one in which a set of observations of the physical system was collected and a second one in which an inductive assertion about the behaviour of the system—a hypothesis—was generated. The observations represent *specific knowledge*, whereas the hypothesis represents a *generalization* of these data that *implies* and *describes* the observations. One may argue that through this process of hypothesis generation, one fundamentally economizes thought, as more compact ways of describing observations are proposed.

Today, in the beginning of the twenty-first century, we are experiencing yet another change in the scientific process as outlined. This latest scientific approach is one in which state-of-the-art information technology is employed to assist the human analyst in the process of hypothesis generation. This computer-assisted analysis of large, multidimensional data sets is referred to as a process of *data mining*. Obviously, the means for data collection have never been as advanced as they are today. SCADA (supervisory control and data acquisition) systems form the principal data-collection components of many hydroinformatics systems. Nor have the numerical models we use today ever been so advanced. Feeding and calibrating models against collected measurements, however, only represent a one-way flow: from measurements to the model.

The observations of a physical system can be further analyzed in the search for the information that they encode.

© 2000 *Computer-Aided Civil and Infrastructure Engineering*. Published by Blackwell Publishers, 350 Main Street, Malden, MA 02148, USA, and 108 Cowley Road, Oxford OX4 1JF, UK.

Techniques such as artificial neural networks (ANNs) and evolutionary algorithms (EAs) have already entered the modeling arena and are currently understood as the principal means for merging various sources of information. This, in very few words, constitutes a new direction that can be identified as that of *data mining in hydroinformatics*. It can be expected that in the years to come we shall concentrate our efforts more and more on the analysis of the data we acquire from natural or artificial sources and that we shall *mine for knowledge from the data so acquired*.

This article describes some of the very first efforts under the D2K (Data to Knowledge) research project currently being conducted at the Danish Hydraulic Institute with support from the Danish Technical Research Council (STVF). The article first outlines elementary data mining principles, particularly when applied to the analysis of scientific data. In the second section, results obtained through analysis of the data related to the concentration of suspended sediment near a bed are analyzed by means of genetic programming (GP). Finally, induced formulations are discussed in terms of accuracy and physical interpretability.

2 DATA MINING

Knowledge discovery in databases (KDD) is concerned with extracting useful information from data stores. *Data mining* is the step (be it automated or human-assisted) in this larger process called the *KDD process*. The broad KDD process includes retrieving the data from a large data warehouse (or some other source); selecting the appropriate subset with which to work; deciding on the appropriate sampling strategy; selecting target data; dimensionality reduction; cleaning; data mining, model selection (or combination), evaluation, and interpretation; and finally, the consolidation and putting into practical use of the extracted “knowledge.”[9] The data mining step will then adapt the models to the preprocessed data or will extract patterns from the same. The role of the human expert is to provide domain knowledge, to interpret models suggested by the computer, and to devise further experiments that will provide even better data coverage. Clearly, there is an enormous amount of knowledge and understanding of physical processes that should not just be thrown away. Consequently, I strongly believe that the most appropriate way forward is to combine the best of the two approaches: theory-driven, understanding-rich, with a data-driven discovery process.

Thus two main forms of input are required in data mining: an *operator*, preferably a domain expert who is capable of giving advice about the nature of a problem, helping in the configuration of a system, checking the appropriateness of the representation of a problem and other such tasks, and a *database* of correctly classified cases that is used as an experiential foundation for the data mining process.

3 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 solely on observations. These observations, however, always represent (and in principle only) a *limited source of information*. The question emerges as to 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 the sense that it can account for the *entire* range of phenomena encountered within the physical system in question—and to describe even the data outside the range of previously encountered observations.[3]

The present efforts are characterized by the search for a model that is capable of *acquiring semantics from syntax*. Clearly, every model has its own syntax. Artificial neural networks have the syntax of a network of interconnected neurons, whereas genetic programming has the syntax of treelike networks of symbolic expressions in reverse Polish notation.[8] The question is whether such a 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. For example, one may decide to expand Taylor or Fourier series and decrease the error by adding terms in a series. However, in these cases, semantics almost certainly would not be caught.

4 EVOLUTIONARY ALGORITHMS

In this study, I use an approach based on a special kind of evolutionary algorithm—genetic programming—as a model-induction system. Evolutionary algorithms (EAs) are engines simulating grossly simplified processes occurring in nature and implemented in artificial media—such as a computer. The fundamental idea is that of emulating the Darwinian theory of evolution. According to Darwin, evolution is best depicted as the process of the adaptation of species to their environment as one of “natural selection.” Perceived in this way, all species inhabiting our planet are actually results of this process of adaptation.

Evolutionary algorithms effectively provide an alternative approach to problem solving—where solutions of the problem are evolved rather than the problems being solved directly. Today, the family of evolutionary algorithms is divided into four main streams: evolution strategies,[13] evolutionary programming,[10] genetic algorithms,[11] and genetic programming.[12]

Although different and intended for different purposes, all EAs share a common conceptual base (shown in

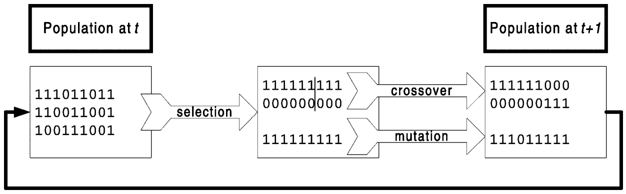


Fig. 1. Schematic illustration of an evolutionary algorithm. The population is initialized (usually randomly). From this population, the most fit entities are selected to be altered by genetic operators exemplified by crossover (corresponding to sexual reproduction) and mutation. The selection is performed based on certain fitness criteria in which the more "fit" are selected more often. Crossover simply combines two genotypes by exchanging substrings around randomly selected points. In the illustration, parental genotypes are indicated as either all 1's or all 0's for the sake of clarity. Mutation simply flips the randomly selected bit.

Figure 1). In principle, an initial population of individuals is created in a computer and allowed to evolve using the principles of inheritance (so that the offspring resembles the parents), variability (the process of offspring creation is not perfect; some mutations occur), and selection (more fit individuals are allowed to reproduce more often and less fit less often so that their "genealogic" trees disappear in time).

One of the main advantages of EAs is their domain independence. EAs can evolve almost anything, given an appropriate representation of evolving structures. Similar to processes observed in nature, one should distinguish between an evolving entity's genotype and its phenotype. The *genotype* is basically a code to be executed (such as a code in a DNA strand), whereas the *phenotype* represents a result of the execution of this code (such as any living being). Although the information exchange between evolving entities (parents) occurs at the level of genotypes, it is the phenotypes in which one is really interested.

The phenotype is actually an interpretation of a genotype in a problem domain. This interpretation can take the form of any feasible mapping. For example, for optimization and constraint-satisfaction purposes, genotypes typically are interpreted as independent variables of a function to be optimized. Along these lines, one can employ mapping in which genotypes are interpreted as roughness coefficients in a free-surface pipe-flow model with the genetic algorithms (GAs) directed toward the minimization of the discrepancies between the model output, the measured water level, and the discharge values. The resulting GA represents an automatic calibration model of hydrodynamic systems.[7] Several other applications of GAs, which make use of various kinds of genotype-phenotype mappings and with a specific emphasis on water resources, are also described in literature.[2]

5 GENETIC PROGRAMMING

In genetic programming (GP), the evolutionary force is directed toward the creation of models that take a symbolic form. In this evolutionary paradigm, evolving entities are presented with a collection of data, and the evolutionary process is expected to result in a closed-form symbolic expression describing the data. GP iteratively applies variation and selection on a population of evolving tree structures representing symbolic expressions in reverse Polish notation (RPN). Standard variation operators in GP are subtree mutation (replace a randomly chosen subtree with a randomly generated subtree) and subtree crossover (replace a randomly chosen subtree from a formula with a randomly chosen subtree from another formula).[5] The types of functions used in this tree structure are user-defined. This means that they can be algebraic operators, such as sin, log, $+$, $-$, etc., but they also can take the form of if-then-else rules, making use of logical operators such as OR, AND, etc. A number of applications of GP has been reported, e.g., studies in which salt intrusion data were analyzed,[8] studies related to experimental data on the concentration of suspended sediment near a bed,[1] and studies related to rainfall runoff modeling.[4] In all of these studies, GP-induced relationships provided more accurate descriptions of data than those obtained using more conventional methodologies. An extensive survey of the applications of GP in water resources is provided in Babovic and Abbott.[6]

6 AN APPLICATION TO SEDIMENT-TRANSPORT MODELING

6.1 Data

This study makes use of experimental flume data utilized by Zyserman and Fredsøe.[14] The experimental data consisted of a 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.[14] The hydraulic conditions were represented by the *Shields parameter* θ, defined as

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

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

$$\frac{v}{u_{f'}} = 6 + 2.5 \ln\left(\frac{D'}{k_N}\right) \qquad (2)$$

v = mean flow velocity

k_N = bed roughness $= 2.5d$

The study makes use of the so-called Rouse number z, defined as

$$z = \frac{w_s}{\kappa u_f} \qquad (3)$$

where

w_s = settling velocity of suspended sediment

κ = von Karman's constant (≈ 0.4)

It is interesting to observe that not all "directly measurable" quantities correlate too well with the concentration of the sediment c_b as well as the derived dimensionless quantities θ and θ' (see Table 1). This phenomenon also can be inspected graphically as in Figure 2.

6.2 Human analyst–induced relationship as a benchmark

Bearing these strong correlations in mind, it is a little surprising that after dimensional analysis, Zyserman and Fredsøe formulated the following expression[14]:

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

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

Table 1
Correlation coefficients for directly observable and derived, dimensionless quantities

Directly observable		*Derived quantities*	
Correlation pairs	*Correlation coefficient*	*Correlation pairs*	*Correlation coefficient*
c_b–u'_f	0.784	c_b–θ'	0.894
c_b–u_f	0.628	c_b–θ	0.711
c_b–ws	0.430		
c_b–nu	0.152		
c_b–d_{50}	−0.232		

6.3 Formulations induced by the means of genetic programming

By way of comparison, a genetic programming environment was set up in such a way as to comprehend all measured data and their corresponding parameters based on the measurements, namely, θ, θ', and z. These dimensionless quantities were chosen in order to ensure the dimensional consistency. The evolutionary process resulted in a number of expressions, of which only the best-performing and the most interesting ones are presented. The best-performing expression is presented as an RPN tree in Figure 3. This expression can then be written in an ordinary (infix) notation as

$$c_b = 0.07\left[1.93\theta' + 0.38 \ln \sqrt{\theta} + 0.18 - \frac{\theta'}{0.35(\theta - 1.56^{0.11})}\right] - 0.0487 \qquad (5)$$

The degree of accuracy of the induced expression is quite satisfactory. A statistical measure of conformity, such as the adjusted coefficient of determination, gives a value of 0.651. This provides an improvement of 6.5 percent over the value of 0.611 based on the Zyserman-Fredsøe relationship (Equation 4). The total error over the data set is reduced by 14.55 percent. All other statistical measures such as average deviation, coefficient of efficiency, robustness, and 95 percent confidence interval show an improvement over the benchmark expression.

The summary of the statistical information comparing the performances of the Zyserman-Fredsøe formulation (4) and the GP-induced formulation (5) is presented in Figure 4 and is seen to resemble a normal distribution. This behavior, when compounded with the analysis of robustness, indicates that the induced formula can replace and generalize the training set. Further verification tests, however, are necessary before this GP expression can be used as a substitute for the model used to generate it.

The induced relationship can be further analyzed for the sensitivity of induced relationships to variations in the input parameters. The results of this analysis show that to a great extent the accuracy of Equation (5) depends on Shield's parameter related to skin friction (96.16 percent) and, to a considerably lesser extent, on Shield's parameter related to the average velocity (36.62 percent). Such additional information can be used to select the formulations that have the highest sensitivity rates to the most reliable input parameters.

It may be argued that the improvement in accuracy created by Equation (5) is due to the increased complexity of the expression and primarily due to the fact that this formula uses two independent variables θ and θ', whereas the Zyserman-Fredsøe expression uses θ' only. In order to test

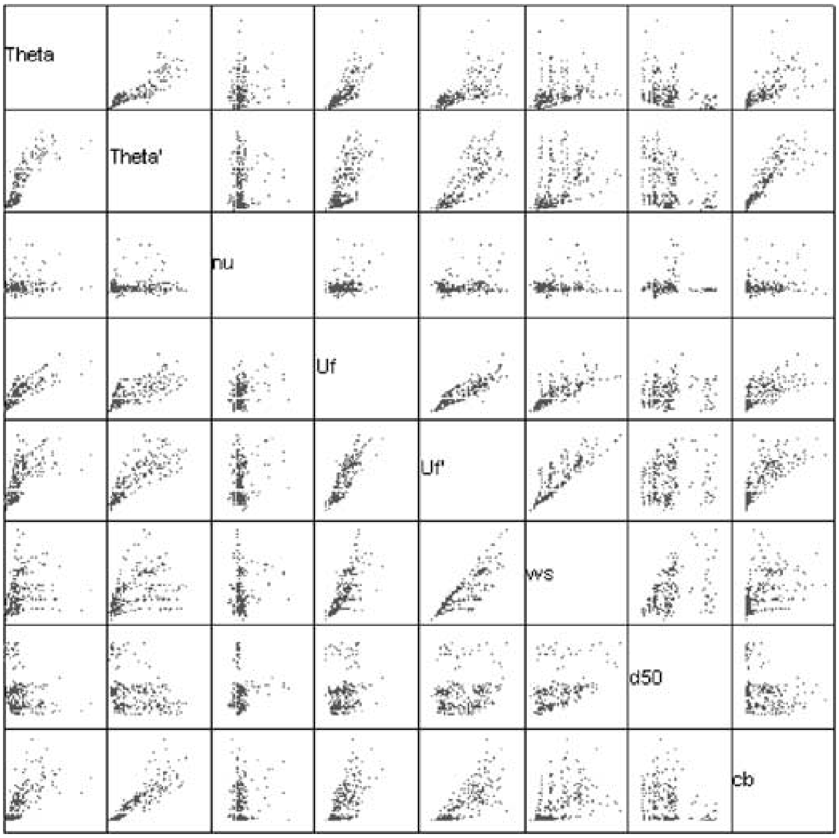
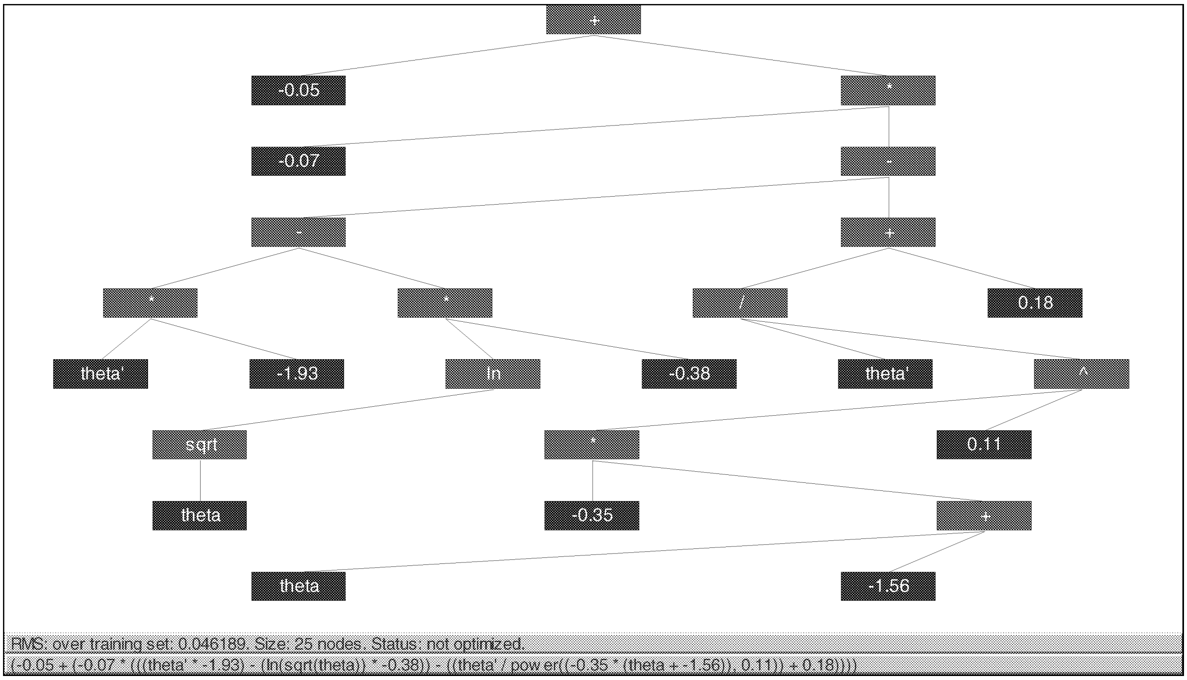
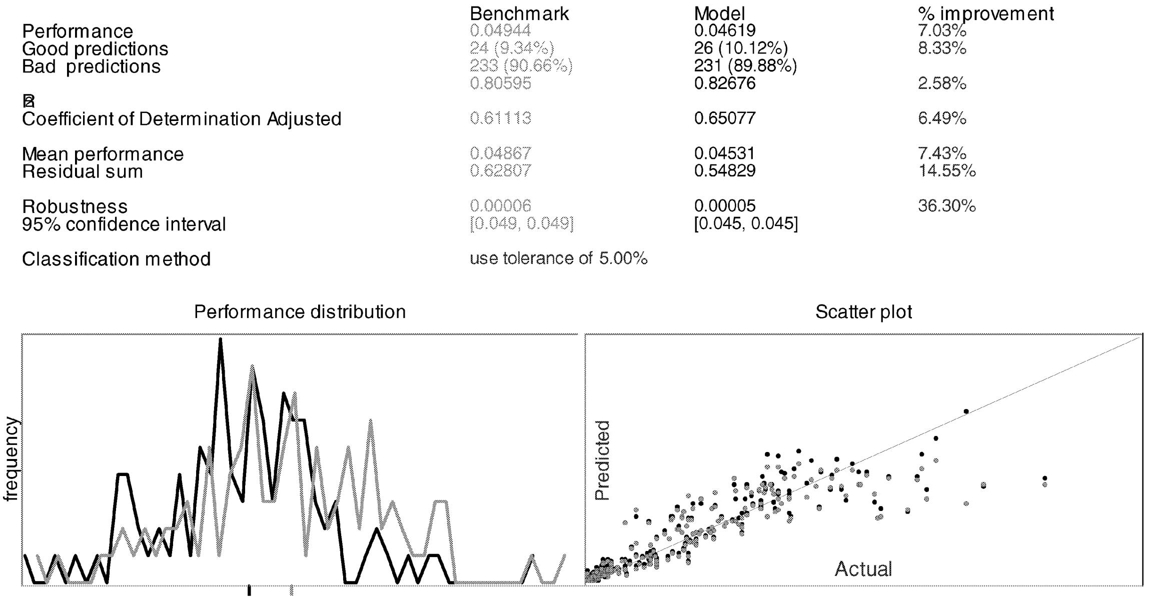


Fig. 2. Scatter plots of all variable pairs.



```
RMS: over training set: 0.046189. Size: 25 nodes. Status: not optimized.
(-0.05 + (-0.07 * (((theta' * -1.93) - (ln(sqrt(theta)) * -0.38)) - ((theta' / pow er((-0.35 * (theta + -1.56)), 0.11)) + 0.18))))
```

Fig. 3. The best-performing expression induced by the GP system—Expression (5).



	Benchmark	Model	% improvement
Performance	0.04944	0.04619	7.03%
Good predictions	24 (9.34%)	26 (10.12%)	8.33%
Bad predictions	233 (90.66%)	231 (89.88%)	
R^2	0.80595	0.82676	2.58%
Coefficient of Determination Adjusted	0.61113	0.65077	6.49%
Mean performance	0.04867	0.04531	7.43%
Residual sum	0.62807	0.54829	14.55%
Robustness	0.00006	0.00005	36.30%
95% confidence interval	[0.049, 0.049]	[0.045, 0.045]	
Classification method	use tolerance of 5.00%		

Fig. 4. Summary of statistical measures of performances for both the benchmark expression (4) and the GP-induced expression (5).

whether it is possible to evolve a formulation of comparable accuracy using θ', a new GP experiment was set. In this new environment, only θ' was allowed to be chosen by the system as an independent variable. The result of this induction process is presented in the sequel.

$$c_b = -0.31^{e^{\frac{1}{0.24\left[(\theta'-0.70)([\ln(\theta'/\theta')]+e^{\theta'})^{0.30\theta'}+0.38\right]}}} \tag{6}$$

A summary of statistical measures of accuracy and an intercomparison of performances between the Zyserman-Fredsøe formulation and Equation (6) is provided in Table 2. Albeit minor, this formulation provides satisfactory results, since there is an improvement in performance in all statistical measures of accuracy.

The open question, however, remains: Which formula to choose? However, the primary purpose of this article is not to propose an improved formulation for the concentration of suspended sediment near bed c_b but rather to demonstrate the power of some of the modern, intelligent data-analysis techniques. The physical interpretation of Equations (5) and (6) requires a wealth of knowledge in sediment transport, and I feel ill equipped to perform this task in a fully qualified way. Thus the choice of the best formula is not made here. Such a choice is still to be made by experts in this field, by the people who can competently judge the quality of the data sets used, provide an interpretation (semantics) of the induced relationships, and in the end choose the one that makes the most sense. It is only in this way that one can take full advantage of data mining and advance our understanding of physical processes.

Table 2
Comparison of some statistical measures of the accuracy for Zyserman-Fredsøe and Equation (6)

	R^2	*Coefficient of determination*	*Residual Sum*
Zyserman-Fredsøe	0.611	0.806	0.628
Equation (6)	0.612	0.808	0.609

The argument forwarded here is that mining of scientific data fundamentally alters the scientific process and is characterized by two principal qualities: (1) this is the first time in the history of computing that scientists can ask the computer to do something without explicitly instructing it how to do it, and (2) it is within such a framework that the human analyst takes the central position in a model-induction process as he or she is now able to concentrate on the analysis of the essence of the processes and their representation rather than the techniques for addressing and solving particular numerical matters.

7 CONCLUSIONS

This article addresses the utilization of GP in the process of scientific discovery. Throughout science, experimental data form a basis for the description of physical phenomena. The presently described framework is part of a research effort aimed at providing a new and automated hypotheses-building process on the basis of data alone. The ultimate objective is to build a knowledge-discovery environment in which models that can be interpreted by the domain experts are generated by a computer. Once a model is interpreted, it can be used with confidence. It is only in this way that one can take full advantage of automated knowledge discovery and advance our understanding of physical processes.

ACKNOWLEDGMENTS

I wish to acknowledge two anonymous reviewers for their constructive criticism of a previous version of this article. This work was funded in part by the Danish Technical Research Council (STVF) under Talent Project No. 9800463 entitled “Data to Knowledge—D2K.” Their support is greatly appreciated. For more information on the project, visit http://projects.dhi.dk/d2k.

REFERENCES

1. Babovic, V., Genetic model induction based on experimental data, in *Proceedings of the XXVIth Congress of the International Association for Hydraulic Research, Hydra 2000*, London, 1995.
2. Babovic, V., *Emergence, Evolution, Intelligence: Hydroinformatics*, Balkema, Rotterdam, 1996.
3. Babovic, V., Can water resources benefit from artificial intelligence? in *Proceedings of the Internationales Wasserbau Symposium—Computational Fluid Dynamics; Bulte Bilde in der Praxis*, ed. by J. Köngeter, Aachen, Germany, 1996.
4. Babovic, V., On the modelling and forecasting of nonlinear systems, in *Operational Water Management*, ed. by J. C. Refsgård & E. A. Karalis, Balkema, Rotterdam, 1997.
5. Babovic, V. & Abbott, M. B., The evolution of equations from hydraulic data: I. Theory, *Journal of Hydraulic Research*, **35** (3) (1997), 1–14.
6. Babovic, V. & Abbott, M. B., The evolution of equations from hydraulic data: II. Applications, *Journal of Hydraulic Research*, **35** (3) (1997), 15–34.
7. Babovic, V., Larsen, L. C. & Wu, Z, Calibrating hydrodynamic models by means of simulated evolution, in *Proceedings of the 1st International Conference on Hydroinformatics*, ed. by A. Verwey et al., Balkema, Rotterdam, 1994, pp. 193–200.
8. Babovic, V. & Minns, A. W., Use of computational adaptive methodologies in hydroinformatics, in *Proceedings of the 1st International Conference on Hydroinformatics*, ed. by A. Verwey et al., Balkema, Rotterdam, 1994, pp. 201–10.
9. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., From data mining to knowledge discovery: An overview, in *Advances in Knowledge Discovery and Data Mining*, ed. by U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy, MIT Press, Cambridge, MA, 1996, pp. 1–36.
10. Fogel, L. J., Owens, A. J. & Walsh, M. J., *Artificial Intelligence Through Simulated Evolution*, John Wiley & Sons, Needham Heights, 1966.
11. Holland, J. H., *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, 1975.
12. Koza, J., *Genetic Programming: On the Programming of Computers by Natural Selection*, MIT Press, Cambridge, MA, 1992.
13. Schwefel, H. P., *Numerical Optimisation of Computer Models*, Wiley, Chichester, England, 1981.
14. Zyserman, J. A. & Fredsøe, J., Data analysis of bed concentration of suspended sediment, *Journal of Hydraulic Engineering*, ASCE, **120** (9) (1994), 1021–42.