International Journal of Earth, Energy and Environmental Sciences
ISSN: 2517-942X
Vol:9, No:2, 2015

Predicting Residence Time of Pollutants in Transient Storage Zones of Rivers by Genetic Programming

Rajeev R. Sahay

***Abstract*—**Rivers have transient storage or dead zones where injected pollutants or solutes are entrapped for considerable period of time, known as residence time, before being released into the main flowing zones of rivers. In this study, a new empirical expression for residence time, implementing genetic programming on published dispersion data, has been derived. The proposed expression uses few hydraulic and geometric characteristics of rivers which are normally known to the authorities. When compared with some reported expressions, based on various statistical indices, it can be concluded that the proposed expression predicts the residence time of pollutants in natural rivers more accurately.

***Keywords*—**Parameter estimation, pollutant transport, residence time, rivers, transient storage.

I. Introduction

POLLUTANT transport in rivers is largely influenced by surface hydrodynamics and mass exchanges between the river's main flowing zones and transient storage zones. Transient storage zones, also called retention or dead zones are formed by irregular bed forms, pools and riffles, and interstitial sediment voids within the river beds. Some parts of accidental spills of pollutant get entrapped in these retention domains for considerable period of time before being released to the free flowing river zone, thus reducing the effective advection velocity in the longitudinal direction of the stream. Sorbing solutes experience longer residence time than nonsorbing solutes. Knowledge of dispersion is important to hydraulic and environmental engineers for ascertaining quality of river water, managing water resources schemes and water rights, and designing outfalls. Although, the importance of transient storage zones in pollutant dispersion is well recognized, however, relatively few studies have quantified the residence time of pollutants in these zone. The transient storage zone model (TSM) is reported to be a better approach to include the effect of transient storage in longitudinal solute transport in rivers than the classical first order advection-dispersion model [1]. TSM incorporates following two mass conservation equations, one for the solute concentration dissipation in the free-flowing water zone and another in the transient storage zone

$$\frac{\partial C_f}{\partial t} + U_f \frac{\partial C_f}{\partial x} - K_f \frac{\partial^2 C_f}{\partial x^2} = \varepsilon T^{-1}\left(C_s - C_f\right) \tag{1}$$

Rajeev R. Sahay is with Birla Institute of Technology (Mesra) Patna Campus, Patna, India (phone: +919431382737; e-mail: rajeev_sahay@yahoo.com).

$$\frac{\partial C_s}{\partial t} = T^{-1}\left(C_f - C_s\right) \tag{2}$$

where C_f is the pollutant concentration in the free flowing water zone; C_s is the transient storage zone solute concentration; t is the time elapsed since injection of the solute/pollutant; x is the longitudinal distance from the place of injection of the pollutant to the investigation site; U_f is the mean flow velocity in the free flowing water zone; K_f is the longitudinal dispersion coefficient in the free-flowing water zone; ε is ratio of cross-sectional area of the transient storage to the total cross-sectional area of the channel and T is the residence time of the tracer in the transient storage zone.

Appropriate values of TSM parameters i.e., U_f, K_f, T and ε, are required for model's successful implementation. These parameters can best be estimated using tracer concentration profile taken from a particular reach of the stream, but such tracer investigation is expensive and rarely done for every reach of a stream. For this reason, many investigators have derived empirical expressions for estimation of these parameters. However, the present study by the author shows the reported expressions for the residence time to be inadequate with large deviations between measured and predicted values.

Reference [2], after fitting experimental dispersion data compiled by [3] on the solution of [4], derived the following simple expression for residence time T of solutes in transient storage zones of rivers as

$$T = \frac{5 f W^2}{H U_*} \tag{3}$$

where f is the Darcy-Weisbach friction factor, W is the stream width, H is the mean flow depth and U_* is the shear velocity.

Reference [5], employing the weighted one-step Huber nonlinear multi-regression method on published field data, derived the following expression in non-dimensional format as

$$\frac{T}{H/U_*} = 56.68 \left(\frac{W}{H}\right)^{0.767} \left(\frac{U}{U_*}\right)^{-0.884} \tag{4}$$

Reference [6] developed robust minimum covariance determinant method and applied it on hydraulic and geometric information available for natural rivers to derive the following empirical expression for prediction of residence time T and showed its superiority over other reported expressions of the time

$$\frac{T}{H/U_*} = 20.595 \left(\frac{W}{H}\right)^{0.6639} \left(\frac{U}{U_*}\right)^{-1.4625} P_e^{0.3232} S_i^{1.9132} \quad (5)$$

where P_e is the Peclet Number and S_i is the channel sinuosity. Reference [7] satisfactorily estimated parameters of the transient storage model using neural networks, however, direct expressions are handy and more suitable for predicting these parameters.

II. Derivation of New Expression for T by Genetic Programming

The Performance of the above-mentioned expressions, as would be shown later in this article, is far from satisfactory. The claim of [5] and [6] about accuracy of their expressions are found to be rather misplaced. In the present work, an alternative expression for the residence time of solute/pollutant in dead zones of rivers is advanced. This has been done implementing genetic programming on the published dispersion data of USA Rivers [Table I]. Comparable datasets were chosen for deriving and verifying the new expressions to avoid any bias in modeling. There are two reasons for selecting this data: (1) it represents a wide range of geometric and flow characteristics of streams and (2) it had earlier been used by [5] and [6]. Thus, results from the proposed and other reported expressions could be compared well.

A. Genetic Programming

Genetic Programming invented by [8] and further developed by [9], is a biologically inspired machine learning method that evolves computer programs to perform a task. It does this by randomly generating a population of computer programs (represented by tree structures) and then mutating and crossing over the best performing trees to create a new population. This process is iterated until the population contains programs that solve the task well [10]. Though GP does not use chromosomes, it works on the principle of genetic algorithm (GA) as both employ populations of individuals, select them according to fitness values and introduce variation in population for next generation using genetic operators. The only important difference between GP and GA lie in the nature of individuals and in the way they are reproduced to allow adaptation. A basic flow-chart for the GP system is given in Fig. 1.

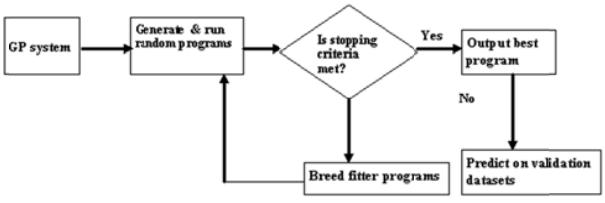

Fig. 1 Working structure of GP

In the present work, GPTIPS ([10]), an open source GP program was applied on published dispersion data (Table I) with the following settings: population size = 300, number of generation = 250, tournament size = 4, elitism = 0.02% of population, maximum depth of tree = 3, maximum number of genes allowed in an individual tree = 4 (W/H, U/U*, P_e and S_i) and function node set = {plus, minus, times, tanh}. The above setting was used to minimize the root mean square prediction error on the derivation datasets. The model that performed the best on the verification datasets was chosen. Accordingly, the best expression for dimensionless parameter $\frac{T}{H/U_*}$ was derived as

$$\frac{T}{H/U_*} = 82.69 + \frac{294.3\tanh(P_e)\left(\frac{W}{H}-10.19\right)}{\frac{U}{U_*}} - \frac{0.16\left(\frac{W}{H}+S_i\right)^2}{(P_e-2.19)^{0.5}} + 0.111\frac{\left(\frac{W}{H}+\frac{U}{U_*}\right)\left(\frac{W}{H}+8.83\right)}{\left(\frac{U}{U_*}-2.16\right)^{0.5}} \quad (6)$$

The above expression appears to have successfully been derived, with coefficient of correlation between measured and predicted values being equal to 0.89. The expression successfully predicted the highest and the lowest $\frac{T}{H/U_*}$ values of 2235.4 and 16.7 respectively as 2104.2 and 17.2 respectively, and most of the predicted values are evenly distributed about the ideal line, showing no bias for over or under prediction (Fig. 2).

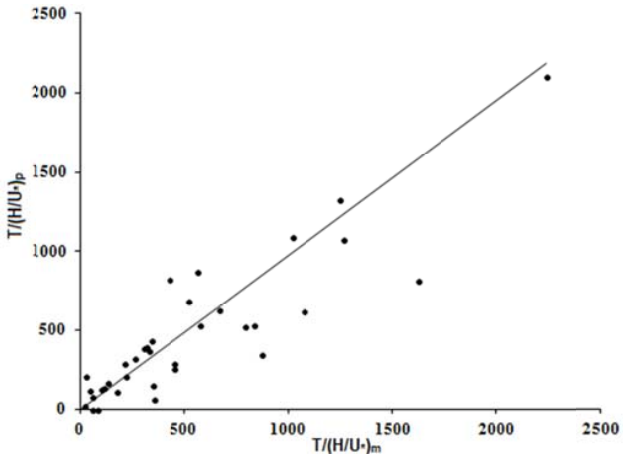

Fig. 2 Measured and predicted dimensionless residence time (derivation dataset)

III. Verification of the New Expression

The comparison of the new and other reported expressions for residence time was accomplished using 20 measured datasets of Table I that were not used for deriving the new expression. The comparison models used here are [2], [5] and [6]. For brevity, these comparison models are denoted as Pedersen, C-S, and C-Y-S, respectively. The performance indices used for comparison of the models are coefficient of correlation (CC), root mean square error (RMSE), discrepancy ratio (DR) and accuracy. They are defined as

$$CC = \frac{\left[\sum_{i=1}^{N}\left(\frac{T}{H/U_*}\right)_p\left(\frac{T}{H/U_*}\right)_m - \sum_{i=1}^{N}\left(\frac{T}{H/U_*}\right)_p \sum_{i=1}^{N}\left(\frac{T}{H/U_*}\right)_m\right]}{N S_p S_m} \quad (7)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left[\left(\frac{T}{H/U_*}\right)_p - \left(\frac{T}{H/U_*}\right)_m\right]^2}{N}} \quad (8)$$

$$\text{DR} = \log\frac{\left(\frac{T}{H/U_*}\right)_p}{\left(\frac{T}{H/U_*}\right)_m} \quad (9)$$

$$\text{Accuracy} = 1 \text{ (accurate, if DR lies between -0.3 to 0.3), otherwise 0.} \quad (10)$$

where $\left(\frac{T}{H/U_*}\right)_p$ and $\left(\frac{T}{H/U_*}\right)_m$ are predicted and measured dimensionless residence time parameters, respectively and S_p, S_m are standard deviations of predicted and measured values, respectively.

Fig. 3 shows comparison of the predicted values and the measured values for the verification datasets. As can be observed from this figure, the maximum number of predicted values by the new expression is closer to the measured values. The highest 3 values of the parameter of 2066.9, 1734.4 and 1269.2 are successfully predicted as 2024.3, 1357.5 and 766.6 whereas; deviations are larger in case of other expressions. Table II summarizes performance indices of the models for the verification data. Prediction of parameter $\left(\frac{T}{H/U_*}\right)$ by the newly derived expression is more realistic in comparison to all other models, the predictions' RMSE being the smallest and CC, the largest. The performance of Pedersen model is seen to be the least satisfactory.

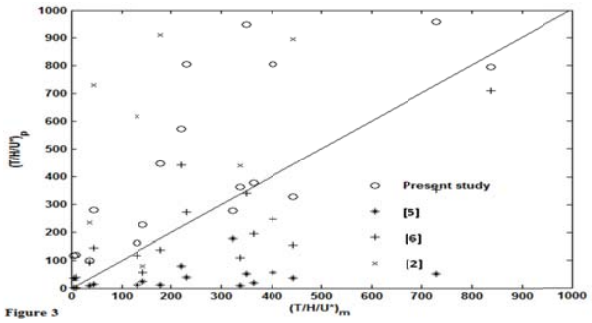

Fig. 3 Measured and predicted dimensionless residence time (verification dataset)

If accuracy of a model can be defined as the percentage of predicted values lying between 50% and 200% of the measured values, i.e., discrepancy ratios falling between -0.3 and 0.3, then it can be observed from Table II that the proposed expression has 55 % accuracy, the highest among the models compared. Accuracies of C-Y-S, Pedersen and C-S models are 5%, 35%, and 15% respectively (Fig. 4).

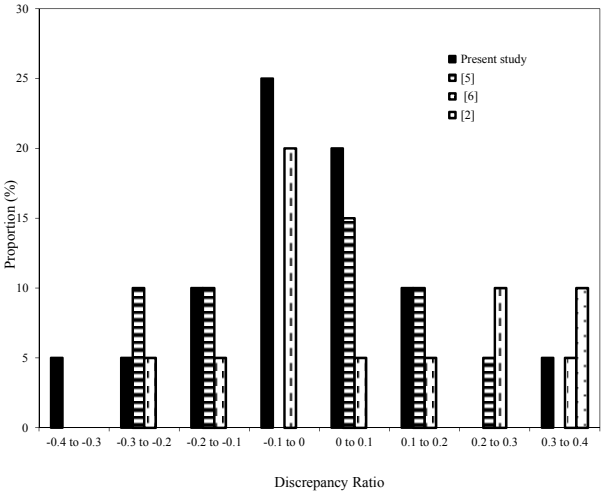

Fig. 4 Comparison of discrepancy ratios of models (verification dataset)

IV. Conclusion

A pollutant's dispersion in rivers is greatly influenced by the presence of transient storage zones formed largely at the beds of rivers where it might be retained for considerable period of time before being released slowly into the main flowing zone of rivers. The available empirical expressions for estimating the residence time in transient zones are evaluated and found inadequate. In the present work, implementing genetic programming on published field data, a new empirical expression for residence time of pollutant/solute in transient storage zone of rivers has been derived. The proposed expression uses few hydraulic and geometric characteristics of a river, i.e., stream width, stream sinuosity, mean flow velocity, mean flow depth, shear velocity and Peclet number. The performance of new expression is compared with those of [2], [5] and [6]. Based on various performance indices, i.e., coefficient of correlation, root mean square error, discrepancy ratio and accuracy, the new expression is found superior to other considered models.

TABLE I
HYDRAULIC DATA OF TRANSIENT STORAGE MODEL OBSERVED AT 55 SITES OF NATURAL STREAMS IN USA [6]

River	W(m)	H(m)	U(m/s)	S	Pe	Si	T(s)
Green & Duwamish, WA*	21.77	1.58	0.31	0.00022	0.08	1.39	1032.43
Green & Duwamish, WA	29.61	1.08	0.36	0.00022	2.22	1.39	386.82
Copper creek, VA	16.30	0.50	0.23	0.00134	0.05	1.23	2865.38
Copper creek, VA*	17.12	0.38	0.18	0.00373	0.09	1.23	2761.27
Copper creek, VA	18.23	0.83	0.12	0.00127	0.02	1.23	1086.15
Powel river, TN	34.46	0.84	0.09	0.00037	0.03	1.38	4643.32
Clinch river, VA*	47.62	1.03	0.19	0.00050	0.09	1.73	3264.22
Copper creek, VA	19.81	0.82	0.50	0.00130	0.07	1.23	886.21
Clinch river, VA	53.30	2.08	0.76	0.00061	0.10	1.73	851.482
Coachella canal, CA*	23.90	1.56	0.66	0.00011	0.58	1.04	206.85
Coachella canal, CA*	24.43	1.55	0.67	0.00011	0.58	1.04	355.67
Clinch river, VA	53.30	2.08	0.69	0.00061	0.10	1.73	1842.24
Copper creek, VA	16.00	0.50	0.20	0.00134	0.03	1.23	2865.38
Missouri river	180.80	3.28	1.25	0.00020	0.14	2.00	3265.31
Antiem creek, Md	12.20	0.37	0.21	0.00143	0.32	1.91	8566.31
Antiem creek, Md*	18.90	0.73	0.52	0.00136	0.33	1.91	3404.43
Antiem creek, Md	11.89	0.66	0.43	0.00143	0.56	1.91	1841.87
Antiem creek, Md*	23.09	0.45	0.41	0.00130	0.39	1.91	7684.56
Monocacy river, MD	48.77	0.55	0.26	0.00050	0.17	1.39	8870.40
Monocacy river, MD	92.96	0.71	0.16	0.00037	0.10	1.39	17524.90
Monocacy river, MD	46.13	0.80	0.32	0.00050	0.26	1.39	8612.81
Monocacy river, MD*	64.08	0.93	0.05	0.00037	0.01	1.39	5235.69
Conococheague creek, MD	46.28	0.52	0.22	0.00052	0.16	2.27	5261.21
Conococheague creek, MD*	40.88	0.43	0.10	0.00065	0.05	2.27	17150.40
Conococheague creek, MD	51.20	0.95	0.68	0.00063	0.38	2.27	3915.62
Chattahoochee river, GA	54.78	1.80	0.74	0.00045	0.50	1.23	18067.20
Chattahoochee river, GA	86.27	2.40	0.30	0.00030	0.17	1.23	9928.67
Salt creek, NE	31.30	0.34	0.18	0.00043	0.15	1.17	7174.80
Difficult Run, VA*	14.58	0.30	0.22	0.00127	0.30	1.38	3650.40
Bear creek, CO*	13.72	0.85	1.29	0.00173	4.36	2.04	1084.32
Little Pincy Creek, MD	15.85	0.22	0.39	0.00130	0.37	1.25	4502.11
Bayou Anacoco, LA*	17.50	0.45	0.23	0.00055	1.05	1.76	16241.60
Comite river, LA*	12.44	0.30	0.25	0.00057	0.26	1.35	2726.12
Tickfau river, LA	21.58	0.68	0.07	0.00093	0.30	1.2	19898.20
Tangipahoe river, LA	32.14	0.96	0.26	0.00051	0.49	1.29	7982.50
Tangipahoe river, LA	29.73	0.55	0.36	0.00058	0.70	1.29	12660.10
Red river*	205.10	3.20	0.31	0.00008	0.40	1.59	14923.20
Red river*	185.50	2.65	0.29	0.00012	0.34	1.59	16826.40
Red river	152.40	3.66	0.45	0.00012	0.30	1.59	18461.80
Red river	200.40	1.75	0.33	0.00008	0.35	1.59	9770.81
Sabin river, LA	68.97	1.27	0.65	0.00016	0.36	1.61	9867.10
Sabin river, LA*	158.40	2.26	0.98	0.00014	0.61	1.61	5434.40
Sabin river, TX	12.90	0.73	0.08	0.00018	0.40	1.52	12210.10
Sabin river, TX	14.88	0.75	0.04	0.00021	0.28	1.52	8482.26
Sabin river, TX*	33.98	1.20	0.13	0.00013	0.50	1.52	12754.90
Mississippi river, LA	735.30	18.10	0.52	0.00001	0.37	1.73	9479.20
Mississippi river, MO	569.80	4.77	0.99	0.00012	0.52	1.38	3312.90
Mississippi river, MO*	587.20	8.47	1.43	0.00012	0.91	1.38	15385.60
Wind/Bighorn river, WY	59.13	1.17	0.75	0.00135	0.70	1.15	9726.43
Wind/Bighorn river, WY*	72.85	2.40	1.58	0.00135	0.58	1.15	4677.66
Colorado river, AZ	106.10	6.10	0.79	0.00013	0.18	1.76	12376.40
Colorado river, AZ	71.60	8.20	1.20	0.00141	0.28	1.76	9385.88
Botna river*	2.05	0.10	0.22	0.00235	0.33	1.12	98.64
Kogilnik river	2.30	0.40	0.56	0.00211	0.05	1.31	262.82
Byk river	2.55	0.14	0.22	0.00084	0.40	1.23	932.21

*. Verification dataset

TABLE II
PERFORMANCE INDICES OF MODELS

Parameter	Model	Whole datasets		Derivation datasets		Verification datasets				
		CC	RMSE	CC	RMSE	CC	RMSE	CC (W/H>50 ignored)	DR Range	Accuracy (%)
$T/(H/U_*)$	Present	0.87	253.3	0.89	233.8	0.88	283.7	0.91	-0.22 to 1.36	55
	C-S	0.62	668.2	0.56	664.1	0.70	675.4	0.67	-0.26 to -1.60	5
	C-Y-S	0.45	509.8	0.59	472.3	0.38	569.4	0.46	-0.87 to -0.86	35
	Pedersen	0.24	38253.2	0.56	15325.6	0.19	60108.7	0.33	-0.30 to 2.90	15

Pedersen = [2], C-S = [5] and C-Y-S = [6]

REFERENCES

[1] K. E. Bencala and R. A. Walters, "Simulation of solute transport in a mountain pool-and-riffle stream", *Water Resources Research*, Vol. 19, pp. 718-724, 1983.
[2] F. B. Pedersen, *Prediction of longitudinal dispersion in natural streams*. Series Paper 14, Technical University of Denmark, Lyngby, 1977.
[3] C. F. Nordin and G. V. Sabol, *Empirical data on longitudinal dispersion*. US Geol Survey Water Resour Invest Report, 1974, pp. 20–74.
[4] A. Okubo, "Effect of shoreline irregularities on stream wise dispersion in estuaries and other embayments", *Netherlands J of Sea Research*, Vol. 6, pp. 213–224, 1973.
[5] T. S. Cheong and I. W. Seo, "Parameter estimation of the transient storage model by a routing method for river mixing processes", *Water Resources Research*, Vol. 39, pp.1074–1084, 2003.
[6] T. S. Cheong, B. A. Younis and I. W. Seo, "Estimation of key parameters in model for solute transport in rivers and streams", *Water Resources Management*, Vol. 21, pp. 1165–1186, 2007.
[7] P. M. Rowiński and A. Piotrowski, "Estimation of parameters of the transient storage model by means of multi-layer perceptron neural networks", *Hydrological Sciences Journal*, Vol. 53, pp. 165-178, 2008.
[8] Cramer NL (1985) A representation for the adaptive generation of simple sequential programs. In J. J. Grefenstette, ed., Proceedings of the first international conference on genetic algorithms and their applications Erlbaum 183-187.
[9] J. R. Koza, *Genetic Programming: On the programming of computers by means of natural selection*. Cambridge, MIT Press, 1992.
[10] D. P. Searson, D. E. Leahy and M. J. Willis, "GPTIPS: An open source genetic programming toolbox for multigene symbolic regression", in *Proc. Inter. Multi-conference of Engineers and Computer Scientists*, Hong Kong. 2010.