Information Technology and Management Science

doi: 10.2478/itms-2013-0021

2013 / 16

Regression-based Daugava River Flood Forecasting and Monitoring

Vitaly Bolshakov, *Riga Technical University*

***Abstract* – The paper discusses the application of linear and symbolic regression to forecast and monitor river floods. Main tasks of the research are to find an analytical model of river flow and to forecast it. The challenges are a small set of flow measurements and a small number of input factors. Genetic programming is used in the task of symbolic regression. To train the model, historical data of the Daugava River monitoring station near Daugavpils city are used. Several regression scenarios are discussed and compared. Models obtained by the methods discussed in the research show good results and applicability in predicting the river flow and forecasting of the floods.**

***Keywords* – Genetic programming, floods, forecasting, monitoring, regression**

I. Introduction

Early forecasting of river floods and prediction of areas to be flooded is an actual problem in territories located on banks of big rivers with regular or irregular flood behaviour. Solution to this problem allows preventing damages and possible losses in advance on inhabited or agricultural territories in risk areas. Essential part of forecasting is monitoring, which allows collecting data of river behaviour parameters in long time periods. In turn, the application of mathematical methods allows finding relations and patterns in this behaviour. Moreover, the monitoring of river physical parameters allows forecasting the river behaviour in the near future and correspondingly gives predictions of the flood.

In research [1], the modelling for the evaluation of aftermath of spring floods of the Daugava River is discussed with the estimation of the flooded areas. For this estimation a heightmap model of the investigated area is applied together with data from local river monitoring station. The main river monitoring data used for the flood estimation are river flow or river discharge, i.e., the volume of water flowing through the current cross-section of the river in a defined time interval. The problem is that the local station performs the evaluation of river flow rarely and irregularly. Nonetheless, a river water level is monitored regularly, which can be used for the estimation of unknown data.

In this study, the development of an analytical model of river flow depending on the current or recent river level is proposed for the determination of river flow discharge values, which are used in [1]. Application of linear and symbolic regression is applied to find such a model.

The problem statement is defined in Section II. Section III discusses regression methods proposed to find the flow analytical model. Section IV gives an overview on the problem input data. The application of linear and symbolic regression methods is discussed in Sections V and VI, correspondingly. Conclusions are summarised and given in Section VII.

II. Problem Statement

The goal of the research is to find how to calculate a river flow, which is based on the data of number of samples of river water level at the same river monitoring station. Data of the river level are dynamic.

The available input data of the stated problem are given in two tables:

1. A table of river water level values for each hour.
2. The data of the river discharge measurements.

Both data tables for the considered area are taken from an open data source [2]. The data describe the condition of the Daugava River at the monitoring station near Daugavpils city.

The data of river water level are given with precision of centimetre for each hour of each day in the analysed time interval (see Table I). The river level is estimated from a referencing level, thus, can have negative values.

TABLE I

Water Level Data

Daugavpils, Ūdens līmenis, stundas vidējais, cm

Datums \ Laiks	00:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00	08:00	09:00	10:00	11:00	12:00	13:00	14:0
01.01.2008	-26	-25	-25	-25	-25	-26	-25	-25	-25	-25	-26	-25	-24	-25	
02.01.2008	-14	-12	-11	-10	-8	-6	-4	-3	-2	-1	0	1	2	4	
03.01.2008	12	11	12	14	16	19	23	28	34	38	43	45	44	45	
04.01.2008	43	42	41	39	38	37	37	36	36	36	36	36	35	35	
05.01.2008	26	25	24	23	21	19	18	17	15	14	13	12	11	10	
06.01.2008	12	13	14	15	16	17	18	18	20	21	22	23	24	25	
07.01.2008	35	36	36	37	38	38	39	39	41	41	42	43	43	44	
08.01.2008	47	48	48	48	48	49	49	49	49	49	50	50	50	50	
09.01.2008	53	54	54	54	54	55	56	56	56	56	57	57	57	58	
10.01.2008	64	65	65	66	66	67	67	68	68	68	70	70	71	71	
11.01.2008	75	75	76	76	76	77	77	77	77	77	77	78	77	78	
12.01.2008	77	78	78	78	78	77	77	77	77	77	76	76	76	76	
13.01.2008	73	73	72	72	72	71	71	71	70	70	70	70	69	69	
14.01.2008	67	67	67	67	67	66	66	66	65	66	66	65	65	65	
15.01.2008	63	63	63	62	62	62	62	61	62	61	61	61	61	60	
16.01.2008	61	61	62	62	62	62	62	62	62	63	62	63	62	62	

In turn, the data table of river flow discharge (see Table II) contains dates, when the measurements are sampled, as well as measurement time and the determined water flow in m^3/s. Flow measurements are highly irregular: between some measurements there are intervals of several months, but other measurements are performed within one week interval.

TABLE II

Water Flow Measurements

Daugavpils, Caurplūdums, izmērītais, m3/s

Datums \ Laiks	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00
02.01.2008		182,32								
10.01.2008								197		
14.02.2008			310							
27.02.2008		408								
29.02.2008		525								
11.03.2008										103
28.03.2008				1153						
04.04.2008										130
25.04.2008										1180,
10.05.2008									477,67	
06.06.2008						673				
30.06.2008				141						
06.08.2008							125			
23.09.2008				290,98						
02.10.2008						203				
01.11.2008					382					

In the research it is proposed that the water discharge in the river is related to the water level. Thus, it is possible to determine the required water flow value if the current and recent river levels are known. Due to high availability of river level measurements, it is proposed to apply a number of recent level measurements, which are separated by a specific time interval to calculate the flow.

An analytical mathematical model in the closed form of algebraic expression, which will relate the water level to the forecasted water flow with a reliable precision, has to be found in the research.

The following solution steps and subtasks are planned and described in this research:

1. to analyse the input data to reveal the patterns and data incompleteness;
2. to process and prepare data for the following analysis methods;
3. to perform data statistical analysis;
4. to solve the regression task with the application of least squares method and to analyse the results;
5. to solve the symbolic regression task with the application of genetic programming;
6. to compare results and to make conclusions.

III. Solution Methods

A. Linear Regression

One of the most common and widely used approaches to find the relationships between one dependent variable and a number of explanatory variables is a linear regression. Linear regression implies that a dependent variable can be expressed in a form of linear equation from explanatory variables. The task is to find such coefficients of linear equation, which will fit the data with the smallest error [3].

The commonly used method to fit the linear regression data is the least squares method. The least squares method is an exact mathematical method and its goal is minimisation of the sum of squared residuals, where residual is the difference between the observed value and the value provided by the regression model [4].

In the current research, a linear regression is performed by statistical tools embedded in the Microsoft Excel spreadsheet application, which fits the coefficients for linear model.

B. Symbolic Regression

Symbolic regression or function identification is an approach to find mathematical expressions in a symbolic form, which will fit regression data in the best way and predict a dependent variable from explanatory variables with the smallest error. In the symbolic regression both the symbolic form of a model and coefficients for model variables are found. The symbolic regression differs from a traditional linear or polynomial regression, where only the best coefficients for linear or polynomial models should be found [5].

The symbolic regression approach is closely related to the genetic programming, which is the natural choice to find symbolic expressions that fit the data.

C. Genetic Programming

Genetic programming (GP) is an evolutionary algorithmic approach to find computer programs that perform the defined task in the best way [5]. Genetic programming is derived from a genetic algorithm: it works with a population of solution candidates (i.e., individuals) and performs evolution via iterative execution of selection, crossover and mutation operators. The main distinguishing feature of GP is that the individuals are represented in form of functional trees and the fitness function determines how well the solution candidate's program performs a given task.

In a symbolic regression, the mathematical expression that should be identified is interpreted as a computer program, whose input data are explanatory variables and output is a dependent variable. The following evaluators can be used as fitness function in a symbolic regression: mean squared error, mean average error, Pearson R squared (R^2) coefficient of determination [6].

In the current research, the implementation of genetic programming based symbolic regression in HeuristicLab optimisation framework [7], [8] is applied.

IV. Analysis and Preparation of Input Data

A. Input Data Preparation

For the identification of the regression model, the following data pre-processing tasks are performed.

The measurements from Table II are taken as values of the dependent variable in the training dataset. As the number of measured samples is relatively small, all data in the corresponding table will be used. Explanatory variables are derived from Table I, but the table is transformed in the following way.

An analysis of dataset from Table I shows that an hourly water level has only small changes between neighbour samples; moreover, the water level often does not change each hour. Thus, for the regression task only a small part of input data will be selected.

It is assumed that the following data are related to each sample of water flow discharge measures:

1. current water level in the river (measurement taken at time, when the corresponding water flow is measured);
2. water level of the river several hours before the flow measurement;
3. water level of the river several days before the flow measurement;
4. water level one week before the flow measurement is carried out.

In the transformation process of the input data, the following table with prepared data is obtained (see Table III). The table has the following attributes:

- *flow* – the river water discharge in m^3/s;
- $h0$ – the water level at water flow measuring time;
- $h3$ – the water level 3 hours before the water flow measurement ($h6$, $h12$, $h18$ are levels at 6, 12 and 18 hours, respectively, before the flow is measured);

- *d*1 – the water level measured 24 hours before the flow measurement (*d*2 and *d*3 are river levels 2 and 3 days before the flow measurement, respectively);
- *d*7 – the water level measured a week before the current flow is measured.

TABLE III

THE DATASET OF THE REGRESSION TASK (FRAGMENT)

flow	h0	h3	h6	h12	h18	d1	d2	d3	d7
182,32	-1	-4	-10	-17	-24	-25	-24	-16	24
197	72	71	68	66	62	59	51	44	44
310	30	31	38	29	29	32	31	29	34
408	67	65	62	58	56	52	33	26	51
525	119	117	115	111	106	95	67	52	42
1030	277	278	279	281	283	284	289	290	264
1153	300	299	298	297	297	297	302	311	333
1303	346	348	350	354	359	362	369	362	303
1180,5	347	349	351	356	361	368	389	408	428

The regression experiments are performed with the dataset described in Table III. A full table contains 93 records with the Daugava water flow discharge measures from January 1, 2008 to February 28, 2013.

B. Input Data Statistical Analysis

A correlation analysis is performed on the data of Table III with a statistical tool of Microsoft Excel application. Results of this analysis are shown in Table IV:

TABLE IV

CORRELATION BETWEEN INPUT DATA ATTRIBUTES

	flow	*h0*	*h3*	*h6*	*h12*	*h18*	*d1*	*d2*	*d3*	*d7*
flow	1									
h0	0.94808	1								
h3	0.94805	0.9995	1							
h6	0.94921	0.99946	0.99971	1						
h12	0.94925	0.99838	0.99911	0.99954	1					
h18	0.94918	0.99622	0.99726	0.99809	0.99937	1				
d1	0.94845	0.99546	0.99649	0.99753	0.99895	0.99965	1			
d2	0.93949	0.98526	0.98667	0.98868	0.99097	0.99306	0.99523	1		
d3	0.92102	0.96784	0.97006	0.97292	0.97644	0.98018	0.98384	0.99482	1	
d7	0.77774	0.83352	0.83736	0.84241	0.8483	0.85461	0.86223	0.89484	0.92647	1

The correlation analysis shows that the lowest correlation is between a water flow value and a water level measured a week ago. The highest correlation with a water flow value has water level values, which are collected during a day before the flow is measured. For the analysed dataset the highest correlation has a water level measured 12 hours before water discharge is measured. Moreover, explanatory values in the training dataset have a high correlation between each other; thus, most of them can be omitted.

V. LINEAR REGRESSION

Several experiments to obtain a linear regression model of the water flow value were performed within a regression toolbox of Microsoft Excel application.

At first, the model with only one explanatory variable *h*12, which has shown the highest correlation with the dependent variable in the correlation analysis, is fitted with the least squares method. The obtained model has the following mathematical expression:

$$flow \approx 82.386 + 3.482 \cdot h12 \qquad (1)$$

The model with one coefficient has a coefficient of determination $R^2 \approx 0.9011$, which shows its high reliability.

For the linear model with all dataset attributes listed above, except one week old measurement of water level, the following linear approximation has been obtained:

$$flow \approx 92.703 - 2.770 \cdot h0 - 3.078 \cdot h3 + 13.033 \cdot h6 - 12.401 \cdot h12 + 1.817 \cdot h18 + 7.932 \cdot d1 + 2.276 \cdot d2 - 3.391 \cdot d3 \qquad (2)$$

Such a model fits the data with coefficient $R^2 \approx 0.908$; thus, it does not increase the accuracy of the regression model, but the model becomes large and uses a big number of attributes.

For the validation of the obtained regression models, the dependency (1) of the river flow discharge from the river level is visualised in a chart and shown in Fig. 1. As it can be seen, the linear regression poorly fits the empirical data. The flow data have a trend, which looks close to the polynomial trend (see Fig. 1). The linear model is bad for high water levels; therefore, it can be concluded that it is not feasible in flood forecasting, where river levels are high.

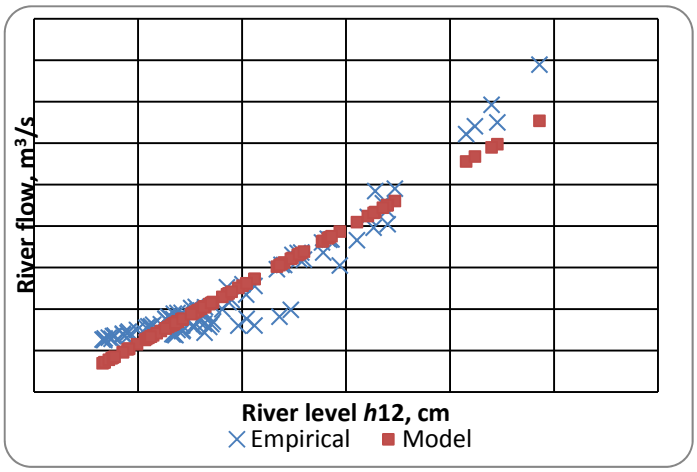

Fig. 1. Dependence of a river flow on the river level *h*12 for the regression model (1) and empirical data

The linear regression model (2) applied to the training set also shows the chart, which is visually close to Fig. 1.

VI. SYMBOLIC REGRESSION WITH GENETIC PROGRAMMING

In search for more precise models, the symbolic regression experiments with the application of genetic programming are performed. In the following experiments, *HeuristicLab* optimisation framework is applied [7]. The following parameters are defined for the algorithm of genetic programming: a population size equal to 200 individuals; a subtree swapping crossover and all GP mutation operators implemented in HeuristicLab with a mutation rate of 5% [9]. The proportional selection operator is used in the algorithm with one elite individual. Fitness function is evaluated by Pearson R^2 coefficient. Available tree nodes are: real value constants in a range [–100, 100], explanatory variables, arithmetic functions (+, -, *, /), trigonometric functions (sin, cos, tg), exponential, logarithm and power functions. Maximal

tree depth is limited to 10 nodes and a maximal tree length – to 25 nodes. The termination condition of the GP algorithm is defined as 1000 generations. Input dataset was randomly divided into a training set and a test set. The training set includes 80% of records and test set includes 20%.

Results obtained in genetic programming show that the found symbolic models in most cases are close to linear models and actually are polynomial models. In 20 GP experiments the optimisation framework has found several models with very close fitness. The model with the best found fitness, which is expressed as a formula, is given in (3) and is shown in tree representation in Fig. 2. Here tree leaves correspond to an explanatory variable multiplied by coefficients or to constants, and intermediate nodes are multiplication and addition operators.

$$flow \approx (((-0.033 \cdot d3 - 0.032 \cdot h0) + (((0.677 \cdot h0 - 0.033 \cdot d3) - \\ - 0.031 \cdot h6) - (0.666 \cdot h12 - 0.031 \cdot h6))) + 0.361 \cdot h18) \cdot \\ \cdot ((-0.361 \cdot h18 + 0.230 \cdot h12) - 24.078) \cdot (-0.202) + \\ + 216.678 \qquad (3)$$

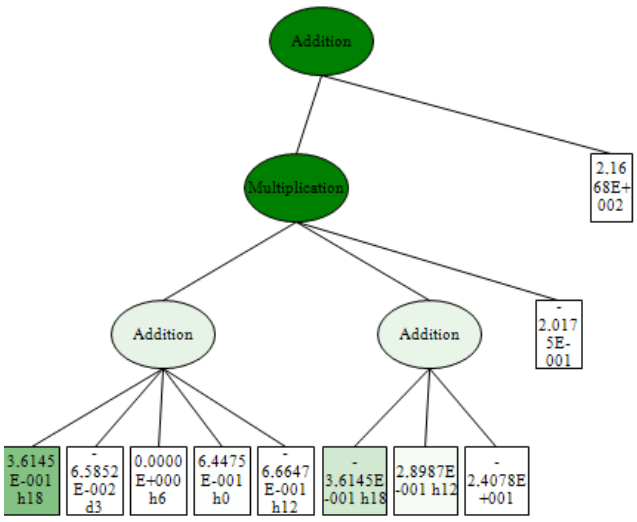

Fig. 2. Regression model (3) represented in form of a tree in an optimisation framework

The model (3) can be easily transformed in a more readable form, such as:

$$flow \approx 216.678 - 0.202 \cdot (0.645 \cdot h0 - 0.666 \cdot h12 + 0.361 \cdot h18 - \\ - 0.066 \cdot d3) \cdot (-0.361 \cdot h18 + 0.290 \cdot h12 - 24.078) \qquad (4)$$

As it can be seen, the main factors, which affect the water flow, are the same attributes $h12$ and $h18$. The model fits data with coefficient $R^2 \approx 0.963$ for the training set and $R^2 \approx 0.953$ for the test set. The model expresses the river flow in the polynomial form and has higher accuracy than a linear model. Line chart of the model is shown in Fig. 3 and a scatter plot is shown in Fig. 4, respectively. Both charts are obtained in the output of the experiment in the HeuristicLab framework. In the line chart, a blue line corresponds to the empirical values, red and yellow lines – to the model response. The division of the data in training and test datasets can be observed. In the scatter plot, the mapping of the model output to a dataset empirical flow is shown.

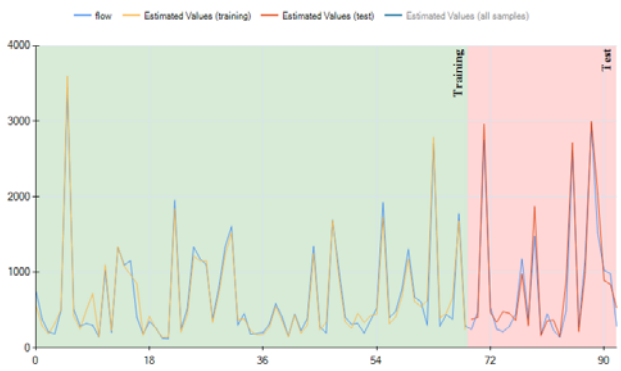

Fig. 3. Line chart of the model (3)

Data visualised in the scatter plot (Fig. 4) shows that the model is accurate for the majority of the records. High mismatch between the estimated and target values for several records is determined for low water data.

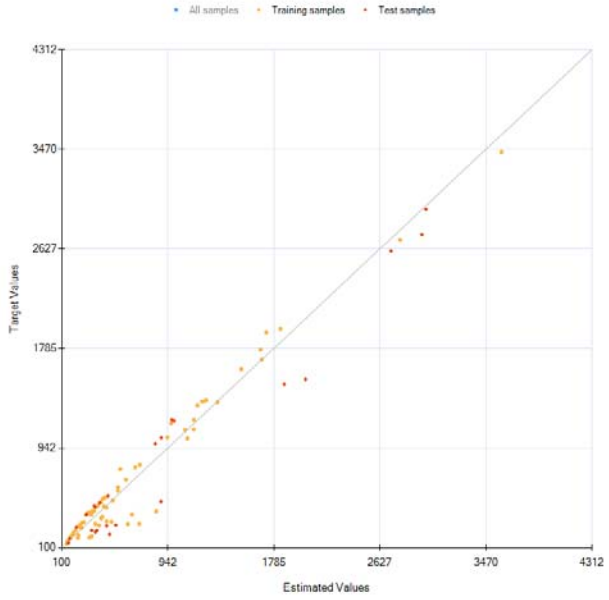

Fig. 4. Scatter plot of the symbolic model (3)

It should be noted that all models obtained after the termination condition of genetic programming include only simple mathematical operators, such as multiplication, addition and subtraction, but the solutions with trigonometric, exponential and logarithmic functions have bad fitness. Thus, the flow discharge should be described as a polynomial model.

Results of the validation experiments for the regression model (4) are shown in Fig. 5. It can be seen that the model is well fitted, and in the dataset there are only a few records that are very different from the model. These outliers probably are caused by some other river physical parameters, which were not included in the input data of the problem statement.

Nevertheless, the model (4) shows good results for records with a high water level; thus, it is applicable for flood forecasting.

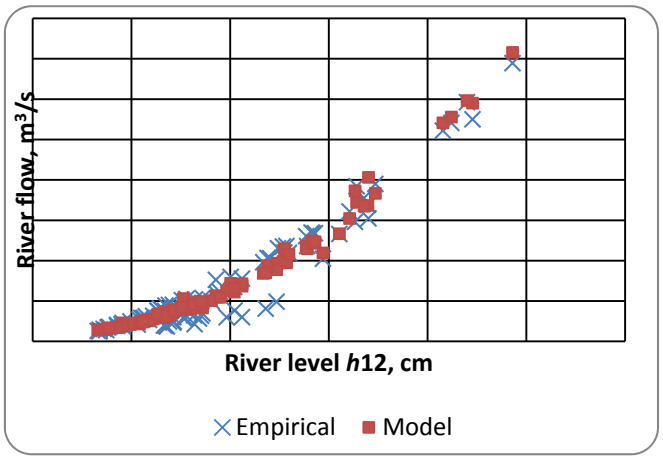

Fig. 5. Dependence of a river flow on the river level *h*12 for the regression model (4) and empirical data

As a majority of the explanatory variables are chosen as the measurements taken in the near past, a symbolic regression can be applied to the flow forecasting in the near future. In this case a regression model is obtained in GP that is applied to a dataset with the excluded current river level measurements (e.g., *h*0, *h*3). In the series of 20 GP experiments with the above-mentioned algorithm parameters and dataset with excluded *h*0, the following regression model was obtained:

$$\text{flow} \approx 208.214 - 0.011 \cdot h12 + 0.014 \cdot (1.123 \cdot h6 - \\ - 0.415 \cdot h3 + 1.503 \cdot h18 - 1.262 \cdot h12) \cdot (-0.814 \cdot h12 + \\ + 1.123 \cdot h6 + 98.845); \quad (5)$$

The model (5) can be applied to situations, when the current water level *h*0 in the river is not known, but it is possible to operate with measures that are made at least 3 hours ago. The model (5) has the Pearson's $R^2 \approx 0.957$ for the training set and $R^2 \approx 0.981$ for the test set.

It should be noted that the obtained models, when compared between different runs of GP, have different algebraic forms and values of coefficients, but at the same time they describe the training dataset in the same way and with a very close error. It can be concluded that the search algorithm performs well and converges to similar models that are just expressed in different forms.

VII. Conclusion

The main result of the research is that river flow discharge can be estimated through water level recent measurements taken at a particular monitoring station. To obtain the analytical model of the flow discharge, the regression model has to be fit with an application of genetic programming. The obtained river flow regression models used in the real life validation of the river flood prediction [1] have shown good results and the proposed methods are applicable for the solution of similar tasks.

The linear model obtained in the Microsoft Excel tool can be used as the simple equation for the flow calculation at a medium river level, but the model is not feasible in flood situations, when a water level is high. For a higher accuracy of output data, the model obtained in genetic programming has to be applied.

However, the best models obtained in a symbolic regression also have small errors and do not fit perfectly several records in the dataset. This can be explained by a small number of input factors, which include only values from river level measurements. Thus, more parameters obtained at a river monitoring station, which affect the water flow, should be included in the dataset to search for more precise flow discharge models in the future.

Acknowledgments

The research has been supported by the project 2.1/ELRI-184/2011/14 "Integrated Intelligent Platform for Monitoring the Cross-border Natural-Technological Systems" as part of "Estonia-Latvia-Russia Cross-border Cooperation Programme within European Neighbourhood and Partnership Instrument 2007–2013".

References

[1] V. Zelentsov, J. Petuhova, S. A. Potryasaev, S. A. Rogachev, "Technology of operative automated prediction of flood during the spring floods", Tr. SPIIRAN, vol. 29, pp. 40–57, 2013.

[2] Latvian Environment, Geology and Meteorology Centre, "LVGMC / Observations / Hydrology / Observation Stations," 2013. [Online]. Available: http://www.meteo.lv/en/hidrologijas-staciju-karte/?nid=613 [Accessed: Sep. 2, 2013].

[3] N. R. Draper, H. Smiths, *Applied Regression Analysis*. Wiley-Interscience, 1998.

[4] D. C. Montgomery, *Design and Analysis of Experiments*. John Wiley & Sons, inc., 2005.

[5] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, USA: MIT Press, 1992.

[6] W. Mendenhall, T. Sincich, *Statistics for Engineering and The Sciences: Fifth Edition*. New Jersey: Pearson Prentice Hall, 2006

[7] "HeuristicLab: A Paradigm-Independent and Extensible Environment for Heuristic Optimization", 2013. [Online]. Available: http://dev.heuristiclab.com/ [Accessed: Sep. 2, 2013]

[8] G. Kronberger, S. Wagner, M. Kommenda, A. Beham, A. Scheibenpflug, and M. Affenzeller, "Knowledge Discovery Through Symbolic Regression with HeuristicLab," in *Proceedings of the 2012 European conference on Machine Learning and Knowledge Discovery in Databases - V. Part II (ECML PKDD'12)*, 2012, pp. 824-827.

[9] M. Kommenda, G. Kronberger, S. Wagner, S. Winkler, and M. Affenzeller, "On the Architecture and Implementation of Tree-based Genetic Programming in HeuristicLab," in *Proceedings of the Fourteenth International Conference on Genetic and Evolutionary Computation Conference Companion (GECCO Companion '12)*, 2012, pp. 101-108.

Vitaly Bolshakov is a doctoral student at Riga Technical University, Latvia. In 2009 he received Mg.sc.ing. degree in Information Technology. At the moment, he is completing the doctoral thesis. The thesis is devoted to a simulation-based fitness landscape analysis in optimisation of complex systems. His major areas of research include metaheuristic optimisation, fitness landscape analysis and simulation-based optimisation.

Since 2009 Vitaly Bolshakov has been a Researcher at the Department of Modelling and Simulation of Riga Technical University (1 Kalku Street, Riga LV-1658, Latvia). E-mail: vitalijs.bolsakovs@rtu.lv

Vitālijs Boļšakovs. Regresijā bāzēta Daugavas upes plūdu prognozēšana un pārraudzīšana
Rakstā ir apskatīta lineāras un simbolu regresijas pielietošana upes plūdu prognozēšanas un pārraudzīšanas uzdevumos. Apskatītajā pētījumā tiek risināti uzdevumi, kuru atrisināšanas rezultāti ir pielietoti daudz plašākā upes plūdu seku noteikšanas metodikā. Pētījuma galvenie uzdevumi ir upes plūsmas caurplūduma analītiskā modeļa noteikšana, kurā caurplūdumu var noteikt, balstoties uz esošo un neseno upes ūdens līmeni, kā arī caurplūduma vērtību prognozēšanu tuvākā nākotnē. Formulētas problēmas galvenie izaicinājumi ir upes caurplūduma mērījumu mazā kopa un mazs ieejas faktoru skaits. Caurplūduma modeļa apmācīšanai ir izmantoti Daugavas pārraudzīšanas stacijas vēsturiskie dati, kas ir savākti netālu no Daugavpils pilsētas. Uzdevuma ieejas dati ir transformēti tādā veidā, lai esošo upes caurplūdumu būtu iespējams aprēķināt, balstoties uz vairākiem ūdens līmeņa mērījumiem nesenā pagātnē. Rakstā ir apskatīti un salīdzināti vairāki regresijas modeļu ieguves scenāriji. Pirmajā scenārijā uzdevuma risināšanai ir paredzēts pielietot lineāru regresiju un tai atbilstošas metodes. Otrajā scenārijā tiek risināts simbolu regresijas uzdevums, pielietojot ģenētisko programmēšanu un konfigurējot esošo programmlīdzekli. Simbolu regresijā iegūtie analītiskie modeļi rāda mazāku apmācības kļūdu un labāk tuvina datus, nekā lineārie modeļi. Iegūto modeļu precizitāte ir papildus validēta grafiskā veidā, veicot empīrisko un iegūto datu salīdzināšanu. Pētījumā iegūtie aprēķinu modeļi rāda labus rezultātus un ir teicami pielietojami upes plūsmas vai caurplūduma prognozēšanas uzdevumos, kā arī tos ir iespējams pielietot, prognozējot upes plūdu sekas.

Виталий Большаков. Прогнозирование и мониторинг наводнений для реки Даугава на основе регрессии
В статье рассмотрено применение линейной и символьной регрессии в задачах прогнозирования и мониторинга речных наводнений. В статье решаются задачи, результаты решения которых используются в более обширной методике определения последствий речных наводнений. Основной задачей данного исследования является нахождение аналитической модели определения расхода воды в реке в зависимости от текущего и предыдущего уровня воды, а также прогнозирование предполагаемого расхода воды в ближайшем будущем. Главными проблемами в нахождении такой модели являются малое количество исторических замеров расхода воды, а также малое количество входных переменных. Для получения модели использованы исторические данные со станции наблюдения реки Даугавы возле города Даугавпилс. Входные данные задачи преобразованы таким образом, чтобы текущий расход воды в реке можно было бы рассчитывать на основании нескольких замеров уровня реки в ближайшем прошлом. В статье рассмотрены и сравниваются несколько сценариев получения регрессионных моделей. В первом случае предполагается решение задачи линейной регрессии и применение соответствующих методов. Во втором случае решается задача символьной регрессии с применением генетического программирования и существующего оптимизационного инструмента. Аналитические модели, найденные путём символьной регрессии, показывают меньшую ошибку обучения и аппроксимируют данные лучшим образом, чем линейные модели. Точность найденных моделей дополнительно провалидирована графическим сопоставлением эмпирических и полученных значений. Полученные в исследовании модели расчёта показывают хорошие результаты и высокую применимость в задачах прогнозирования расхода воды в реке и в решении задач прогнозирования последствий наводнений.