Using genetic programming to determine Chèzy resistance coefficient in corrugated channels

Orazio Giustolisi

ABSTRACT

Genetic Programming has been used to determine Chèzy resistance coefficient for full circular corrugated channels. Three corrugated plastic pipes have been experimentally studied in order to generate data. The tests aim at measuring hydraulic parameters of the open-channel flow for some slopes, from 3.49–17.37% (2–10°), in order to discover the dependence of the channel resistance coefficient when wake-interference flow occurs. The monomial formula for the Chèzy resistance coefficient performs well on experimental data, both from measurement errors and from a technical point of view. In this paper, we present some very parsimonious formulae that have been created by Genetic Programming with few constants and which fit the data better than the monomial formula. Moreover, two of the Genetic Programming formulae, after 'physical post-refinement', seem to better explain the role of the roughness in the Chèzy resistance coefficient for corrugated channels with respect to its traditional expression for rough channels. This fact suggests that at least the structure of those formulae can be extrapolated to other types of corrugated channels. Finally, the work stresses the fact that the Genetic Programming hypothesis can be easily manipulated by means of 'human' physical insight. Therefore, Genetic Programming should be considered more than a simple data-driven technique, especially when it is used to perform scientific discovery.

Key words | genetic programming, evolutionary strategies, data mining, corrugated pipes

Orazio Giustolisi
Engineering Faculty of Taranto,
Technical University of Bari,
via Turismo no 8,
Paolo VI,
74100,
Taranto,
Italy
Email: *o.giustolisi@poliba.it*;
oraziogiustolisi@libero.it

INTRODUCTION

Three artificially corrugated channels have been experimentally studied. Laboratory tests aim at determining the Chèzy resistance coefficient. For this reason, about fifteen discharges and flow depths have been measured at slopes from 2–10°. In fact, it is observed that in this range of slopes wake-interference flow occurs and that the Chèzy resistance coefficient is a function of the hydraulic radius, channel slope, roughness-elements longitudinal spacing and height (Giustolisi 2001).

The wake-interference flow, also referred to in Morris's theory (Morris 1955, 1959) as hyper-turbulent flow, is caused by the particular channel surface that is constituted by macro wall-roughness elements which have constant longitudinal spacing and height along the pipe.

In hyper-turbulent flow, these roughness elements generate vortices which interfere with each other. Therefore, from a technical viewpoint, the hyper-turbulent flow is interesting because of the abnormal turbulence along the channel wall-roughness elements which generates additional dissipations, with respect to traditional rough channels, that cause a decrease of the average velocity of flow. This phenomenon is useful in drainage networks installed on sloping soils where usually the use of chutes to dissipate energy implies a higher cost of the project because they substitute standard manholes and, moreover, the installation of pipes has higher excavation costs compared to the situations in which it is possible to set pipes more or less at the same slope as the ground surface itself.

Therefore, it is very interesting to set up a formula for the resistance coefficient that, starting from its traditional expressions, is able to demonstrate the role of

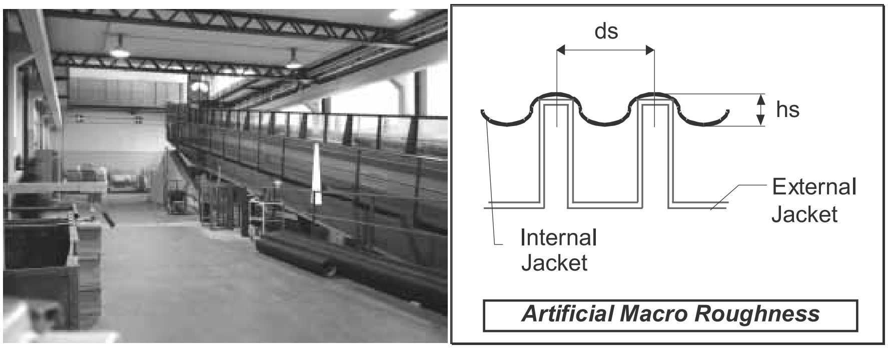

Figure 1 | Pictures of experimental apparatus and a detail of the artificial roughness.

the roughness index, ratio of the height to longitudinal spacing of the roughness elements and slope in corrugated channels.

For this reason, Genetic Programming (GP) is used as a tool to explore the domain of the resistance coefficient formulae that fit the experimental data well. Then, the user of GP has to select by means of input parameters the area of the domain where the machine must perform evolutionary searching of formulae that fit the data well and, finally, he has to choose the best formulae both from the fitting and the physical point of view.

In this way, GP allows us to use the knowledge based on data by means of the integration of 'human' physical insight and computer capability to explore the domains of the formulae.

DESCRIPTION OF EXPERIMENTAL TESTS AND HYDRAULIC PROBLEM

Artificially corrugated plastic pipes have been tested to slow down the open-channel flow in very steep culverts (Shipton & Graze 1976). Tests have been carried out at the Hydraulic Laboratory of the Technical University of Bari in Italy. Experimental data show a strong slow down of the average velocity of flow with respect to commercial rough pipes (Giustolisi 2001).

Good technical results make it interesting to determine the functional dependence for the Chèzy resistance coefficient in corrugated channels. Then experiments have been performed on three corrugated pipes. Therefore, this work concerns the setting up of a formula based on experimental data from three different corrugated channels.

Actually, experiments on corrugated pipes or large-scale roughness in channels of non-circular section already exist (Streeter 1936; Powell 1944; Morris 1955; Perry *et al*. 1969; Marone 1970; Shipton & Graze 1976; Pyle & Novak 1981; Ead *et al*. 2000), but a formula for the Chèzy resistance coefficient for corrugated circular channels does not exist.

Experimental facility and set-up

A closed circuit hydraulic apparatus of a rectangular channel of about 30 m in length with variable slope, see Figure 1, has been used to perform experiments on the three corrugated circular pipes in polyethylene. Discharge, flow depth and slope have been measured in uniform flow. The single corrugated channel length is 28.5 m and it is laid in the rectangular channel in order to conduct experiments. The three pipes have internal diameters of 182, 231.5 and 285 mm. The wall-roughness elements are respectively 5.5, 6.5 and 7 mm in height and they have a longitudinal spacing of 22.3, 26.5 and 30.4 mm respectively. The slope of the rectangular channel varies from 0–10°, but wake interference, i.e. hyper-turbulent flow, has been observed (Giustolisi 2001) at slopes from 2–10° (from 3.49–17.37%) (see in Table 1).

The profile of the flowing fluid is measured through transparent and easily removable lids in the pipe. A

Table 1 | Experimental slopes characterised by wake-interference flow occurring.

***S* (degrees)**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
S (%)	3.49	5.23	6.98	8.72	10.45	12.19	13.92	15.64	17.37

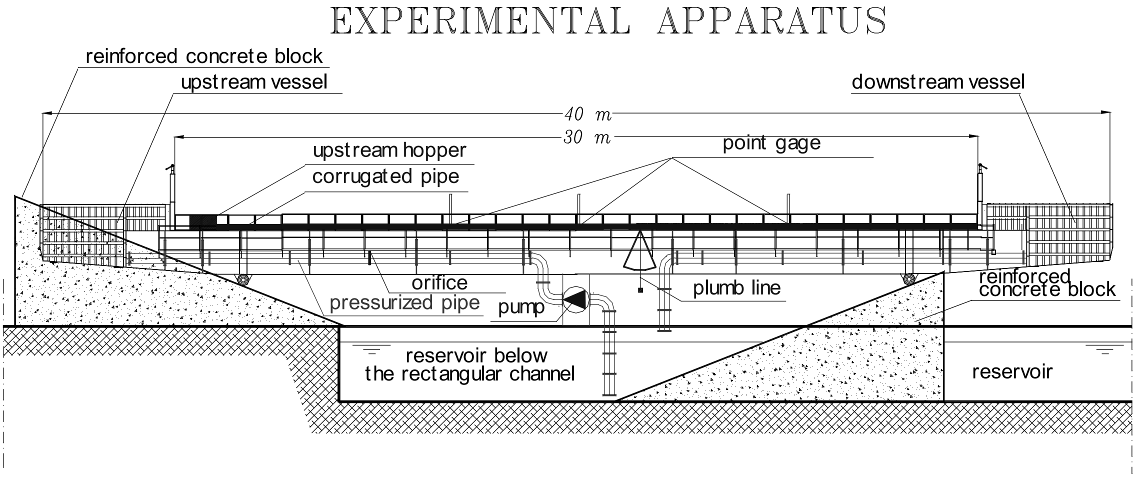

Figure 2 | Schematic of the experimental apparatus.

hopper, located upstream of the rectangular channel, conveys water into the corrugated channel. Water enters the hopper through a large upstream case. The flow at the end of the pipe overflows into the large downstream case which discharges into a reservoir below. A pump carries the flow from the reservoir to the upstream case, closing the hydraulic circuit, see Figure 2. Actually, two pressurised pipes carry the water to the upstream case; the smaller pipe has an internal diameter of 155 mm while the larger one has an internal diameter of 300 mm. Each contains an orifice to determine the discharge. The pipes contain valves which divert the flow into the appropriate orifice that is selected considering the discharge range (see Giustolisi (2001) for more details).

Measurements

For each slope, about fifteen measurements have been done of both the discharge and flow depth for the three channels. Then, 379 data values of hydraulic parameters have been collected; they correspond altogether to three values of the roughness index and to nine values of the slope. The measurement of each discharge has been performed by setting the filling level at about 20%, gradually increasing up to 80%, so covering a large range of Reynolds numbers, velocities and associated hydraulic parameters. During tests, down-flow scales have been plotted in real time, as well as the average velocity of flow and other hydraulic parameters, in order to better control the experiments and to avoid possible coarse errors. For this reason, the maximum estimated percentage error on flow depth and discharge measurements have been respectively less than 2% and 1%. Finally, slopes have been measured by means of a plumb line and the total error has been found to be absolutely negligible (see Giustolisi (2001) for more details).

Hydraulic notes

In open-channel uniform flow, the average velocity of flow, V, is related to resistance coefficient by (Yen 2002)

$$
\begin{aligned}
&V = \frac{K_n}{n} R^{2/3} S^{1/2} && \text{Manning}\\
&V = \sqrt{\frac{8g}{f}} \sqrt{RS} && \text{Darcy and Weisbach}\\
&V = C\sqrt{RS} && \text{Chèzy}
\end{aligned}
\tag{1}
$$

in which n, f and C are respectively the Manning, Darcy and Weisbach and Chèzy resistance coefficients; $R =$ hydraulic radius; $S =$ slope; $g =$ gravitational acceleration. $K_n = 1\ \mathrm{m}^{1/3}\ \mathrm{sec}^{-1}$ or $1.49\ \mathrm{ft}^{1/3}\ \mathrm{sec}^{-1}$ when n is chosen dimensionless and V and R are respectively expressed in SI units or in English units (Chow 1973).

Equations (1) allow us to relate the resistance coefficients of the Manning, Darcy and Weisbach and Chèzy formulae:

$$
\sqrt{\frac{f}{8}} = \frac{n}{R^{1/6}} \frac{\sqrt{g}}{K_n} = \frac{\sqrt{g}}{C} = \frac{\sqrt{gRS}}{V} \tag{2}
$$

In particular, the Darcy–Weisbach formula is derived from the pressurised pipes with the position $4R = D$, $D =$ diameter of the pipe

$$
\begin{aligned}
&S = \frac{h_f}{L} = f\frac{V^2}{8gR}\\
&f = F(R,K)
\end{aligned}
\tag{3}
$$

where $h_f =$ frictional loss; $L =$ length of the pipe; $R =$ Reynolds number; $K =$ relative roughness; $F =$ functional. K is equal to the ratio between $k_S =$ equivalent surface roughness and the hydraulic radius R.

Equation (3) says to us that, in uniform flow, the energy gradient h_f/L of the open-channel flow corresponds to its slope and, moreover, it shows better that, in open-channel flow, the Weisbach resistance coefficient must vary with the hydraulic radius, for a given wall roughness of the pipe, if the fully turbulent flow regime occurs, while it must also be a decreasing function of the Reynolds number, as experimentally determined in commercial pipes and shown in the Moody diagram, if turbulent flow or laminar regimes occur.

Moreover, relations in Equation (2) can be transformed into

$$
f = \frac{8n^2}{R^{1/3}} \frac{g}{K_n^2} = \frac{8g}{C^2} \quad \Leftrightarrow \quad \frac{C}{\sqrt{g}} = \frac{K_n}{n\sqrt{g}} R^{1/6} \tag{4}
$$

which shows that in fully turbulent open-channel flows K depends on $R^{-1/3}$, derived from the Manning formula, n is a constant related to surface roughness and the Chèzy resistance coefficient depends on $R^{1/6}$.

In summary, if fully turbulent flow occurs in our corrugated channel, it is expected that the Chèzy resistance coefficient is an increasing function of R, the Weisbach resistance coefficient is a decreasing function of R and n is constant. Then, at different slopes and fixed R, it is expected that C and f are constant.

Moreover, it is well known that fully turbulent flow occurs when the following inequality holds (French 1983):

$$
\frac{h_S u^*}{\nu} = \frac{h_S V\sqrt{g}}{C\nu} = \frac{h_S \sqrt{gRS}}{\nu} \geq 70 \tag{5}
$$

where $u^* =$ shear velocity; $h_S =$ roughness height; $\nu = \mu/\rho =$ kinematic viscosity; $\mu =$ viscosity of the fluid; $\rho =$ density of the fluid. In our pipe, Equation (5) allows us to estimate if fully turbulent flow occurs from the measured data, the hydraulic radius derived from the flow depth, the slope and the artificial roughness height.

Morris's wake-interference flow

Our corrugated channels showed a physical behaviour different from normal commercial rough pipes. This behaviour is caused by the wall-roughness elements on the pipe surface which have a constant longitudinal spacing and height. Moreover, the elements are macroscopic, having a radial height of 5.5, 6.5 and 7 mm. It is to be remembered that the Colebrook–White formula for full circular pipes (Colebrook & White 1937; Colebrook 1939; Yen 2002):

$$
\sqrt{\frac{1}{f}} = -2\log\left(\frac{k_S}{14.83R} + \frac{2.52}{4R\sqrt{f}}\right) \tag{6}
$$

is related to the Moody diagram, which differs from the Nikuradse diagram since the latter is obtained using experimental pipes with sand-grain surface roughness having a constant diameter. In fact, in the Moody diagram, the *f*–*R* (Weisbach resistance coefficient against Reynolds number) curves, after departing from the Blasius smooth pipe curves, show a dipping characteristic in the turbulent zone and, then, become horizontal in the fully turbulent one. In contrast, in the Nikuradse diagram, the *f*–*R* curves show, before becoming horizontal in the fully turbulent zone, a typical dip and rise. Therefore, the Colebrook–White formula does not predict the Nikuradse experiments because of the artificial roughness. For corrugated pipes, the situation is worse because it is known (Morris 1955; Marone 1970; Shipton & Graze 1976) that the Moody diagram does not predict at all the Weisbach resistance coefficient as well as the Chèzy and Manning coefficients.

Morris (1955, 1959) describes three regimes of flow that can originate in corrugated pipes. One of these explains very well our corrugated channel behaviour at slopes higher than 3.49%.

It is wake-interference flow that occurs at high Reynolds numbers, i.e. at higher velocities and slopes, when the longitudinal spacing of the roughness elements becomes not large enough with respect to the velocity at their crests. In this situation, the wake zone and the vortex-generating zone at each element interfere with each other and, in the zone near the wall, an abnormally intense turbulence appears. For this reason, Morris (1955) defines this regime as wake-interference flow and in his second paper (1959) he calls it hyper-turbulent flow. The name refers to the generation of a particular sub-layer, substituting for the viscous one, of intense flow mixing.

In fact, this hyper-turbulent sub-layer is characterised by an intense and complex vorticity and turbulent mixing while, separated by a transitional zone, in the central region of the channel, the normal turbulence, described by the Von Kàrmàn universal constant of turbulence *k*, prevails.

For this reason, in the flow that is subjected to this regime, it is possible to distinguish two fairly separated regions: the central zone of the channel with fully turbulent flow and the zone near the wall with hyper-turbulent flow.

The former region has a logarithmic velocity distribution, while the latter has a more flattened one, due to the intense flow mixing (Shipton & Graze 1976) which is given in logarithmic form in Morris's hypothesis. For this reason, Morris (1955) writes the expressions

$$
\begin{aligned}
&\frac{u}{u^*} = A + \frac{1}{k}\ln\left(\frac{y}{d_S}\right) = a + \frac{1}{k}\ln\left(\frac{y}{cd_S = y_0}\right) \quad \text{normal turbulent region} \\
&\frac{u}{u^*} = A_P + \psi\ln\left(\frac{y}{d_S}\right) = a + \psi\ln\left(\frac{y}{cd_S = y_0}\right) \quad \text{hyper-turbulent region}
\end{aligned}
\tag{7}
$$

where ln = natural logarithm; u = velocity at any point of the crossing section at distance y from the crests of the roughness; u^* = shear velocity; A and A_P = values of u/u^* at $y = d_S$ in the fully turbulent region and wall region respectively; ψ = slope of the logarithmic velocity distribution in the hyper-turbulent region; a = value of u/u^* at a distance y_0 that is the thickness of the hyper-turbulent sub-layer, which is assumed proportional to the longitudinal spacing d_S throughout the constant c. Morris integrates the previous expressions, the former for ($y > y_0$) and the latter for ($y < y_0$), and then the expression for the Weisbach resistance coefficient in wake-interference flow is

$$
\begin{aligned}
&\frac{1}{\sqrt{f}} = \frac{2.3}{k\sqrt{8}}\log\frac{2R}{d_S} + \frac{1}{\sqrt{8}}\left(A - \frac{3}{2k}\right) + \varphi\left(R_W = \frac{d_S\sqrt{f}\,R}{2R} = \frac{\sqrt{32}d_S\sqrt{gRS}}{\nu}, \text{element shape}\right) \\
&\varphi\left(R_W\right) \cong \frac{1}{\sqrt{8}}\left(\frac{1}{k} - \psi\right)\frac{cd_S}{R}
\end{aligned}
\tag{8}
$$

where log = decimal logarithm; R_w = Reynolds wall number calculated by roughness longitudinal spacing; and φ is an additive function due to the hyper-turbulent flow. It is stressed that, d_S/h_S being constant in our pipe, R_w is proportional to the value $h_S\,u^*/\nu$ of Equation (5).

The first of Equations (8), i.e. adopting $k = 0.40$ and $A = 8.7$, becomes

$$k_a = \frac{1}{\sqrt{f}} - 2\log\frac{2R}{d_S} = 1.75 + \varphi\left(R_W, \text{element shape}\right) \tag{9}$$

Actually, Morris uses the pipe radius because he is concerned with corrugated pipes more than with channels. For this reason, the pipe radius is substituted by $2R$ and Equation (9) corresponds to the fully turbulent flow equation if we make the substitution $k_S = d_S$ and $\varphi = 0$.

Equation (9) shows that, for a given roughness an additive function φ of the Reynolds wall number and of the roughness-element shapes, exists in hyper-turbulent flow. Clearly, the Morris resistance function k_a is constant in fully turbulent flow. However, it is known (Rand 1955; Perry *et al*. 1969; Pyle & Novak 1981; Marchi & Rubatta 1981) that the value 1.75 is not universal because k and A depend on roughness geometrical characteristics (longitudinal spacing, height, shape, etc.) and channel shape.

Moreover, Morris's theory predicts a positive decreasing value of the additive function—see the second of Equations (8)—and this explains the fact that, in the Moody diagram, the f–R curves rise in the region of fully turbulent flow. He infers that the function φ appears to be linear over a large range, but it should tend to zero at very high Reynolds wall numbers, or Reynolds numbers, as is clear from the first of Equations (8).

In fact, with increasing Reynolds numbers, the hyper-turbulent sub-layer of thickness y_0 should become more and more flattened by the central fully turbulent flow. For this reason, with increasing R_w, ψ should tend to $1/k$ and c should tend to zero, as $cd_S = y_0$.

For this flow regime, the hydraulic parameters must be correctly computed with reference to a datum at the crests of the roughness elements.

Range of hydraulic parameters and experimental monomial formula for Chézy resistance coefficient

Table 2 reports the range of some hydraulic parameters of flow computed by means of measured discharge Q, flow depth H and slope S. D is the internal diameter computed according to Morris's theory. Table 2 also shows that, in our experiments, fully turbulent flow always occurs, as $h_S u^*/\nu > 480$, while f is not constant, but it is quite variable with slope. Consequently, the Weisbach resistance coefficient, i.e. the Chèzy resistance coefficient, see Equations (2) and (4), does not depend on R alone as in fully turbulent flow (Giustolisi 2001). Then the study of the dependence of C has been carried out starting from the dimensionless expression of C in Equation (4) in rough channels adding to the function in Equation (3) for the dependence on S that is shown by our tests (Giustolisi 2001):

$$\frac{C}{\sqrt{g}} = \text{Cadim} = F(K,S) \tag{10}$$

Actually, being the equivalent surface roughness constant in our tests, the dependence of $K = k_S/R$ in Equation (10) means a dependence on hydraulic radius. Therefore, the selection of a monomial formula for the Chèzy resistance coefficient makes for easy parameter estimation and it gives

$$\begin{aligned} \text{Cadim} &= \frac{707}{\sqrt{g}} R^{0.12} S^{-0.116} \left(\frac{ds}{hs}\right)^{-2.235} \\ &= 226 R^{0.12} S^{-0.116} \left(\frac{ds}{hs}\right)^{-2.235} \end{aligned} \tag{11}$$

Table 3 reports the fitting properties of Equation (11) according to AVG (Average error), CoD (coefficient of determination) and RMS (root mean squared error) error functions:

$$\begin{aligned} \text{AVG} &= \frac{\sum\limits_{\text{No. of data}} \sqrt{\left(1 - X_{\text{computed}}/X_{\text{experimental}}\right)^2}}{\text{No. of data}} \\ \text{RMS} &= \sqrt{\frac{\sum\limits_{\text{No. of data}} \left(X_{\text{computed}}/X_{\text{experimental}}\right)^2}{\text{No. of data}}} \\ \text{CoD} &= 1 - \frac{\text{No. of data} - 1}{\text{No. of data}} \\ &\quad \frac{\sum\limits_{\text{No. of data}} \left(X_{\text{computed}}/X_{\text{experimental}}\right)^2}{\sum\limits_{\text{No. of data}} \left(X_{\text{experimental}} - \text{Mean}\left(X_{\text{experimental}}\right)\right)^2} \end{aligned} \tag{12}$$

Table 2 | Hydraulic parameters.

Measured				**Calculated**				
S **(degrees)**	***D*** **(mm)**	***Q*** **(l/sec) (min–max)**	***H/D*** **(min–max)**	***h***$_S$**u*/ν ×10**$^{-3}$ **(min–max)**	***V*** **(m/sec) (min–max)**	**R ×10**$^{-5}$ **(min–max)**	***R*** **×10**^{2} **(m) (min–max)**	***f*** **(min–max)**
2	271	12.6–74.89	27.3–78.8	0.84–1.17	0.99–1.54	1.67–5.02	4.28–8.24	0.095–0.121
	218.5	13.2–50.2	34.1–79.2	0.77–0.97	1.17–1.57	1.92–4.16	4.14–6.64	0.073–0.082
	171	2.90–25.46	28.7–76.1	0.48–0.73	0.90–1.36	0.70–2.78	2.22–5.17	0.079–0.104
3	271	9.17–90.07	16.9–81.3	0.93–1.43	0.88–1.79	0.98–5.87	2.81–8.25	0.105–0.150
	218.5	4.7–55.8	18.5–76.2	0.73–1.19	0.98–1.82	0.96–4.76	2.46–6.61	0.079–0.103
	171	7.02–32.04	32.0–81.5	0.69–0.89	1.10–1.60	1.35–3.30	3.08–5.20	0.083–0.110
4	271	7.59–97.52	18.5–77.5	1.00–1.65	1.04–2.03	1.25–6.63	3.04–8.22	0.109–0.155
	218.5	4.8–62.2	17.6–74.4	0.82–1.37	1.08–2.08	1.00–5.43	2.34–6.58	0.082–0.108
	171	3.25–36.07	19.9–83.2	0.65–1.03	1.10–1.66	0.81–3.64	2.05–5.20	0.088–0.111
5	271	12.1–96.37	15.4–71.3	1.21–1.82	1.29–2.19	1.06–7.02	2.57–8.07	0.115–0.169
	218.5	5.3–70.4	17.6–77.3	0.92–1.53	1.19–2.26	1.11–5.94	2.35–6.61	0.088–0.113
	171	4.84–39.87	23.0–83.7	0.77–1.15	1.21–1.92	1.12–3.96	2.33–5.20	0.096–0.108
6	271	9.43–105.9	18.5–72.9	1.23–2.00	1.29–2.35	1.55–7.57	3.04–8.12	0.121–0.150
	218.5	5.8–73.5	17.6–74.6	1.00–1.68	1.30–2.45	1.21–6.42	2.34–6.58	0.089–0.113
	171	8.10–40.18	29.5–80.2	0.94–1.26	1.43–2.03	1.64–4.20	2.88–5.20	0.103–0.115
7	271	7.18–118.6	15.7–76.4	1.23–2.17	1.24–2.51	1.29–8.17	2.63–8.20	0.121–0.164
	218.5	5.8–80.4	16.9–74.3	1.06–1.82	1.38–2.56	1.24–6.72	2.26–6.63	0.097–0.114
	171	3.49–40.06	19.1–76.2	0.84–1.36	1.13–2.13	0.89–4.38	1.98–5.17	0.109–0.146
8	271	9.04–122.4	17.1–74.9	1.47–2.32	1.38–2.64	1.55–8.57	2.84–8.17	0.128–0.162
	218.5	5.6–82.3	16.0–77.0	1.11–1.94	1.44–2.65	1.23–6.96	2.16–6.62	0.100–0.114
	171	4.99–40.04	22.6–75.5	0.97–1.45	1.28–2.17	1.17–4.45	2.30–5.16	0.119–0.153
9	271	11.4–119.4	18.6–70.4	1.51–2.44	1.54–2.75	1.87–8.78	3.06–8.04	0.128–0.159
	218.5	5.6–83.6	15.7–73.9	1.16–2.05	1.48–2.81	1.25–7.34	2.11–6.56	0.102–0.118
	171	4.98–40.31	21.6–73.0	1.00–1.53	1.36–2.24	1.19–4.56	2.20–5.12	0.125–0.145
10	271	12.2–129.1	18.7–73.7	1.69–2.59	1.63–2.83	1.99–9.16	3.08–8.15	0.135–0.159
	218.5	5.4–85.9	15.2–74.6	1.21–2.16	1.50–2.86	1.23–7.48	2.05–6.58	0.109–0.124
	171	3.62–40.37	18.3–72.0	0.98–1.61	1.26–2.28	0.95–4.62	1.90–5.10	0.134–0.158

Table 3 | Errors of the hydraulic parameters computed by Equation (11). AVG_% and CoD_% are, respectively, AVG and CoD in percentages.

Hydraulic parameter	AVG_%	RMS	CoD_%
Discharge Q	2.5342	9.368×10^{-4}	99.99987
Velocity of flow V	2.5342	6.343×10^{-2}	99.407
Chèzy coefficient Cadim	2.5342	2.823×10^{-1}	88.2543

where X is the dimensionless Chèzy resistance coefficient, Cadim, of Equation (4).

The monomial formula of Equation (11) fits sufficiently well the data but it has some limits:

1. It has four parameters, but a more parsimonious expression should be more easily extrapolated, especially outside the range of the experimental roughness index [0.230 0.247].
2. Equation (11) allows us to explain the role of the new parameter S, due to the rise of the f–$\boldsymbol{R}$ curves in the Moody diagram for wake-interference flow that alters the quadratic dependence of Equation (3) in fully turbulent flow (Giustolisi 2001), but which leaves some interpretations open about the joint role of slope and roughness index.

For this reason, in this work GP is used to better explore the domain of Chèzy resistance formulae in order to confirm the dependencies of the monomial formula and to have a more parsimonious one that could be a better candidate for performing extrapolation, especially with respect to the roughness index. In fact, the experimental ranges of parameters R and $\boldsymbol{R}$, i.e. S, are quite large.

GENETIC PROGRAMMING APPROACH

A Genetic Algorithm (GA) is a machine learning paradigm which derives its behaviour from a metaphor for the processes of evolution in nature. This is done by the creation, within a machine, of a population of individuals represented by chromosomes, essentially a set of character strings that are analogous to chromosomes observed in our own DNA. The individuals in the population go through a process of evolution. Evolution is not a purposeful or directed process. Indeed, the processes of nature seem to boil down to different individuals competing for resources in the environment. Some fit better than others. Those that fit better are more likely to survive and propagate their genetic material. In nature, the encoding for our genetic information (genome) is done in a way that admits sexual reproduction. Asexual reproduction (such as by budding) typically results in offspring that are genetically identical to the parents. Sexual reproduction allows the creation of genetically radically different offspring that are still of the same general flavour (species). At the molecular level strings of chromosomes bump into one another, exchanging chunks of genetic information and drift apart (this is a wild oversimplification of course). This is the *recombination* operation, called *crossover*, because of the way that genetic material crosses over from one chromosome to another. The crossover operation happens in an environment where the selection is related to the *fitness* of the individual, i.e. how good the individual is in competing in its environment. *Mutation* also plays a role in this process, though it is not the dominant role that is popularly believed to be the process of evolution, i.e. random mutation and survival of the fittest. It cannot be stressed too strongly that GAs (as a simulation of a genetic process) do not perform *a random search* for a solution to a problem. GAs use stochastic processes, but the result is distinctly non-random (better than random).

GP is the extension of the genetic model of learning into the area of programs (Koza 1992), i.e. formulae for the Chèzy resistance coefficient in our case. That is, the objects that constitute the population are not fixed-length character strings that encode possible solutions to the problem at hand. In our case, they are symbolic regressions that, when executed, are the candidate solutions to the Chèzy resistance coefficient modelling problem. These symbolic regressions are expressed as parse trees of variable length. The formulae in the population are composed of elements from a *function set*, a *terminal set* and *constants*. The function set is selected by the user and, in our case, it is related to the choice of function types that are

Table 4 | Input parameters in GP. The optional parameters are respectively related to the first and second approaches.

Parameter	Value	Parameter	Value
Number of experiments	120	Number of generations to run	1,000 or 500
Population size	1,000	Number of children	1,000
Maximum length of parse tree	20 or 30	Training percentage	100% or 75%
Cost functions	CoD, RMS	Set of functions	*, /, ln, pow
Target	Cadim	Inputs	R, S, h_S and d_S
Breeding method	Tournament	Tournament size for replacement	3
Constant mutation probability	0.9	Crossover rate	0.9
Self-crossover	0.3	Subtree mutation probability	0.3
Swap mutation rate	0.3	Constant probability	0.5

candidates to play a role in the symbolic regression. The terminal set is composed by inputs or constants that are arguments of the function set.

In this work, GP is used in the hydraulic field (Babovic 1996; Babovic & Abbott 1997; Davidson *et al*. 1999; Babovic & Keijzer 2000; Babovic *et al*. 2001) to perform knowledge based on data, then, in order to find a symbolic expression for the dimensionless Chèzy roughness coefficient. For this reason, the target of GP is the experimental dimensionless Chèzy resistance coefficient and inputs are experimental values of slope, hydraulic radius, roughness height and longitudinal spacing of the roughness elements. The function set is composed of product (*), ratio (/), natural logarithm (ln) and power (pow), and this is an implicit bounding of the formula space domain where GP must search the best. GP searches, in an evolutionary manner, formulae that better fit the target data. The user is free to select the cost function to measure fitting.

Naturally, selection of a function set limits the searching area in the space of formulae by physical choice. The choice of natural logarithm, product, ratio and power in the function set is supported by formulae given in Equations (8) and (11).

Moreover, a so-called dimensionally aware GP (Keijzer & Babovic 1999) has been used and, therefore, dimensional information of d_S and h_S have been used while the hydraulic radius R has been treated as dimensionless, as explained above. In fact, the geometric factor k_S, describing the micro-roughness of the three pipes, should be constant and therefore it does not affect the functional dependence of Cadim. Dimensional information in GP acts as a human preferential bias (Keijzer & Babovic 1999, 2002) that supplies good results in improving GP exploration without limiting the search area.

On the maximum length of the parse tree, two approaches have been tested. In the first approach GP has been forced to find a formula that is no longer than Equation (11). Therefore we have chosen the maximum length of the parse tree equal to 20, see Table 4. This avoidance overfitting technique is based on prior limiting the complexity of the formula by expert choice, that is, performing a sort of Minimum Description Length principle. This should have the same effect as a cross-validation technique and it is useful when the amount of data is not high. Therefore, at first cross-validation is not performed and during the evolution the fitness of the cost

Table 5 | Original best GP formulae. AVG_% and CoD_% are, respectively, AVG and CoD in percentages

Expt. no.	**AVG_%**	**RMS**	**CoD_%**	**Hypotheses/formulae**
11	2.649	0.2777	88.634	$\text{Cadim} = \left[x(3)x(1)^{R} / x(2)^{(d_S S)}\right]^{x(4)h_S/d_S}$
99	2.561	0.2838	88.135	$\text{Cadim} = \left(x(1)\, S / R^{x(3)}\right)^{x(2)h_S/d_S}$
83	2.700	0.2839	88.127	$\text{Cadim} = x(1)\left\{\mathbf{ln}\left[S/R\; h_S{}^{-x(2)d_S/h_S + 2}\right]\right\}^{-x(3)}$
95	2.609	0.2851	88.027	$\text{Cadim} = (x(1)R^{x(2)}S^{x(2)-1})^{(h_S/d_S)x(3)}$
66	2.700	0.2907	87.553	$\text{Cadim} = x(2)\left(S/R\right)^{x(1)}\left(d_S/h_S\right)^{2x(3)}$
101	2.614	0.2932	87.338	$\text{Cadim} = \left(-\,x(3)\,\mathbf{ln}\left(x(1)S/R\right)\right)^{x(2)h_S/d_S}$

Expt. no.	***x*(1)**	***x*(2)**	***x*(3)**	***x*(4)**
11	1.577e + 01	4.471e − 01	1.337e + 01	2.715e + 00
99	1.569e − 08	− 5.150e − 01		
83	6.726e + 01	8.257e − 01		
95	4.9000e − 01	1.015e + 04		
66	− 1.160e − 01	1.596e + 02		
101	2.047e − 04	4.335e + 00		

functions has been calculated by the whole set of data. Subsequently, in the second approach cross-validation has been performed and we have chosen the maximum length of the parse tree equal to 30, see Table 4, because of use of the overfitting avoidance technique.

On cost functions, after the first experiments where RMS or CoD has been used to calculate the fitness without getting results, the choice to simultaneously optimise RMS and CoD, using a Pareto front, supplied the best results.

On the other parameters of GP, the choice of performing a lot of experiments with one thousand runs, high mutation, crossover, constant probabilities, population size and number of children provided good results. However, the opposite choice of fewer experiments with higher numbers of runs and low mutation, crossover and constant probabilities did not work well: see some input parameters of GP in Table 4.

Finally, the second approach seems to require a smaller number of generations to run in order to provide results.

RESULTS AND DISCUSSION

Results without cross-validation (first approach)

Tables 5 and 6 report the best GP formulae that it has supplied during the 120 experiments. The values of the constants of the formulae in Table 6 are not equal to those

Table 6 | Best GP formulae after mathematical post-refinement

Expt. no.	AVG_%	RMS	CoD_%	Hypotheses/formulae
11	2.6027	0.2691	89.329	$\text{Cadim} = \left[x(3)x(1)^R / x(2)^{(d_S S)}\right]^{x(4)h_S/d_S}$
99	2.5038	0.2801	88.444	$\text{Cadim} = \left(x(1)\, S / R^{x(3)}\right)^{x(2)h_S/d_S}$
83	2.7676	0.2790	88.527	$\text{Cadim} = x(1)\left\{\mathbf{ln}\left[S/R\, {h_S}^{-x(2)\,(d_S/h_S)\,+\,2}\right]\right\}^{-\,x(3)}$
95	2.5257	0.2806	88.400	$\text{Cadim} = (x(1)R^{x(2)}S^{x(2)\,-\,1})^{(h_S/d_S)x(3)}$
66	2.6992	0.2901	87.600	$\text{Cadim} = x(2)\left(S/R\right)^{x(1)}\left(d_S/h_S\right)^{2x(3)}$
101	2.4836	0.2789	88.540	$\text{Cadim} = \left(-\,\underline{x(3)}\,\mathbf{ln}\left(x(1)S/R\right)\right)^{x(2)h_S/d_S}$

Expt. no.	*x*(1)	*x*(2)	*x*(3)	*x*(4)
11	2.503e + 01	4.637e − 01	1.035e + 01	2.944e + 00
99	5.449e − 09	− 4.861e − 01	1	
83	7.129e + 01	8.463e − 01	1	
95	5.053e − 01	1.079e + 04	1	
66	− 1.153e − 01	1.599e + 02	1	
101	1.020e − 04	4.170e + 00	1	

Expt. no.	*x*(1)	*x*(2)	*x*(3)	*x*(4)
11	2.503e + 01	4.637e − 01	1.035e + 01	2.944e + 00
99	6.015e − 09	− 4.879e − 01	9.904e − 01	
83	8.774e + 01	8.677e − 01	1.071e + 00	
95	5.005e − 01	1.015e + 04	1.004e + 00	
66	− 1.151e − 01	1.597e + 02	9.998e − 01	
101	8.440e − 05	4.263e + 00	9.338e − 01	

of the formulae in Table 5—the original hypothesis from GP—because the formulae are re-optimised by traditional nonlinear least squares, LSQ, using as the initial search point the constant vector $x(i)$ of Table 5, to give a 'mathematical post-refinement' of the results. This post-refinement generally improves the formulae from the point of view of all statistical coefficients. The exception is formula no. 83, which appears to be worse for AVG_%, but this is possibly because, actually, LSQ optimises the RMS cost function. From Table 6, the six formulae

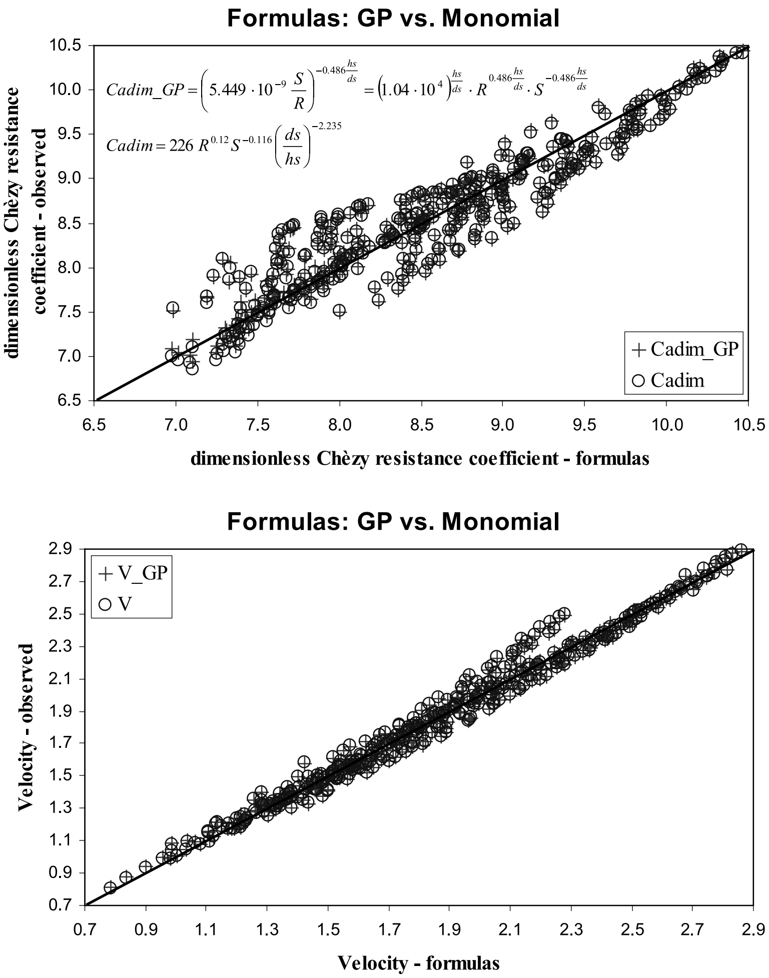

Figure 3 | Comparison between GP (expt. no. 99) and monomial formulae.

generally perform better than the monomial of Equation (11) and GP confirms the dependence of the Chèzy roughness coefficient on R, S, d_S and h_S. Moreover, Table 6 shows that GP supplies five formulae having two constants, compared to four constants of the formula in Eq. (11).

Also, if GP supplies two constants $x(i)$ in the last five formulae, a deeper post-refinement of GP results is tried by adding parameter $x(3)$ to be estimated by LSQ optimisation. Values in Table 6 near unity of $x(3)$ indicate that the five formulae do not need a constant $x(3)$.

In conclusion, GP generally supplies formulae that, after mathematical post-refinement, generally fit experimental data better than the monomial formula, but in a more compact way and including geometrical information about the roughness index in a different way.

Selection of the best formula

From the formulae reported in Table 6, no. 99 has been selected as it performs better than the monomial formula, see Figure 3 and the comparison of AVG_%–RMS–CoD_% in Tables 3 and 6. It is also characterised by a structure that has some aspects that will be useful.

In fact, the two formulae, Equation (11) and experiment no. 99, have similar structures, but they differ in the way in which the geometric roughness index acts in it:

$$\mathrm{Cadim} = 226R^{0.12}\,S^{-0.116}\left(\frac{d_S}{h_S}\right)^{-2.235}$$

$$\mathrm{Cadim|GP} = \left(5.449 \times 10^{-9}\frac{S}{R}\right)^{-0.448h_S/d_S}$$

$$= \left(1.04 \times 10^{4}\right)^{h_S/d_S} R^{0.486h_S/d_S}\,S^{-0.486h_S/d_S} \quad (12)$$

as an exponent of the constant term, hydraulic radius and slope in the GP formula, and as an isolated term in the monomial expression. Starting from the second of Equation (12), we can write

$$\mathrm{Cadim|GP} = x(1)^{h_S/d_S}\,R^{x(2)h_S/d_S + 1/6}\,S^{x(3)h_S/d_S} \quad (13)$$

which is more suitable from a physical point of view because the exponent of the hydraulic radius should tend to 1/6 and the exponent of the slope should tend to zero, as already correctly indicated in Equation (13), when h_S is negligible with respect to d_S, see Equation (4). Optimising by GA the vector $x(i)$ in Equation (13), we finally obtain

$$\mathrm{Cadim|GP} = 15770^{h_S/d_S}\,R^{-0.1039h_S/d_S + 1/6}\,S^{-0.4460h_S/d_S} \quad (14)$$

that has AVG_% = 2.5458% RMS = 0.2955 CoD_% = 87.1353%. This last ‘physical post-refinement’ of the GP formula does not degrade much the fitness properties while it introduces a physical insight into the formula for the Chèzy roughness coefficient.

It is clear that both the formulae in Equation (14) and the GP formula in Equation (12) are valid in the range of hydraulic parameters of Table 2 and inside the experimental roughness index range of [0.230 0.247].

Despite this fact, the two formulae of GP are very interesting because they explain the effect of the geometry of the macro-roughness more clearly and both of them are more parsimonious than the monomial formula.

From a technical point of view, we can consider the use of either the second of Equation (12) and Equation (14) or the monomial formula that fit more or less equally well in the range of the experimented roughness indexes. Out of this range, we believe that the formula in Equation (14) is to be recommended because it fits sufficiently well the experimental data, it is parsimonious and it is physically interpretable. In fact, the presence of the roughness index is the physical condition that generates wake-interference flow and this varies the traditional dependence on hydraulic radius of the Chèzy resistance coefficient in a rough pipe, see Equation (4), and causes the increase of the ***f*–*R*** curves in the Moody diagram, and therefore the dependence on slope S (Giustolisi 2001).

Results with cross-validation (second approach)

When cross-validation is performed, GP supplies formulae that have in general a less parsimonious structure, due to the higher value of the maximum length of the parse tree, without good fitting properties. Here we report the best fitting formula

$$\mathrm{Cadim|GP} = \ln\frac{R{h_S}^{(d_S/h_S)-2}}{S} + 20.2005 \quad (15)$$

that, despite this, is very short and fits well the whole set of data (AVG_% = 2.4496% RMS = 0.2499 CoD_% = 90.7957%). However, there is no evidence that this result is strictly related to the use of cross-validation because, for example, experiment no. 83 in Table 5 is quite similar and Equation (15) could have been supplied by the first GP approach before the 120th experiment considering its good fitting properties regarding the whole set of data and its parsimonious structure. Perhaps we could argue that cross-validation makes GP more selective during exploration of hypothesis/formulae space.

Also in this situation, a mathematical and physical post-refinement has been performed on Equation (15), as done above, but now considering the formulae in Eq. (8) and using logarithmic properties. The result is the formula

$$\mathrm{Cadim|GP} = \frac{1}{0.534}\,\mathbf{ln}\,\frac{2R}{h_S} + \left(1.0973\,\mathbf{ln}\,\frac{{h_S}^{d_S/h_S}}{RS} + 20.6863\right) \quad (16)$$

that fits very well the whole set of data (AVG_% = 2.1709% RMS = 0.2295 CoD_% = 92.2412%), see Figure 4.

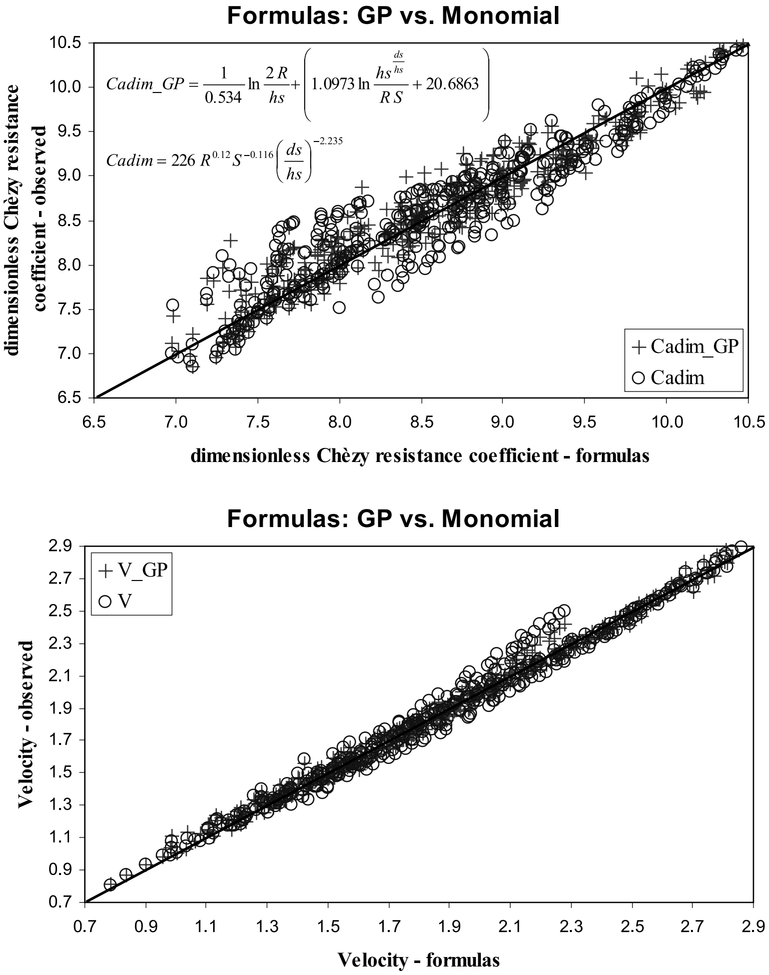

Figure 4 | Comparison between GP and monomial formulae.

Moreover, re-writing the first of Equation (8) thanks to the equalities in Equations (2) and (10):

$$\text{Cadim} = \sqrt{\frac{8}{f}} = \frac{1}{k}\,\mathbf{ln}\,\frac{2R}{d_S} + \left[\left(A - \frac{3}{2k}\right) + \varphi\left(R_W, \text{element shape}\right)\right] \tag{17}$$

We can compare Equations (16) and (17) in order to stress the fact that GP has supplied a formula similar to the theoretical one where $k = 0.534$, a physically compatible value considering its dependence on the roughness of the channel as mentioned above, and d_S is substituted by h_S. Naturally, it is possible to find the ratio R/d_S in Equation (16) by means of mathematical manipulation based on logarithmic properties, but Equation (16) is the most parsimonious structure.

In conclusion, Equation (16), as well as Equation (14), is recommended because it fits very well the experimental data, it is parsimonious, physically interpretable directly from Morris's theory and is technically feasible requiring a knowledge of d_S, h_S, R and S to compute the Chèzy resistance coefficient.

GP: DATA-DRIVEN TECHNIQUE OR MORE?

A common framework to classify modelling techniques divides them into three categories:

- white box, i.e. completely physically based models, whose mathematical structure and parameters are previously known;
- grey box, i.e. conceptual models, whose mathematical structure is known from physical insight or conceptualisation, but whose parameter estimation is needed by means of data;
- black box, i.e. data-driven models, whose mathematical structure and parameters are not previously known and data are needed to determine both a mathematical structure and parameters.

The above classification is not exhaustive but it is useful for our discussion. It demonstrates the fact that white box models theoretically do not need to monitor data information, while grey box and black box models use this information at different levels.

The problem in the use of data is that they have a limited information content, i.e. we have sparse points in the domain of the model (the curse of dimensionality) that are corrupted by noisy measurements and secondary physical effects whose input is not known. Highly flexible data-driven models tend to fit noisy data, that is they are generally able to fit training data well but this, called overfitting, causes poor generalisation properties (the model is not able to predict unseen data).

For this reason, the aim is to improve modelling by means of other physical information about the mathematical structure of the model, that is shifting from completely black box models toward grey box ones. In this way, the same information content of the data is used to determine fewer characteristics of the model.

For example, in Multilayer Perceptron avoidance, overfitting techniques (Giustolisi 2002) improve generalisation, i.e. by a phenomenon of information smoothness, called Tikhonov regularization (Tikhonov 1963). In Support Vector Machines (Vapnik 1995) the definition of the ε-tube is external information regarding the level of non-Gaussian noise; cross-validation pragmatically allows us to avoid overfitting, controlling it by means of information in the validation subset of the data.

In this context, GP can be classified as a data-driven technique because it uses data to find both a mathematical structure and parameters, while in Multilayer Perceptron and Support Vector Machines the structure is a mathematical expansion based on a prior selected kernel, which is general in approximating functions, and the data are used to estimate parameters.

The advantage of the GP technique is that it easy to deal with physical prior knowledge perhaps because it works in a similar way as humans, especially when applied to experimental data from the laboratory used to perform scientific discovery, as in this work. Moreover, GP can be forced to use information in data to find a mathematical feasible structure more than parameters and this supplies a parsimonious hypothesis. In fact, in this work, prior knowledge about physical phenomena are inserted into GP during its setting up in terms of the selection of function set type and dimensional information (Keijzer & Babovic 2002). Moreover, the selection of the maximum length of the tree, related to the monomial formula length, requires expert physical insight and an application of the Minimum Description Length principle, while performing cross-validation is an alternative. Prior choices in GP before an evolutionary search has shifted it from being data-driven towards a grey box technique, allowing us to avoid typical problems, such as overfitting, of the traditional overparametrized black box techniques. The parsimony of the formulae/hypotheses supplied by GP, which at the same time fit the experimental data well (considering the measurement errors), is the result.

The mathematical post-refinement of the formulae/hypotheses uses one more time data to improve fitting, but now in a context of parsimonious formulations and, therefore, far from the overfitting problems generated by an excess of parameters. The final physical post-refinement of the expert-selected formula by means of manipulation (Keijzer & Babovic 2002) is very important because it can be a decisive shift of GP toward physically based modelling. In fact, the final physical post-refinement is in the direction of a physical formula that could fit the worst experimental data, for example Equation (14), but it supplies more information about the effect in a case study of hydraulic/geometric variables in the resistance coefficient. Therefore, if overfitting occurs the physical post-refinement could be seen as a *post-avoidance overfitting technique*.

In summary, we believe that, strictly speaking, GP is a data-driven technique, but prior knowledge during the setting up of the evolutionary search and final physical post-refinement (Keijzer & Babovic 2002) of the hypothesis should make it very close to a white box technique, especially when GP is used in scientific discovery problems.

CONCLUDING REMARKS

GP (Babovic 1996; Babovic & Abbot 1997; Davidson *et al*. 1999; Babovic & Keijzer 2000; Babovic *et al*. 2001) has been applied to the determination of the Chèzy roughness coefficient for corrugated channels in wake-interference flow, i.e. hyper-turbulent flow (Giustolisi 2001).

The novelty of this application is that the author, more trained in the specific hydraulic problem, takes advantage of the knowledge discovery technique based on his integration with GP to improve the Chèzy resistance coefficient formula with respect to the monomial one. It is notable that GP quite easily and quickly supplies at least two good formulae that fit the experimental data better and are more parsimonious than the monomial formula. Moreover, GP has supplied six parsimonious expressions (one or two constants compared to four for the monomial formula) for the Chèzy resistance coefficient, all confirming the dependencies on hydraulic radius, slope and roughness index.

Finally, the two new formulae for the Chèzy resistance coefficient, derived from these GP formulae by means of 'mathematical/physical post-refinement', are suitable for explaining the effect of the macro-roughness elements, with respect to the behaviour of the rough commercial channels and their traditional expressions for resistance coefficients (Morris's theory and monomial expressions). Therefore, the work seems to indicate that this approach, which combines data-mining techniques together with a theoretical understanding, provides very good results. In fact, Equations (14) and (16) show the balance between experimental data and the physical interpretation of the roughness index that generates hyper-turbulent flow and fitting properties.

ACKNOWLEDGEMENTS

This work has used software developed under the Talent Project No. 9800463 entitled 'Data to Knowledge–D2K' funded by the Danish Technical Research Council (STVF) and the Danish Hydraulic Institute (DHI). For more information see http:/www.d2k.dk.

NOTATION

$A =$ value of u/u^* for normal turbulent region velocity distribution at $y = dS$ in Morris (1955)

$A_p =$ value of u/u^* for wall-velocity distribution at $y = d_S$ in Morris (1955)

$a =$ value of u/u^* at distance y_0 (boundary between turbulent and hyper-turbulent regions) in Morris (1955)

$C =$ resistance coefficient of Chèzy formula

Cadim = dimensionless resistance coefficient of Chèzy formula

$c =$ coefficient such that the thickness of hyper-turbulent region is proportional to d_S in Morris (1955)

$D =$ internal diameter of the pipe in wake-interference flow

$d_S =$ longitudinal spacing of the wall-roughness elements

$f =$ resistance coefficient of Darcy–Weisbach formula

$g =$ gravitational acceleration

$H =$ flow depth in wake-interference flow

$h_f =$ frictional loss

$h_S =$	height of the wall-roughness elements
$L =$	length of the pipe
$n =$	resistance coefficient of Manning formula
$k =$	Von Kàrmàn universal constant of turbulence
$k_S =$	equivalent roughness height
$\mathbf{K} =$	relative roughness
$Q =$	discharge of flow
$R =$	hydraulic radius of flow
$\mathbf{R} =$	Reynolds number of flow
$\mathbf{R_W} =$	Reynolds wall number of flow
$S =$	slope of the channel
$V =$	average velocity of flow
$u =$	velocity at any point of the cross section of flow
$u^* =$	shear velocity of flow
$y =$	radial distance from the crests of the roughness
$y_0 =$	thickness of the hyper-turbulent region
$\varphi =$	additive element function in hyper-turbulent flow in Morris (1955, 1959)
$\varkappa_a =$	resistance function in Morris (1955)
$\psi =$	slope of logarithmic velocity distribution in the hyper-turbulent region in Morris (1955).

REFERENCES

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

Babovic, V. & Abbott, M. B. 1997. The evolution of equations from hydraulic data, Part I: Theory. *J. Hydraulic Res.* **35** (3), 1–14.

Babovic, V. & Keijzer, M. 2000. Genetic programming as a model induction engine. *J. Hydroinform.* **2** (1), 35–61

Babovic, V., Keijzer, M., Aguilera, D. R. & Harrington, J. 2001. *An Evolutionary Approach to Knowledge Induction: Genetic Programming in Hydraulic Engineering. Proceedings of the World Water & Environmental Resources Congress, May 21–24, Orlando, Fl. USA*. ASCE. Available for download from *www.d2k.dk*

Chow, V. T. 1973. *Open Channel Hydraulics*. McGraw-Hill, New York.

Colebrook, C. F. 1939. Turbulent flow in pipes with particular reference to the transition region between the smooth and the rough pipe laws. *J. Inst. Civil Eng., (London)* **11**, 133–156.

Colebrook, C. F. & White, C. M. 1937. Experiments with fluid friction in roughened pipes. *Proc. R. Soc. London, Ser. A* **161**, 367–381.

Davidson, J. W., Savic, D. & Walters, G. A. 1999. Method for the identification of explicit polynomial formulae for the friction in turbulent pipe flow. *J. Hydroinform.* **1** (2), 115–126.

Ead, S. A., Rajaratnam N., Katopodis, C. & Ade, F. 2000. Turbulent open-channel flow in circular corrugated culverts. *J. Hydraulic Engng.* **126** (10), 750–757.

French, H. R. 1983. *Open Channel Hydraulics*. McGraw-Hill, New York.

Giustolisi, O. 2001. Experimental study of artificially roughened sewer. *29th IAHR Congress, (Beijing), China, 17–21 September*. Theme D, Vol. I, pp. 301–309. Tsinghua University Press, Beijing, China.

Giustolisi, O. 2002. Some techniques to avoid overfitting of Artificial Neural Networks. *Hydroinformatics 2002*, IWA Publishing, London. pp. 1465–1477.

Keijzer, M. & Babovic, V. 1999. Dimensionally aware genetic programming. In: *GECCO-99: Proceedings of the Genetic and Evolutionary Computation Conference, July 13–17* (ed. Banzhaf, W., Daida, J., Eiben, A. E., Garzon, M. H., Honavar, V., Jakiela, M. & Smith, R. E.). Morgan Kaufmann Publishers. Available for download from *www.d2k.dk*

Keijzer, M. & Babovic, V. 2002. Declarative and preferential bias in GP-based scientific discovery. *Genetic Program. Evolvable Machines* **3** (1), 41–79.

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

Marchi, E. & Rubatta, A. 1981. *Meccanica dei fluidi, principi ed applicazioni idrauliche,* UTET, Torino, Italy.

Marone, V. 1970. Le resistenze al movimento uniforme in un alveo chiuso o aperto di sezione rettangolare e scabrezza definita. *L'Energia Elettrica* **1**, 1–20.

Morris, H. M. Jr. 1955. Flow in rough conduits. *Trans. ASCE* **120**, 373–398.

Morris, H. M. Jr. 1959 Design methods for flow in rough conduits. *J. Hydraulic Div., Proc. ASCE* **85** (HY 7), 43–62.

Perry, A. E., Schofield, W. H. & Joubert, P.N. 1969. Rough wall turbulent boundary layers. *J. Fluid Mech.* **37** (2), 383–413.

Powell, R. W. 1944. Flow in channel of definite roughness. *Trans. ASCE* **109**, 531–566.

Pyle, R. & Novak, P. 1981. Coefficient of friction in conduits with large roughness. *J. Hydraulic Res.* **19** (2), 119–140.

Rand, W. 1955. Discussion on flow in rough conduits. *Trans. ASCE* **120**, 400–405.

Shipton, R. J. & Graze, H. R. 1976. Flow in corrugated pipes. *J. Hydraulic Div. Proc ASCE* **102** (HY 9), 1343–1351.

Streeter, V. L. 1936. Frictional resistance in artificially roughened pipes. *Trans. ASCE* **101**, 681–704.

Tikhonov, A. N. 1963. Solution of incorrectly formulated problems and the regularization method. *Dokl. Akad. Nauk SSSR* **151**, 501–504.

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*. Springer, New York.

Yen, B. C. 2002. Open channel flow resistance. *J. Hydraulic Engng.* **128** (1), 20–39.