International Conference on Hydroinformatics

2014

A New Methodology For Deriving Regional Time Of Concentration Equations Using GIS And Genetic Programming

Soroosh Sharifi

Mahdi Razaz

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

A NEW METHODOLOGY FOR DERIVING REGIONAL TIME OF CONCENTRATION EQUATIONS USING GIS AND GENETIC PROGRAMMING

SOROOSH SHARIFI (1), MAHDI RAZAZ (2)
(1): School of Civil Engineering, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom
(2): Dept. of Civil and Environmental Engineering, Hiroshima University, 1-4-1 Kagmiyama, Higashi-Hiroshima 739-8527, Japan.

ABSTRACT

In this study, a methodology is proposed for deriving Time of concentration (ToC) equations for watersheds located in a specific geographic region using GIS and Genetic Programming (GP). In this method, “true” ToC values are calculated by integrating GIS data into the TR-55 model and using the travel time method. GP is then used as a data-mining tool for conducting symbolic regression and deriving the most accurate equations for the region’s watersheds. In a case study, the proposed methodology is applied to 72 watersheds and sub-watersheds in Khorasan Razavi province, Iran. The method provides a set of different ToC equations to be used for watersheds in the region. Performance evaluation of the equations mined by GP shows that this approach is able to find ToC equations that are more accurate and robust compared to conventional ToC equations. Also, the derived equations shed some light on the important parameters that influence the ToC of a watershed.

INTRODUCTION

Time of concentration (ToC) reflects the speed at which a watershed responds to rainfall events. ToC is the most frequently utilized time parameter [1] and is of great importance in many hydrological analyses including, inter alia, the design of urban storm water drainage systems using the rational method and predicting the peak discharge using various rainfall-runoff models. Accurate estimates of the ToC are important; if ToC is under-estimated, the result is an over-estimated peak discharge and vice versa [2].

The study of the literature reveals the existence of numerous methods and equations for estimating ToC. These equations are generally empirical regression equations which have been obtained by finding a relationship between measured or estimated ToC values and the other at-hand watershed and rainfall parameters. The applicability of these equations is constrained by the lack of diversity in the data used for their development, and hence, they are not transferable to watersheds located at other regions. As a result, modelers are often baffled by the vast number of ToC equations, and often select an equation without evaluating its accuracy [1, 3].

A few studies [1, 4, 5, 6] have focused on evaluating empirical ToC equations for use in watersheds located at different regions. All studies conclude that a great deal of discrepancy exists between the estimation of different empirical equations, and one must not use an equation without evaluating its performance against observed ToC values. Sharifi and Hosseini [6] found that simple correction factors could be found for many of the conventional equations, which, if applied, will result in enhanced accuracy of ToC estimations.
The limitations and uncertainties associated with conventional equations, which are generally popular for their limited number of input parameters, necessitates the need for developing accurate and robust region specific ToC equations. In this study, a methodology is proposed for deriving accurate ToC estimation method(s) for application in a particular region by using GIS data and Genetic Programming. In a case study, the methodology is applied to a range of watershed conditions in the Khorasan Razavi province of Iran and a set of ToC equations are derived.

METHOD

In summary, the proposed method works as follows: first, “true” or “reference” ToC values are derived using the velocity method incorporated in the TR-55 model. Second, a database is formed from various rainfall and watershed characteristics and measures which are derived automatically by a GIS software. Finally, Genetic Programming (GP) is used as a data-mining tool to search the database and perform symbolic regression to find accurate and robust regional ToC estimation equations.

Obtaining “true” ToC Values

In this methodology, the TR-55 model [7] incorporated in the Watershed Modeling System (WMS) software was employed to obtain the “true” ToC values using the velocity method [7]. The velocity method is a distributed approach where the ToC is derived by sub-dividing the longest flow path into different segments each corresponding to a different flow regime (i.e. sheet flow, shallow concentrated flow and open channel flow regimes). Separate equations are used for calculating the travel time in each of the segments and the ToC is obtained by summing the flow time of all segments. Being based on hydraulics estimates, this method is regarded as one of the most accurate ToC estimation methods. However, its large number of difficult to measure required inputs, such as channel geometry and Manning’s roughness coefficient, makes this method difficult to apply. Employing various GIS tools, the WMS software has the ability to process spatial data and automatically measure various watershed parameters, and hence, can greatly improve the speed and accuracy of the travel times resulting in accurate ToC estimations. Also, the software is able to derive various watershed characteristics and measures (Table 1), and provide a comprehensive database. The details of the approach are further described by Sharifi and Hosseini [6] and Green and Nelson [8].

Genetic Programming

Genetic programming (GP) was first introduced by Koza [9], as a powerful evolutionary computation tool for tackling problems in various fields of artificial intelligence. Like other evolutionary computation methods, GP is based on the principle of Darwin’s theory of evolution. Standard GP starts with an initial population of randomly generated symbolic expressions (also known as pars trees) composed of functions (e.g. arithmetic operations, mathematical functions and standard programming operations) and terminals (e.g. constants, conversion factors) appropriate to the problem domain. A fitness function is then used to measure the performance of each individual symbolic expression in the particular problem environment. Then, a sexual genetic reproduction process is performed on pairs of expressions, which are selected in proportion to their fitness, and offsprings are created. The resulting offsprings are composed of sub-expressions from their parents and form the new generation, which replaces the old population of parents. The fitness function is again used to measure the fitness of each individual in the new population and the process is repeated. Repeating this algorithm will gradually produce populations, which, over a period of generations, reach a high average fitness in dealing with their environment. For an in-depth explanation of GP and its elements the reader is referred to Koza [9].

Table 1. Watershed characteristics and measures derived by the WMS software

N.	**Definition**	**Unit**
A	Area	(m^2)
BC	Basin circularity ratio (area/area of circle with perimeter equal to that of the basin)	-
BL	Basin length: Parallel measure along stream from mouth to boundary	(m)
CN	SCS curve number	-
CR	Compactness ratio (P/circumference of a circle of equal area	-
D	Diameter of a circle having the area equal to the watershed	(m)
DD	Drainage density (channel length/area)	(m^{-1})
ΔH	Elevation difference between start and endpoint of the main channel	(m)
E1	Elevation of channel starting point	(m)
E2	Elevation of watershed outlet	(m)
ER	Elongation ratio	-
H	Average watershed elevation	(m)
L	Basin length among main channel from outlet to upstream boundary	(m)
L_c	Length of main channel	(m)
L_{ca}	Length measured from the concentration point along L to a point on L that is perpendicular to the watershed centroid	(m)
L_o	Average overland flow length	(m)
N	Manning’s roughness coefficient	(%)
P	Watershed perimeter	(m)
RR	Relief ratio (ΔH/BL)	-
S_c	average slope of main channel	(m/m)
S_{ca}	Slope along main channel from outlet to point opposite centroid	(m/m)
Sh1	Shape Factor 1 (watershed length/watershed width)	-
Sh2	Shape Factor 2 (watershed area/main channel length^2)	-
SL	Basin slope among main channel from outlet to upstream boundary	(m/m)
So	Average slope of watershed	(m/m)

CASE STUDY

Study Area

In a case study, the proposed methodology was applied to 72 watersheds of different sizes and characteristics in the Khorasan Razavi province, in the eastern part of Iran (Figure 2). Watershed and rainfall data, including land use, soil type, elevation data, and records of several climate factors, were obtained through personal contact from a number of sources including several water and wastewater companies. After careful examination of available datasets, five main watersheds in the province were selected (Table 2). Also several sub-watersheds within the selected larger watersheds were identified and used in the study, in order to have a wider range of watershed conditions within the chosen geographic region. For more information on the watersheds' main parameters the reader is referred to [6].

Figure 1. Location of the selected watersheds in the Khorasan Razavi province of Iran [6]

Table 2. Summary of the main watersheds' general characteristics [6]

Watershed	Torogh	Kameh	Baig	Roshtkhar	Sheshtamad
A (km^2)	376.7	56.5	33.8	57.7	18.8
S (m/m)	0.044	0.059	0.052	0.087	0.091
L (m)	35309	14407	11609	14397	15335
ΔH	1530.1	834.5	556	1203.1	1381.2
Precipitation station	Torogh dam	Torbat Heidarieh	Torbat Heidarieh	Roshtkhar	Senobar
Latitude	36° 10' N	35° 18' N	35° 18' N	34° 59' N	35° 57' N
Longitude	59° 33' E	59° 13' E	59° 13' E	59° 38' E	57° 46' E

Applying the Methodology

First, the method for obtaining ToC values using the TR-55 model and the NRCS velocity method was applied, and "true" ToC values were obtained for the 72 selected watersheds and sub-watersheds. Then, WMS was used to derive several characteristics and measures (Table 1) of all the watersheds and a database was formed. By means of uniform random sampling, the database was split into three disjoint subsets: training (66%), testing (23%) and validation data (11%). The training data was used as inputs for the GP modeling process, the testing data for model selection and the validation data for model validation. The Genetic programming lab (GPLAB) v.3 toolbox for Matlab [10] was used with slight modifications in the terminal set and fitness function to evolve a relationship between the ToC, and the measured watershed characteristics and computed watershed measures. A sensitivity analysis was first performed in order to obtain a robust algorithm parameter set (Table 3). Then, 250 independent runs of the GP algorithm were performed to limit the effect of randomness on the results.

Results

A weighted average of three measures of fitness, namely, the mean root of sum of squared residuals (MRSS), root mean square of error (RMSE) and coefficient of determination (CoD) for the three data sets were used to rank all the GP results. Table (4) shows the top 4 best performing ToC expressions along with the performance of three conventional ToC equations (California, Kirpich and A.DOT) for the same datasets. Figure 2 illustrates the plot of calculated vs. true ToC values and the residual distributions using two selected equations for the training set.

From Table (4) it is observed that although conventional methods have a slightly higher CoD, their MRSS and RMSE values are significantly higher than the four derived expressions in all datasets. An investigation of residual distributions also shows that the California and Kirpich equations, which each require 2 inputs, drastically underestimate the ToC while the A.DOT method provides much better ToC estimates but requires 4 inputs. On the other hand, the performances of the top 4 expressions are similar for all the datasets, and all clearly outperform the conventional equations (i.e. have much lower MRSS and RMSE values).

The first obtained expression has a more complex structure and requires 4 input parameters but has a slightly better performance especially for the "unseen" validation set. On the other hand, expressions 3 and 4 have a relatively simple structure and only require 2 inputs. As shown in Figure (2) expression 1 tends to underestimate larger ToC values while expression 3 shows no bias in predicting the diverse range of ToCs. Furthermore, the trend line drawn between true and calculated ToC values has almost no deviation from the 45-degree line, which is another indicator of equality in the predicted ToC distribution.

Table 3. GP operators and parameters.

Parameter	**Value**
Function set	Plus, minus, times, power, exp, mydivide, mylog, mysqrt
Terminal set	All variables listed in Table 1 + 3 random numbers
Population size	75
Tree Initialization method	Ramped-half-and-half
Tree size restriction	10 nodes
Genetic operators	Subtree Cross-over and Mutation
Operator probabilities	Variable (minimum equal to 0.20)
Fitness function	Sum of squared distance
Selection method	Lexictour
Number of generations	250

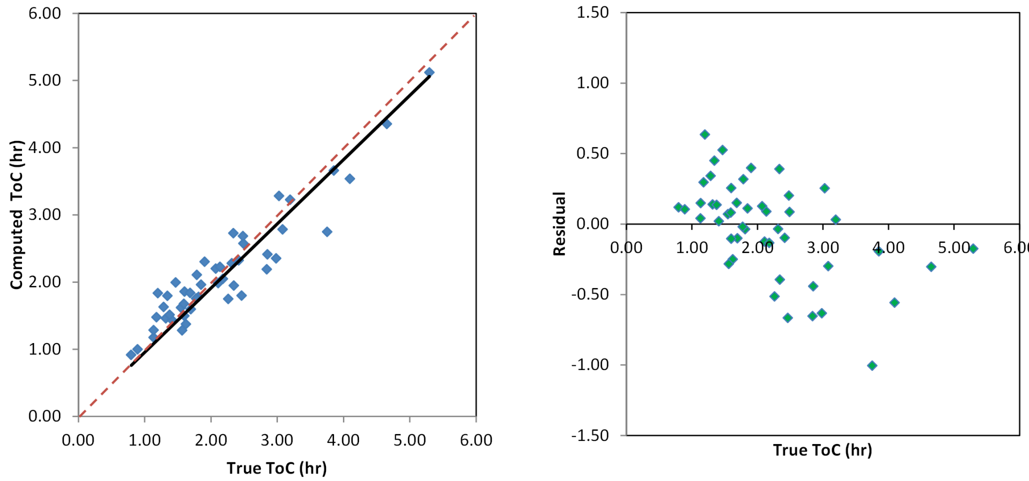
Table 4. Performance of conventional ToC equations and GP derived expressions

No.	Expression	Training			Testing			Validation		
		MRSS	**RMSE**	**CoD**	**MRSS**	**RMSE**	**CoD**	**MRSS**	**RMSE**	**CoD**
	California	0.146	0.417	0.864	0.308	0.429	0.945	0.326	0.399	0.916
	Kirpich	0.147	0.436	0.860	0.307	0.436	0.955	0.329	0.406	0.914
	A.DOT	0.057	0.170	0.869	0.249	0.249	0.948	0.133	0.182	0.923
1	$T_c = \frac{D}{ER + D \times RR - DD + 1.94}$	0.049	0.168	0.883	0.115	0.288	0.929	0.122	0.204	0.959
2	$T_c = 0.34 \times (D + DD \times D^{0.5DD}) + 0.32^D$	0.057	0.190	0.837	0.147	0.352	0.887	0.202	0.190	0.902
3	$T_c = 0.39\sqrt{A} + DD^2$	0.057	0.192	0.835	0.150	0.345	0.886	0.202	0.334	0.904
4	$T_c = (D + DD - 2.11)^{0.61}$	0.056	0.219	0.840	0.120	0.386	0.910	0.214	0.379	0.836

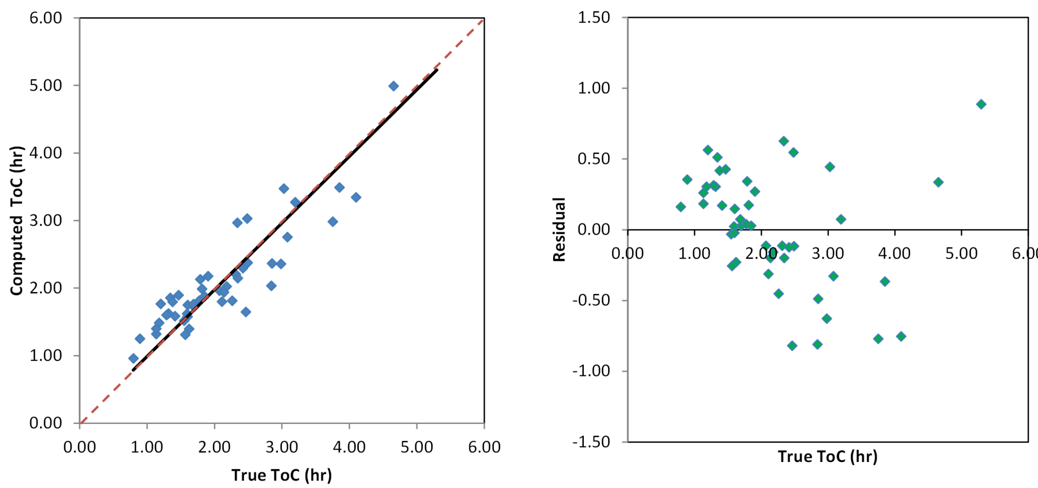
CONCLUSIONS

In the present study, GIS and Genetic Programming have been combined to develop a powerful and efficient tool for deriving equations for the most important watershed time parameter: the time of concentration. In a case study, the proposed method was used to derive a number of equations for estimating the ToC of watersheds in the specific region. The obtained expressions not only outperform the conventional empirical ToC equations, but also generally require fewer inputs. A comparison among the highest ranked expressions revealed that an expression in the form of $ToC = 0.39A^{0.5} + DD^2$ appears to be the simplest and most suitable equation for estimating ToC in the study region. The inputs to this equation are the watershed's area (A) and drainage density (channel length/area), which are generally available and easy to measure. An investigation of all GP simulations (not reported here) identifies the watershed's area (A), drainage density (DD), elongation ratio (ER) and average channel slope (So) as the most influential parameters affecting the ToC. Further work is required to find the probable physics behind this expression and also interpret its coefficients. Additional work is also required to verify the accuracy of this model structure for watersheds of other regions.

Figure 2. Performance of a) expression 1; and b) expression 3 on the training dataset



a) $T_c = \dfrac{D}{ER + D \times RR - DD + 1.94}$



b) $T_c = 0.39\sqrt{A} + DD^2$

REFERENCES

[1] McCuen, R.H., Wong, S.L. and Rawls, W.J., "Estimating Urban Time of Concentration", *Journal of Hydraulic Engineering*, ASCE, Vol. 110, No. 7, (1984), pp 887-904.

[2] Fang, X., Thompson, D.B., Cleveland, T.G., Pradhan, P., et al., "Time of concentration estimated using watershed parameters determined by automated and manual methods", *Journal of Irrigation and Drainage Engineering*, ASCE, Vol. 134, No 2, (2008), pp 202-211.

[3] Wong, T.S.W., "Evolution of Kinematic Wave Time of Concentration Formulas for Overland Flow", *Journal of Hydrologic Engineering*, Vol. 14, No. 7, (2009), pp 739-744.

[4] McCuen, R.H. and Spiess, J.M., "Assessment of kinematic wave time of concentration", *Journal of Hydraulic Engineering*, ASCE, Vol. 121, No. 3, (1995) pp 256-266.
[5] Wong, T.S.W. "Assessment of time of concentration formulas for overland flow", *Journal of Irrigation and Drainage Engineering*, ASCE, Vol. 131, No. 4, (2005), pp 383-387.
[6] Sharifi, S. and Hosseini, S.M. "Methodology for identifying the best equations for estimating the time of concentration of watersheds in a particular region." *Journal of irrigation and drainage engineering*, Vol. 137, No. 11, (2011), pp712-719.
[7] Natural Resource Conservation Service (NRCS), "Urban hydrology for small watersheds", Soil Conservation Service, U.S. Department of Agriculture, Tech. Release No. 20, Washington, D.C. (1986).
[8] Green, J.I. and Nelson, E.J. "Calculation of Time of Concentration for Hydrologic Design and Analysis Using Geographic Information system vector Objects", *Journal of Hydroinformatics*, Vol. 4. No. 2, (2002), pp 75-81.
[9] Koza, J.R. "Genetic Programming: On the Programming of Computers by Means of Natural Selection". MIT Press, Cambridge, MA, USA. (1992).
[10] World Wide Web: Genetic Programming Lab (GPLab, 2007) - ttp://gplab.sourceforge.net/ [Accessed Feb 1st 2014].