

Determining Equations for Vegetation Induced Resistance using Genetic Programming

Maarten Keijzer
WL | Delft Hydraulics
The Netherlands
mkeijzer@xs4all.nl

Martin Baptist
Delft University of Technology
The Netherlands
m.j.baptist@citg.tudelft.nl

Vladan Babovic
WL | Delft Hydraulics
The Netherlands
v.babovic@tectrasys.com

Javier Rodriguez Uthurburu
Uruguay
jruthurburu@hotmail.com

ABSTRACT

Inducing equations based on theory and data is a time-honoured technique in science. This is usually done manually, based on theoretical understanding and previously established equations. In this work, for a particular problem in hydraulics, human induction of equations is compared with the use of genetic programming. It will be shown that even with the use of synthetic data for training, genetic programming was capable of identifying a relationship that was more concise and more accurate than the relationship uncovered by scientists. As such this is a human-competitive result. Furthermore it will be shown that the genetic programming induced expression could be embedded in a line of theoretical work, filling in a few gaps in an already established line of reasoning. The resulting equation is the most accurate and elegant formulation of vegetation induced resistance to date.

1. INTRODUCTION

Proper modelling of flow in wetlands and vegetated floodplains is of great practical importance. Many research initiatives have been undertaken in order to improve on the description of the relationship between flow resistance and the presence and spatial distribution of vegetation. Both analytical and experimental studies of vegetation-related resistance to flow and the equivalent resistance coefficients have shown that the resistance coefficients are water depth dependent. Consequently, the traditional approach of using a single resistance coefficient fails to correctly describe the physics of the phenomenon. One way of improving upon this description is to update the equivalent resistance coefficient based on the computed water depth. In order to do so, a

relation between vegetation characteristics, bed resistance, water depth and equivalent resistance coefficients is needed. Such a relationship can be used in large scale simulations to model and predict the flow characteristics in wetlands and floodplains.

What is searched for in this work is an equation that models these resistance coefficients based purely on measurable parameters.

Two main approaches for creating such an equation are contrasted here. The first approach is the time-honoured method where a scientist uses whatever knowledge is available on the physics of the phenomenon and assembles an equation based on detailed understanding of the phenomena involved in the process. This understanding takes the form of many small models of sub-phenomena that are assembled to create an overall equation. The second approach employs a genetic programming variant to first induce a set of hypothetical relationships that are subsequently selected and *improved* by a scientist. The paper aims to show that the latter process can produce expressions that are in no way inferior to those produced through the former, and in this case they are significantly better, due to both an improved fit and an economy in the amount of detail that is modelled. In a very direct sense these results are an example of human competitive performance of a genetic programming technique, but we will also show that this is not the whole story: the combination of inductive hypothesis generation by genetic programming and subsequent analysis and modifications by a scientist can significantly improve both the understanding and the modelling of the phenomenon that is researched.

This paper shows the end results of the analyses that have been performed on this particular problem. Considerations of space and of suitability of presenting hydraulic material to this audience prevent a thorough description of the derivations. Such a full analysis is currently being prepared for a publication in the field of hydraulics. What is presented here are the theoretical equations describing roughness coefficients, and a comparison of the results in terms of interpretability, simplicity and suitability for further analysis.

Section 2 presents some background information about the problem being studied here. Section 3 describes the ori-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GECCO'05, June 25–29, 2005, Washington, DC, USA.
Copyright 2005 ACM 1-59593-010-8/05/0006 ...\$5.00.

gins of the data used for training the genetic programming induced equations. Section 4.2.1 and Section 4.2.2 describe the equations that were derived by a scientist to model the phenomenon under study. Section 4.3 describes the experiments that give rise to the equation induced by genetic programming. Section 4.4 describes the analysis and further theoretical improvements made on the genetic programming induced equation. Section 5 compares all four expressions on the synthetic data used for training, but also on a set of laboratory flume experiments that were collected from 10 independent studies. Discussion and conclusions sections finalise this paper.

2. BACKGROUND

The effect of vegetation on flow is generally expressed as an effect on the hydraulic roughness. In early measurements (18th century) on flow velocities in channels it was found that the mean velocity, u (m/s), was a function of the water level slope, i ($-$), and the hydraulic radius, R (m). In 1776 Antoine Chézy published a simple equation that includes a factor C , the Chézy value, which was at first thought to be a constant (Vernon-Harcourt, [24]). The well-known Chézy formula is:

$$u = C\sqrt{Ri} \quad (1)$$

In this equation C ($m^{1/2}s^{-1}$) is a parameter that expresses the hydraulic roughness of the bed and banks of a channel. Further investigations, by Nikuradse [20], revealed that the roughness of the bed affects the roughness length, z_0 (m), in the logarithmic velocity profile derived by, among others, Prandtl. Nikuradse showed that for hydraulically rough walls, the roughness length of the logarithmic velocity profile can be expressed as $k_N/30$, where k_N is the Nikuradse equivalent roughness (m) [20]. Using this for uniform flow conditions yields the White-Colebrook formula for the Chézy value:

$$C = 18 \log \frac{12R}{k_N} \quad (2)$$

With an increasing roughness height the value for C decreases. Various alternative expressions for flow resistance exist, for example those of Strickler or Manning. Essentially, using the White-Colebrook formula, vegetation is treated as large bed structures with a logarithmic flow profile above them. In reality, however, there is flow over and through submerged vegetation, and the vertical flow profile deviates from the logarithmic one. This has been established by many authors in the past decades, but even recent researchers attempt to fit a logarithmic profile and conclude that this does not work [21]. A typical velocity profile for submerged vegetation is shown in Figure 2. The White-Colebrook formula fails here and another type of resistance formula should be sought for.

A considerable amount of research has been carried out on the effects of vegetation on the hydraulic resistance, extending the basic ideas of Nikuradse [20]. In a study by Dawson & Charlton [6], a literature search has been carried out on vegetation roughness, resulting in over 360 publications. Since then, many more publications have followed. However, no acceptable formulation for roughness induced by submerged vegetation has been found as of yet. This was the inspiration for this work.

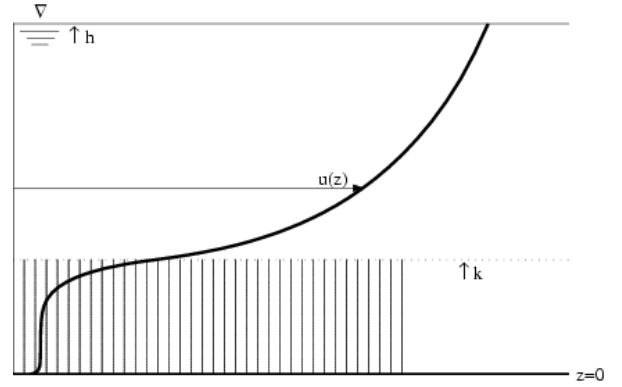


Figure 1: Typical vertical profile of horizontal flow velocity for submerged vegetation, h is water depth, k is vegetation height.

The genetic programming system used here is a variant of a basic genetic programming system, where the units of measurement are used as a guide to the search [10, 11]. This results in expressions that are not only able to fit the data, but also carry the semantic information about the units throughout the calculations. By focussing on expressions that are 'more-or-less' dimensionally correct, it is aimed to find interpretable expressions while not hindering search efficiency [11]. The system used here has been extensively used in the field of hydraulics and numerous results have been obtained using the system [1, 11, 2, 15, 7, 12, 18, 3].

3. DATA AVAILABILITY

Testing expressions for their ability to model a phenomenon such as resistance induced by vegetation needs experimental data. Particularly when using data driven methods, such data is needed for steering the error minimisation process. For manually induced equations, such data is needed to ultimately test the proposed equation for its capacity to model the phenomenon under study.

Even though the variables for this relationship should be measurable in wetlands and vegetated floodplains, this does not hold for resistance coefficients measured at a fine scale, at different water depths and with different types of vegetation. Obtaining data at such a fine scale in realistic circumstances is prohibitive in terms of effort and cost. Therefore, it is chosen to use a finely scaled microscopic dynamical model of the turbulence in the flow to generate data. Such a dynamical model employs all available knowledge about characteristics of plants, turbulence induced by the plants, and resistance caused by the drag forces on the plants. Determining the resistance coefficients from such a microscopic model is trivial, and can be used as a noise free approximation of the phenomenon under study. However, a compact expression describing the phenomenon is not readily available.

It can be argued that using such generated data defeats the purpose of finding an equation. If a dynamic model exists, why not simply use that one instead of going the laborious route of defining an equation. The purpose of finding a *macroscopic* equation lies in the type of modelling that it enables. Vegetation resistance is a typical 3-dimensional problem due to the water depth dependency. A full dynamical

Input	Dimension	Description
D	L	Diameter of the stems
m	L^{-2}	Number of stems per square meter
k	L	Plant height
C_D	–	Drag coefficient of a single stem
C_b	$L^{0.5}/T$	Bed Chézy resistance coefficient
h	L	Water depth

Table 1: Inputs to the 1DV model

model thus operates on a 3D grid, which is computationally expensive. An analytical solution to the problem of resistance induced by vegetation, which includes water depth dependency, makes 2-dimensional, depth-averaged modelling possible, allowing for faster model computations and the possibility to apply the model to larger areas. Furthermore, for the purpose of understanding the phenomenon, the 1DV model is unhelpful. Even though it produces values for resistance coefficients, it is not clear what the major influences are that describe the difference in resistance with different water depths. This is the critical part of interest in this work.

To ultimately test the models created here, a dataset of 177 laboratory flume experiments was collected from 10 independent studies. This data is not used for training, but kept aside to validate the equations induced.

3.1 The 1DV Turbulence Model

To obtain a detailed account of resistance of flow through and above vegetation, detailed numerical simulations based on a 1DV turbulence model [23] were performed. This model has sound theoretical support and has been thoroughly validated on experimental data collected by [17, 19, 16].

The 1DV model is a simplification of the full 3D Navier-Stokes equations in order to account for horizontal flow conditions only. It assumes that the flow is uniform in the horizontal directions, and calculates the orthogonal horizontal velocities $u(z)$ and $v(z)$ as a function of the vertical coordinate z . In order to include the effects of plants into turbulence closure, the following modifications have been included: (i) the decrease of the available cross-section for the exchange of momentum, turbulence kinetic energy and turbulent dissipation, (ii) the drag force exerted by the plants in the horizontal direction, (iii) an additional turbulence production term due to vegetation, and (iv) an additional turbulence dissipation term due to vegetation.

In effect, the 1DV model is a model of turbulence in flow, with the turbulence being influenced by the plants. It is a state-of-the-art model for modelling turbulence induced by vegetation. For a more detailed description the reader is referred to [23].

3.2 Data generated by the 1DV Turbulence Model

The data that were used to initialize the 1DV model is described in Table 1. The water depth dependent resistance coefficient that was modelled was the Chézy coefficient C_r . The coefficient C_r is stated in units $L^{0.5}/T$. In situations with bed resistance only, extensive research has gone into determining the value of C_r . All independent quantities are either measurable in the field or, in the case of C_D (the drag coefficient for a single stem), are well documented for

most vegetation types. From the 1DV model, 990 results for submerged vegetation were produced, and subsequently the resistance coefficient C_r was read from the model. The inputs covered a wide range of situations, including vegetation types (combinations of the input variables) that are not observed in nature.

4. BUILDING THEORETICAL EQUATIONS

Several equations have been induced. The first equation is analytically derived for unsubmerged vegetation only. The subsequent two equations are derived through two different methods for treatment of submerged vegetation, namely the "method of effective water depth" and a method utilising an analytical formula derived from the momentum balance. The third method utilised a genetic programming variant: dimensionally aware GP [10, 11], that evolves an equation based on the data measured from the 1DV model.

4.1 Case of Unsubmerged Vegetation

Unsubmerged flow conditions can be successfully treated analytically and the resistance coefficient can be derived as:

$$C_r = \sqrt{\frac{1}{\frac{1}{C_b^2} + \frac{C_D m D h}{2g}}} \quad (3)$$

Some special cases exist. For this work it is of interest that the bed resistance is negligible (in principle these two conditions are compatible only for tall and dense enough vegetation or very small bed roughness), the equivalent Chézy resistance coefficient can be approximated to:

$$C_r = \sqrt{\frac{2g}{C_{am} D h}} \quad (4)$$

4.2 Case of Submerged Vegetation

Empirical research points out that there four zones in the vertical velocity profile can be observed. This can be seen in Figure 2 where four different curves can be distinguished: near the bed, from this toward a little below the top of the plants, from there to top of the plants and finally a zone above the vegetation.

4.2.1 Method of Effective Water Depth

In the method of effective water depth, only the two most important zones are modelled:

- In the first zone which corresponds to the zone inside the vegetation sufficiently away from the bed and from the top of the vegetation, the velocity is constant.
- The second zone corresponds to the zone above the vegetation, where a logarithmic profile is observed

By summing up the discharges per unit width of each zone, a general analytical expression for the Chézy resistance coefficient can be derived as:

$$C_r = \frac{k\bar{u}_{veg} + (h - k)\bar{u}_u}{h\sqrt{hi}} \quad (5)$$

Where:

\bar{u}_{veg} is the mean velocity inside the vegetation layer, \bar{u}_u is the mean velocity above the vegetation layer, and

i is the energy gradient

The main effort then lies in finding expressions for these velocities.

The method of effective water depth [8, 5, 13, 22, 4], is based on Equation 5, where the velocities are approximated using the case of unsubmerged vegetation for the velocities inside the vegetation, and the resistance induced by the vegetation C_{vegk} for the zone above the vegetation. It results in the expression:

$$C_r = \frac{k \sqrt{\frac{1}{C_b^2 + \frac{1}{2g} C_D m D k}} + (h - k) C_{vegk} \sqrt{(h - k)}}{h \sqrt{h}} \quad (6)$$

The equation makes use of several intermediate, theoretical, expressions to obtain C_{vegk} , the roughness coefficient for the top of the vegetation:

$$C_{vegk} = 18 \log \left(1 + \frac{12(h - k)}{30z_0} \right) \quad (7)$$

The definition for z_0 can be found below.

4.2.2 Analytical Method Based on Momentum Balance

The analytical method attempts to model the velocity inside the vegetation by analytically solving the momentum equation for flow through and over the vegetation, represented as rigid cylinders:

$$\frac{\partial \tau_{xz}}{\partial z} - \rho_0 g \frac{\partial h}{\partial x} - \frac{1}{2} \rho_0 C_D m D u^2(z) = 0 \quad (8)$$

Solving this partial differential equation for the velocity profile inside the vegetation layer using boundary conditions, at the bed and at the top of the vegetation, furthermore assuming a logarithmic velocity profile above the vegetation which extends down in the vegetation layer, connecting with the profile underneath, the full expression for C_r becomes:

$$C_r = \frac{1}{h^{3/2}} \left\{ \begin{array}{l} L \left\{ \begin{array}{l} 2 \left(u_v(k) - \sqrt{a_v + u_{v0}^2} \right) \\ + u_{v0} \ln \left(\frac{(u_v(k) - u_{v0}) \left(\sqrt{a_v + u_{v0}^2 + u_{v0}} \right)}{(u_v(k) + u_{v0}) \left(\sqrt{a_v + u_{v0}^2 - u_{v0}} \right)} \right) \end{array} \right\} \\ + \frac{\sqrt{g(h-k)}}{\kappa} \left\{ \begin{array}{l} (h-d) \left(\ln \frac{h-d}{z_0} \right) \\ -(k-d) \left(\ln \frac{k-d}{z_0} \right) \\ -(h-k) \end{array} \right\} \end{array} \right\} \quad (9)$$

in which (for both Equation 6 and Equation 9):

$$z_0 = (k - d) \exp \left(-\kappa \sqrt{\frac{2L}{c_p \ell} \left(1 + \frac{L}{h - k} \right)} \right) \quad (10)$$

$$d = k - L \left(1 - \exp \left(-\frac{k}{L} \right) \right) \quad (11)$$

$$L = \sqrt{\frac{c_p \ell}{C_D m D}} \quad (12)$$

$$a_v = \frac{2Lg(h-k)}{c_p \ell \exp \left(\frac{k}{L} \right)} \quad (13)$$

$$\ell = \left(\frac{1 - L_p}{m} \right)^{\frac{1}{2}} \quad (14)$$

$$L_p = \frac{\pi}{4} D^2 m \quad (15)$$

$$u_{v0} = \sqrt{\frac{2g}{C_D m D}} \quad (16)$$

$$u_v(k) = \sqrt{u_{v0}^2 + \frac{2Lg(h-k)}{c_p \ell}} \quad (17)$$

It would be difficult in this space to explain the entire derivation of this set of equations, but intuitively what is modelled here is the effect of drag force and turbulence induced by vegetation on the momentum balance of flow. Critical parts of the derivation are the roughness length (z_0), which depends on the length scale L , the mixing length ℓ , and von Karman's constant $\kappa = 0.41$.

A special role is played by the turbulence intensity c_p and its effect on the length scale L . There is no agreement in the literature about analytical expressions for the turbulence intensity. Several attempts were made to model this intensity, including the use of genetic programming, but no satisfactory expression that held up in different circumstances was found. In this study, the turbulence intensity was taken as:

$$c_p = \frac{0.015 \sqrt{hk}}{\ell} \quad (18)$$

Although variants $c_p = 0.05 \frac{(h-k)}{\ell}$, $c_p = \frac{0.0227k^{0.70}}{\ell}$ and the genetic programming induced $c_p = \frac{1}{50} \frac{(h-k)}{\ell}$ were also considered. Results reported here use Equation 18.

4.3 Induced Expressions

A number of genetic programming runs that implement dimensional correctness as an objective next to the goodness of fit [11], have been performed. The system utilises interval arithmetic [9], but no linear scaling. To focus on both error, shape and dimensional correctness the system uses three objectives: Root Mean Squared Error (RMSE), Correlation squared (CoD) and dimension error that measures the dimensional consistency of the formulae [11]. For purposes of dimensional consistency and to focus the system on the geometrical properties of the problem, the roughness coefficients C_b and the dependent variable C_r were replaced with dimensionless values by scaling them with $1/\sqrt{g}$.

The system was run many times with different parameters using various subsets of the 990 cases. Subsequently the front of non-dominated solutions for all the runs was examined to find a suitable formulae. As the interest here was to produce one or more adequate formulations, no effort was made to examine the runtime behaviour and efficiency of the system.

After examining the resulting front of non-dominated solutions, one formula was an obviously good candidate as it combined low RMSE with high CoD without dimension error:

$$\frac{C_r}{\sqrt{g}} = \sqrt{\frac{2}{C_D m D k}} + \ln \left\{ \left(\frac{h}{k} \right)^2 \right\} \quad (19)$$

This formula can be rearranged to:

$$C_r = \sqrt{\frac{2g}{C_{Dm}Dk}} + 2\sqrt{g} \ln\left(\frac{h}{k}\right) \quad (20)$$

It is important to note that the first additive term in Equation 20 is equivalent to the simplified formula for unsubmerged vegetation (Equation 4) when $h = k$, as the logarithmic term would be zero. It appears that the formula induced by genetic programming has modelled a physically significant relationship using the data.

4.4 Subsequent Analysis of the GP-induced formula

A correspondence between the first term of Equation 20 and the simplified formula for unsubmerged vegetation (Equation 4) for the situation where $h = k$ can be identified. Analysis of the residuals reveals that for cases where $h \gg k$, as expected, the genetic programming induced Equation 20 produced large residuals (see Figure 2c). Similarly to the expressions based on the observed profile, Equation 5, the genetic programming induced equation consists of two additive parts: one describing the resistance inside the vegetation and one the resistance above the vegetation. However, it is desirable to have an expression compatible with the general formulation for unsubmerged vegetation, Equation 3. Substituting this in the formula, and replacing the constant 2 with the more theoretically founded von Karmann's constant $1/\kappa = 1/0.41$, the final expression was determined to be:

$$C_r = \sqrt{\frac{1}{\frac{1}{C_b^2} + \frac{1}{2g}C_{Dm}Dk}} + \frac{\sqrt{g}}{\kappa} \ln\left(\frac{h}{k}\right) \quad (21)$$

Although fractionally more complicated than the original formulation, this expression was found to be in good agreement with the data, especially in regions with higher Chézy values (Figure 2d). Furthermore, it is theoretically well founded, combining the (known) resistance for the flow inside the vegetation with the observed logarithmic profile above the vegetation (the sub-formula $\frac{\sqrt{g}}{\kappa} \ln\left(\frac{h}{k}\right)$). Subsequent research on the formulation revealed an agreement with the work by Kouwen [14], where a general formula for resistance induced by vegetation was proposed as:

$$C_r = C_1 + C_2 \ln\left(\frac{h}{k}\right) \quad (22)$$

Kouwen proposed several relationships for C_1 and C_2 , but no definitive conclusions were drawn. Equation 21 gives exact formulations for C_1 and C_2 and as such presents a step forward in the modelling of resistance. Note that the final formulation, Equation 21, is a combination of a computer induced expression that fits the data well, and theoretically based modifications to fit the theory. Further analysis revealed that the equation is equivalent to the equation produced by the method of effective water depth (Equation 6) if, (i) the depth balance in Equation 5 is ignored (i.e., the factors k and $(h - k)$ are considered equal to h), and (ii) the roughness length z_0 is set to $\frac{k}{e}$, where $e = 2.718$ is the base of the natural logarithm. It is still an open question why this particular approximation of the roughness length is optimal.

5. COMPARISON OF THE FORMULATIONS

Equation	RMSE 1DV Model	RMSE Flume Experiments
6	2.2	3.17
9	2.06	3.15
20	0.98	2.11
21	1.13	2.11
1DV Model	0	1.86

Table 2: Errors of the four formulations for the resistance coefficient with submerged vegetation, including the 1DV model that was used for generating the data.

Equation	CoD 1DV Model	CoD Flume Experiments
6	0.831	0.734
9	0.892	0.787
20	0.971	0.838
21	0.977	0.872
1DV Model	1.0	0.873

Table 3: Correlation squared of the four formulations for the resistance coefficient with submerged vegetation, including the 1DV model that was used for generating the training data.

There are several ways to evaluate the formulations. First and foremost is the ability to model the data under study. These results are presented in Table 2. It can be seen that the expressions based on the genetic programming results are in better agreement with the synthetic dataset than the manually induced formulations. Figure 2 presents the scatter plots for the four equations. Even though the manually improved formulation has a higher RMSE than the raw genetic programming formulation, the scatter plot reveals that the improved formulation removes the large residuals associated with high Chézy values, and thus is more reliable over the entire domain than the raw formulation. This is reflected in the correlations between the equations and the data as presented in Table 3. From this it can be seen that there is a small level of systematic error, even though it models the variance admirably.

Up to this point, both training and comparisons were performed on synthetic data, generated by the 1DV model. To ultimately test the approach, 177 experimental runs based on laboratory flume experiments were collected from 10 independent studies and used as a final validation set. The data contained all input information, except the bed roughness C_b , which was assumed to be negligible in the experiments, whereupon it was set to $60m^{1/2}s^{-1}$. For this particular dataset it was possible to also test the 1DV model itself. Results for the comparison can be found in Table 2, and it can be seen that the genetic programming induced equations give a highly competitive agreement with the data. What is particularly startling is that their performance is even competitive with the 1DV model the equations are based upon. Figure 4 presents a comparison between the improved genetic programming equation and the original 1DV model, both applied to the validation set of flume experiments. No serious discrepancies between the dynamical model and the simple equation are observed.

From the perspective of simplicity of the equations, it

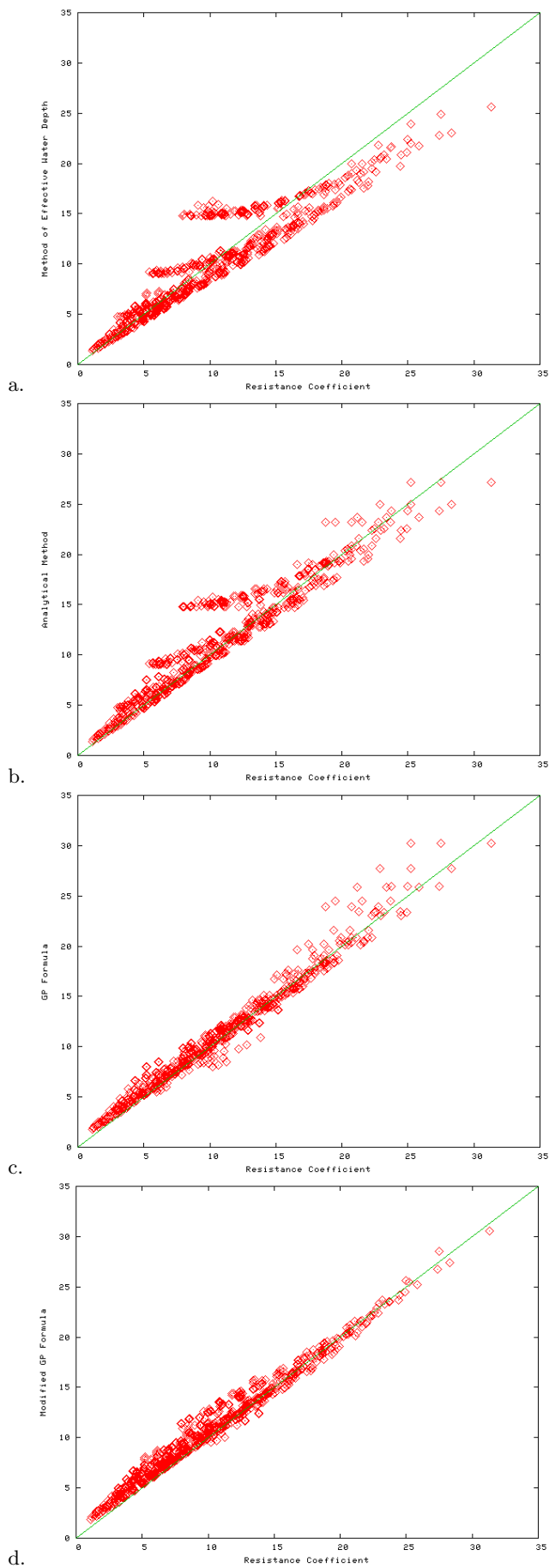


Figure 2: Scatter plots for the four equations: (a) Equation 6, (b) Equation 9, (c) Equation 20, and (d) Equation 21 on the training data.

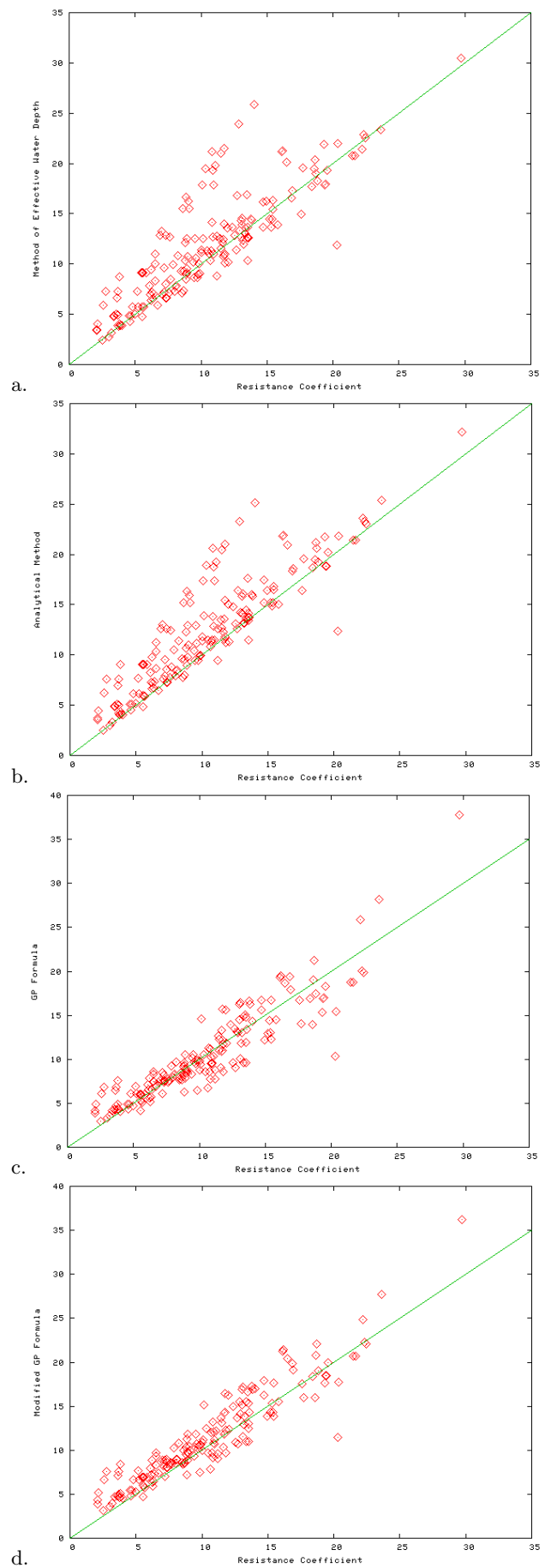


Figure 3: Scatter plots for the four equations: (a) Equation 6, (b) Equation 9, (c) Equation 20, and (d) Equation 21 on the validation data.

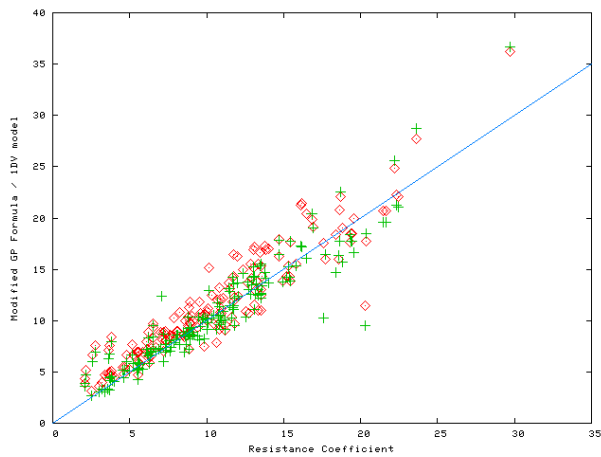


Figure 4: Comparison between the 1DV model (plusses) and the manually improved GP equation (dots) in their ability to model the validation data.

might be enlightening to compare the apparatus of expressions (Equations 7-18) that leads to the definition of Equations 6 and 9, with the conciseness of the genetic programming induced Equations 20 and its human-manipulated variant, Equation 21. The genetic programming induced equations are based purely on the measurements, and ignore considerations about length scale, turbulence intensity and velocity profiles. They focus primarily on obtaining good agreement with the data, and in this case, the problematic interplay between the vegetation and the turbulence induced by the vegetation apparently are of secondary importance to the simple logarithm on the ratio of water depth over plant height.

6. DISCUSSION

This paper presented a necessarily brief treatise of the induction of several equations to model the water depth dependent resistance induced by submerged vegetation in wetlands and floodplains. Two of the equations studied here were created through analysis and a process of derivation by a scientist. One equation was induced using a variant of genetic programming, where extra semantical information in the form of measurement units were processed to aid in subsequent analysis. The equation induced by genetic programming was superior both in the error induced and in the simplicity of the formulation. Subsequently an analysis on the equation was performed and a derivation for the formulation was found, and at the same time a theoretically motivated improvement of the formulation was created. In contrast with the human induced equations, the genetic programming induced equation focusses directly on the problem at hand as it is presented in the data. This potentially avoids elaborate search through phenomena that might be relevant to the intricate physical details, but less so for the particular problem at hand.

The genetic programming induced equation was competitive with the best of the human-induced equations. It was even competitive on experimental data with the dynamic model which was used for training. However, accuracy is not the entire story. Particularly important for this work was the possibility to (i) interpret the equation, and (ii) im-

prove it using theoretically motivated considerations. This in effect means that the genetic programming engine was used as a *hypothesis generator*, where instead of painstakingly deriving an answer through many logical steps, each step meticulously justified, answers are produced through automatic means, not as a black box, but as a tentative expression that can be used as a basis for analysis. This is potentially a much more useful result, as it shows that the symbolic nature of genetic programming can be used to build up knowledge in a problem domain. In contrast with many machine learning algorithms, where the trained model is the *end result* of a problem statement, the genetic programming induced expressions can be used to *start* a new cycle of inquiry.

The equations based on genetic programming are a scientific (or at least an engineering) result in its own right. Due to their simplicity and accuracy, a hydraulic engineer can simply calculate the resistance induced by vegetation using the equation, a few simple calculations, instead of setting up a complicated and computationally expensive model such as the 1DV model, nor does the engineer need to estimate turbulence intensity and related phenomena. The manually improved relationship presented here is both theoretically and experimentally justified, and can be used to estimate the resistance coefficient of submerged vegetation.

7. CONCLUSIONS

Four formulations for waterdepth-related resistance induced by vegetation are studied and compared. Two have been explicitly derived by a scientist building upon the extensive literature on this subject. One expression was derived by dimensionally aware genetic programming, and finally one expression was created by manually analysing and improving the genetic programming equation. It was found that the genetic programming equations were superior to the manually derived equations, both on their performance on synthetic training and laboratory testing data, and in the economy of detail that needs to be modelled. The manually improved expression was found to be in good agreement with previous expressions found in the literature, and performed competitive with the detailed model that was used to generate the training data.

This paper presented a case study in the use of genetic programming as a *hypothesis generator* for use in scientific discovery. Not only is genetic programming capable of producing equations that are comparable or better than their human derived competitors, it can produce expressions that are amenable to further analysis and manual improvement. The equation developed with the aid of genetic programming and modified using theoretical considerations is currently the most accurate and elegant formulation of resistance induced by submerged vegetation.

8. REFERENCES

- [1] Vladan Babovic and Maarten Keijzer. Genetic programming as a model induction engine. *Journal of Hydroinformatics*, 1(2):35–60, 2000.
- [2] Vladan Babovic and Maarten Keijzer. Rainfall runoff modelling based on genetic programming. *Nordic Hydrology*, 34(1), 2003.
- [3] Vladan Babovic, Maarten Keijzer, David Rodriguez Aquilera, and Joe Harrington. An evolutionary

- approach to knowledge induction: Genetic programming in hydraulic engineering. In Don Phelps and Gerald Sehlke, editors, *Proceedings of the World Water and Environmental Resources Congress*, volume 111, pages 64–64. ASCE, 20-24 May 2001.
- [4] M. J. Baptist. *Modelling floodplain biogeomorphology*. Ph.D. thesis, ISBN 90-407-2582-9, 193 pp., Delft University of Technology, Faculty of Civil Engineering and Geosciences, Section Hydraulic Engineering, 2005.
- [5] T. R. Campana. *Hydraulic resistance of submerged floodplain vegetation*. M.Sc. thesis H.E.043, IHE-Delft, 1999.
- [6] F. H. Dawson and F. G. Charlton. Bibliography on the hydraulic resistance of vegetated watercourses. Technical report, Freshwater Biological Association, Occasional Publication No. 25, ISSN 0308-6739, 25 pp., 1988. eigen kopie.
- [7] O Giustolisi. Using genetic programming to determine chézy resistance coefficient in corrugated channels. *Journal of Hydroinformatics*, 6(3):157–173, 2004.
- [8] H. T. M. Hong. *Hydraulic Resistance of Flexible Roughness*. M.Sc thesis H.H.237, IHE Delft, 1995.
- [9] Maarten Keijzer. Improving symbolic regression with interval arithmetic and linear scaling. In Conor Ryan, Terence Soule, Maarten Keijzer, Edward Tsang, Riccardo Poli, and Ernesto Costa, editors, *Genetic Programming, Proceedings of EuroGP'2003*, volume 2610 of *LNCS*, pages 71–83, Essex, 14-16 April 2003. Springer-Verlag.
- [10] Maarten Keijzer and Vladan Babovic. Dimensionally aware genetic programming. In Wolfgang Banzhaf, Jason Daida, Agoston E. Eiben, Max H. Garzon, Vasant Honavar, Mark Jakiela, and Robert E. Smith, editors, *Proceedings of the Genetic and Evolutionary Computation Conference*, volume 2, pages 1069–1076, Orlando, Florida, USA, 13-17 July 1999. Morgan Kaufmann.
- [11] Maarten Keijzer and Vladan Babovic. Declarative and preferential bias in GP-based scientific discovery. *Genetic Programming and Evolvable Machines*, 3(1):41–79, March 2002.
- [12] Soon Thiam Khu, Shie-Yui Liong, Vladan Babovic, Henrik Madsen, and Nitin Muttli. Genetic programming and its application in real-time runoff forecasting. *Journal of the American Water Resources Association*, 37(2):439–451, April 2001.
- [13] G. J. Klaassen, C. Stolker, E. H. Van Velzen, and H. Verheij. Naar een ruwheidsvoorspeller voor moerasvegetatie op basis van riet en gras. Technical report, WL | Delft Hydraulics, RWS/RIZA, 1999.
- [14] N. Kouwen, T. E. Unny, and H. M. Hill. Flow retardance in vegetated channels. *Journal of Irrigation and Drainage Division*, 95(IR2):329–342, 1969.
- [15] Shie-Yui Liong, Tirtha Raj Gautam, Soon Thiam Khu, Vladan Babovic, Maarten Keijzer, and Nitin Muttli. Genetic programming: A new paradigm in rainfall runoff modeling. *Journal of American Water Resources Association*, 38(3):705–718, June 2002.
- [16] F. López and M. H. García. Mean flow and turbulence structure of open-channel flow through non-emergent vegetation. *Journal of Hydraulic Engineering*, 127(5):392–402, 2001.
- [17] D. G. Meijer and E. H. Van Velzen. Prototype-scale flume experiments on hydraulic roughness of submerged vegetation. In *28th International IAHR Conference*, Graz, 1999.
- [18] N. Muttli and S. Y. Liong. Improving runoff forecasting by input variable selection in genetic programming. In Don Phelps and Gerald Sehlke, editors, *World Water Congress 2001*, volume 111, pages 76–76, Orlando, Florida, USA, 20-24 May 2001. ASCE.
- [19] H. M. Nepf and E. R. Vivoni. Flow structure in depth-limited, vegetated flow. *Journal of Geophysical Research*, 105(C12):28,547–28,557, 2000.
- [20] J. Nikuradse. Turbulente strömung in nichtkreisförmigen rohren. *Ing.-Arch.*, 1(306), 1930.
- [21] Z. Shi and M. R. Hughes. Laboratory flume studies of microflow environments of aquatic plants. *Hydrol. Process.*, 16:3279–3289, 2002.
- [22] B. M. Stone and H. T. Shen. Hydraulic resistance of flow in channels with cylindrical roughness. *Journal of Hydraulic Engineering*, 128(5):500–506, 2002.
- [23] R. Uittenbogaard. Modelling turbulence in vegetated aquatic flows. In *International workshop on RIParian FOfest vegetated channels: hydraulic, morphological and ecological aspects*, 20-22 February 2003, Trento, Italy, 2003.
- [24] L. F. Vernon-Harcourt. *Rivers and Canals, Vol. 1 Rivers*. the Clarendon Press, 1896.