Flood Prediction Using Support Vector Machines (SVM)

Shi, Y[1], Cheng T[1] and Taalab K P[1]

[1] SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London, Chadwick Building, Gower Street, London WC1E 6BT, United Kingdom

January 5, 2016

Summary

Flooding is a destructive phenomenon that can risk human life, damage homes and have huge economic impacts. To plan and implement effective mitigation strategies, it is necessary to predict when and where flooding will occur. Based on a combination of rain gauge and river discharge measurement taken from the River Don catchment, UK this study proposes a Support Vector Machine (SVM) based approach to predicting river. The purpose of this work is to show the potential of the SVM method for predicting future flood events.

KEYWORDS: Flooding, Support Vector Machines, River Flow, Rainfall

1. Introduction

Flooding is a destructive phenomenon that can risk human life, damage homes and destroy infrastructure. Furthermore, floods often lead to significant economic losses for the public and governments. The consequences of widespread flooding across Northern England during December 2015 had been estimated to have a total economic cost of nearly £6 billion (Dathan, 2015). In the UK, the main cause of fluvial flooding is intense rainfall causing river flow to exceed capacity. To ensure public safety and implement mitigation strategies in a timely and effective manner, we are interested in predicting when and where floods will occur.

Hydrological models are used to identify areas at risk of flooding, predict the magnitude of floods and determine what measures of anthropogenic protection may be needed in the future (Lundin, et al. 2015). Generally, predictive hydrological models can be classified into two groups: data-driven and physically based models (Jajarmizadeh et al., 2015). Data-driven models only rely on the input data and using mathematical or statistical function to link with the output, such as the artificial intelligence techniques application including Artificial Neural Network and Support Vector Machine (Leavesley et al., 2002). The second group are theoretical models, which aim to represent our understanding of the physical environment based on physical rules, processes and interactions (e.g. Moore et al., 1998). These models are often highly complex and contain large uncertainties. The focus of this paper will be flood forecasting using a data mining approach, specifically Support Vector Machines (SVM), which have been shown to be effective tools for a range of hydrological modelling applications across a number of continents (Suliman et al., 2013).

The data-mining approach to hydrological modelling typically relates a number of expletory variables, such as precipitation, historic river discharge or upstream flow, to a variable of interest such as peak flow downstream or stage height, based on measured observations. This can then be used to predict future river discharge.

2. Methodology

2.1 Support Vector Machines (SVM)

SVMs are a machine learning technique, developed based on statistical learning theory (Vapnik, 1998). In this study, regression is the main applying approach of SVM. The basic idea of support vector regression (SVR) is to map a linear regression. Regression is motivated to seek and optimize the certain bounds between the true value and regressed value, the errors situated within these bounds are ignored by the loss function. This type of function is often called an epsilon intensive loss function. Figure 1 displays an example of one-dimensional non-linear regression function with epsilon (ε) intensive loss function.

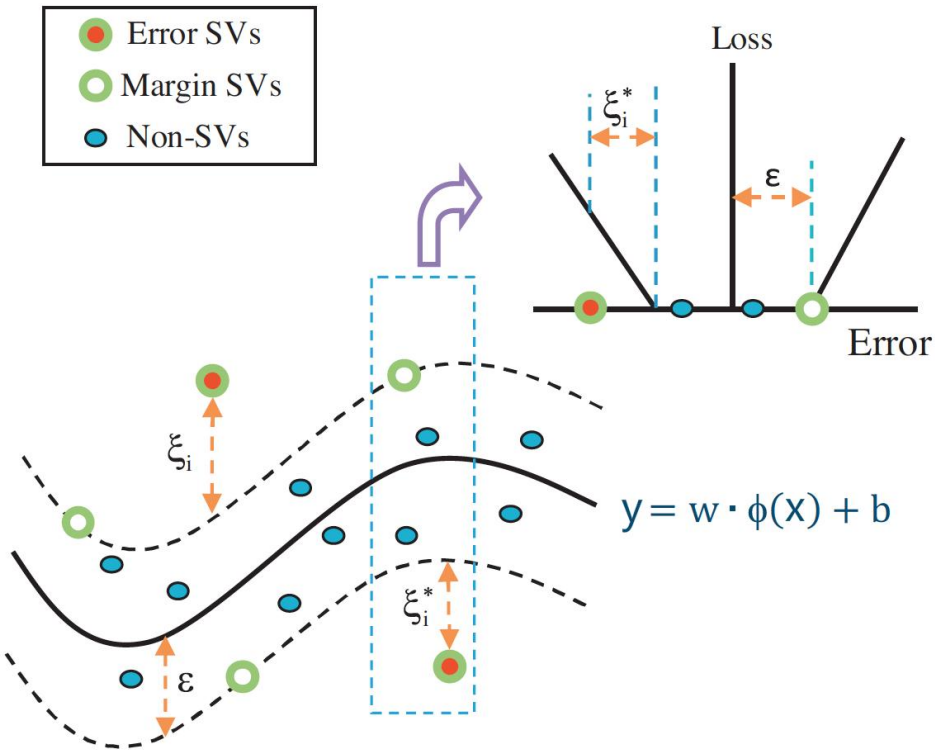

Figure 1: Nonlinear Support Vector Regression ε-insensitive loss function derived from (Deka, 2014)

In SVR, the original input data sets X are converted from the input space to a high dimensional space, and produce the linear or nonlinear regression model between a series of input variables and the dependent variables in this feature space. Using mathematical function, the linear model (in high dimension space) is f (x, w) give by

$$y = f\ (x, w) = \sum_{j=1}^{m} w_j g_j(x) + b = w \times \phi(x) + b \qquad (1)$$

Where $g_j(x), j = 1, \ldots m$, denotes a set of nonlinear transformations. However, w (vector of coefficients) and b (constant 'bias' term) are the regression function parameters and ϕ is the kernel function. The quality of regression estimation is measured by the loss function L (y, f (x, w)) called epsilon insensitive loss function proposed by Vapnik (1998):

$$|\xi|_\varepsilon = |y - f(x)|_\varepsilon = \begin{cases} 0 & if\ |y - f(x) \le \varepsilon \\ |y - f(x)| - \varepsilon & otherwise \end{cases} \quad (2)$$

Here (non-negative) slack variables $\xi_i, \xi_i', i = 1, \ldots n$ are introduced. These measure the deviation of training samples outside epsilon-insensitive zone. Two slack variables specify upper and lower constraints on the outputs of the error tolerance (ε). The minimization function of slack variables is formulated as following:

$$\begin{aligned} Min \quad & \frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i') \\ Subject\ to: & \begin{cases} y_i - f(x_i, w) \le \varepsilon + \xi_i' \\ f(x_i, w) - y_i \le \varepsilon + \xi_i \\ \xi_i, \xi_i' \ge 0, i = 1, \ldots, n \end{cases} \end{aligned} \quad (3)$$

Where $||w||^2$ controls the model complexity (which is to be minimised); C is the cost coefficient (positive constant) which determines the degree of penalized loss when a training error occurs.

2.2 Study Area and Data Description

In this study, the River Don is chosen as the study river, because of the occurrence a severe flooding event in recent decades. This river is located in the Sheffield city of South Yorkshire region in the midland of the United Kingdom. The Don catchment covers an area of 1256.2 km2, and close to the east of Peak District (Figure 2).

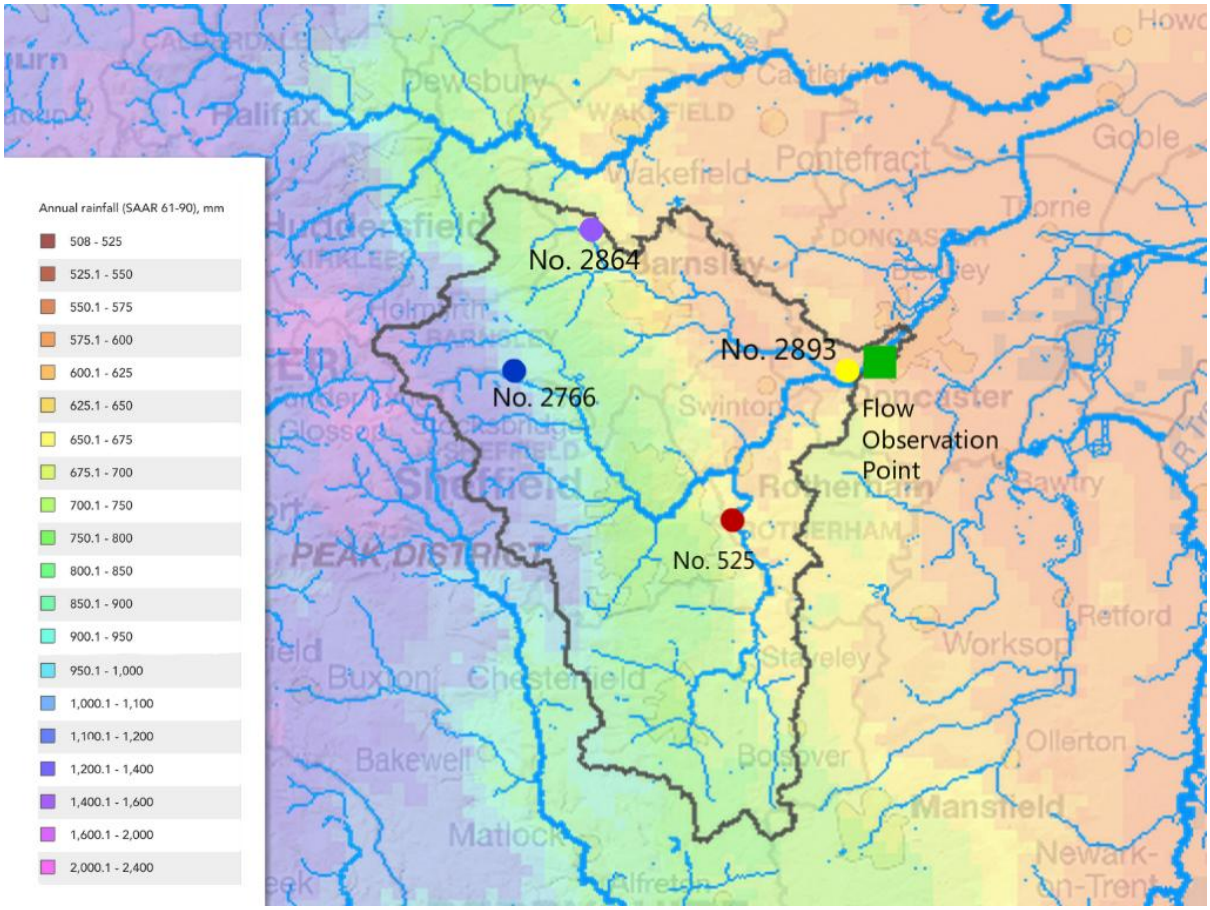

Figure 2: Boundary of Don Catchment and the location of the rainfall gauges stations

Data used in this study is rainfall data and river flow data. The River Don flow is measured by full range ultrasonic, and the observation station of flow is situated close to the outflow of the catchment. The data was downloaded from National River Flow Archive (NRFA). The rainfall data are derived from four rain gauges stations situated across the catchment shown in Figure 2. Rainfall data are obtained from Met Office. Among all the gauges inside Don Catchment, only these four contain the sufficiently complete historical rainfall data. The period used for the study spanned 30 years from January 1985 to June 2015. The relationship between rainfall and runoff are often intertwined and complicated, and it is obvious that flow data at given time involves some information of the past rainfall record, because change of flow data generally are the results of past rainfall events (Seifi & Riahi-Madvar, 2012). Therefore, rainfall and flow in three days are generally used as the variables for the flow prediction in the future week (Jajarmizadeh et al., 2015). An example of the dataset is shown in Table 1.

Table 1: An Example Of The Interpreted Rainfall And Flow Data

Month	Day	Year	R525 (mm)	R2766 (mm)	R2964 (mm)	R2893 (mm)	Flow (m^3/s)
1	1	1985	0	0	0	2.85	13.2

The objective of this study is to predict river flow discharge (Q) over a week time period. To do this, we train a total of 7 models for each day being predicted using a total of 15 predictor variables. These variables are rainfall measurement at each gauge measured at t-1, t-2 and t-3 as well as previous flow data measured at t-1, t-2 and t-3 (these are values measured one, two and three days before the first

prediction is made). Of a total of 10826 records of river flow and associated rainfall measurements, 8600 were used to train the SVM models and 2226 were used to validate model predictions. The accuracy of predictions is assessed using RMSE and residual analysis.

3. Results

3.1 Correlation analysis

The rainfall data of each station and river flow have significant correlations for the prediction of flow. Figure 3 shows the relationship of rainfall at time t-1 from four different stations with river flow at time t, (R525 refers to rainfall data at gauge no. 525). The values above the dotted line imply the variables have significant relationship. Because this study aimed to predict the river flow at given time using previous rainfall data, we are primarily interested in the negative lags. It is evident that the rainfall of all the gauges stations at time t-1 have the highest relationship around 0.6 with river flow at time t. As the lead-time increased, the correlation of rainfall and river flow decreases. Lead-time is the time interval between the stimulation and the response, which in this study means the time interval between the input and output. For instance, the lead-time of $R525_{t-1}$ and the flow at time t (Q_t) is 1. In this study, the largest lead-time is nine (e.g. Q_{t-3} to Q_{t+6})

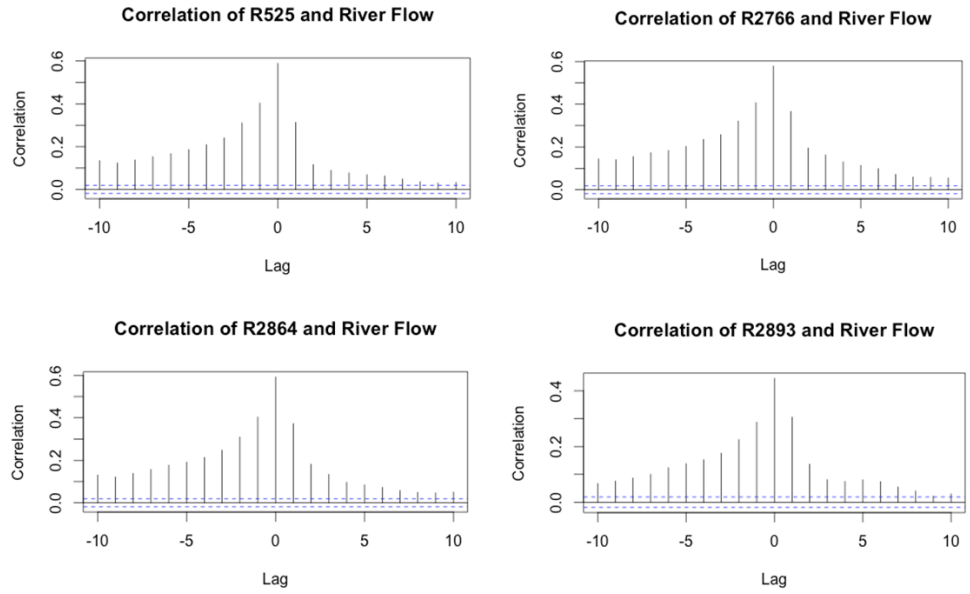

Figure 3: Correlation results of different rain gauges and river flow over various time lags

In addition, the autocorrelation of river flow is shown in Figure 4. This shows that the correlation between Q_{t-1} and Q_t, is higher than between Q_t and any of the rainfall gauges. In summary, all rain gauges in the catchment, as well as previous flow measurements can be feasibly used to predict the seven days river flow. However, as the lead-time increased, the relationship of input variables and output decreased, which implies that the accuracy of regression model may decrease also decrease. Gamma testing suggested that at short lead times, previous river flow is a more significant predictor of future flow, while rainfall becomes more important at larger lead times.

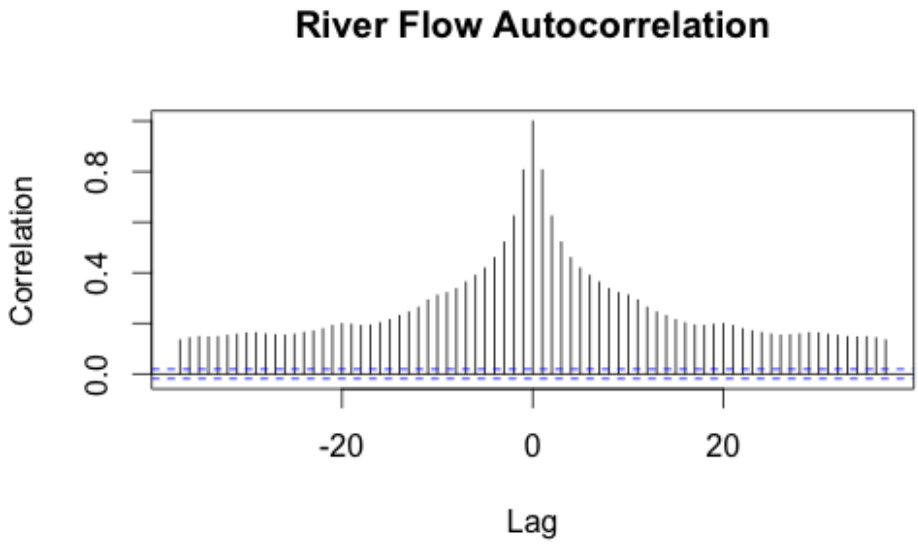

Figure 4: Autocorrelation of the Don River flow

3.1 SVM predictions

The RMSE results for Q_t, Q_{t+1} and Q_{t+6} are shown in Table 2. Based on the validation results, it is clear that shorter lead times are more accurately predicted than longer lead times. However, there is relatively little difference in RMSE between Q_{t+1} and Q_{t+6}

Table 2: RMSE of the calibration model and validation model from seven different outputs

Prediction Target	Calibration (Training)	Validation (Testing)
$\mathbf{Q_t}$	0.326	0.450
$\mathbf{Q_{t+1}}$	0.634	0.779
$\mathbf{Q_{t+6}}$	0.817	0.928

Predictions of flow for Q_t and Q_{t+6} are shown in Figure 5. This confirms that Q_t is predicted more accurately than river flows with larger lead times. The major difference is that at Q_{t+6} the high peak flows (i.e. those likely to cause a flood) are not well predicted.

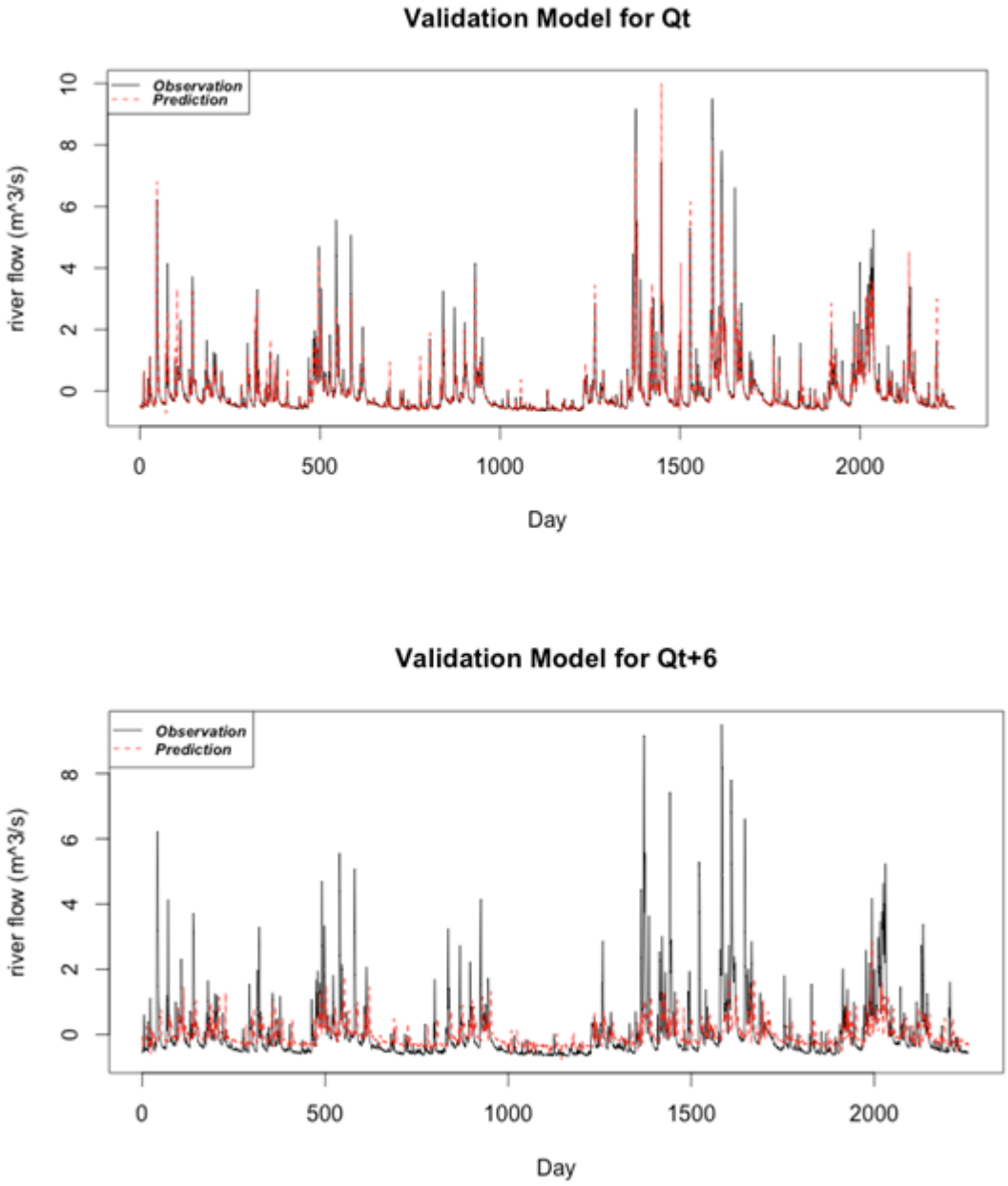

Figure 5: The river flow predictions of Q_t and Q_{t+6}

4. Conclusion

SVMs have been shown to be highly accurate predictors of both river flow change and peak flow over short lead times (up to 48 hours). This method could provide short-notice warnings that flooding was imminent. This compares favorably with similar studies that focus on single day lead times (Behzad, et a;., 2009; Sivapragasam & Liong, 2004) or lead times of several hours (Yu et al, 2006). Beyond 48 hours, predictions of peak flow are not sufficiently accurate to determine flood occurrence, however, the long term predictive accuracy of the method may be improved with the inclusion of additional explanatory variable such as soil moisture and water table data. The drawback of this approach is that it is not possible to establish confidence intervals around predictions. The quality of prediction can only be assessed through validation using data.

5. References

Behzad, M., Asghari, K., Eazi, M., & Palhang, M. (2009). Generalization performance of support vector machines and neural networks in runoff modeling. Expert Systems with applications, 36(4), 7624-7629.

Dathan, M. (2015) UK flooding: Economic cost of storms could hit £6bn, industry experts warn 28 December 2015. The Independent. http://www.independent.co.uk/news/uk/politics/uk-flooding-economic-cost-of-storms-could-hit-6bn-industry-experts-warn-a6788316.html

Deka, P. C. (2014). Support vector machine applications in the field of hydrology: a review. Applied Soft Computing, 19, pp. 372-386.

Jajarmizadeh, M., Lafdani, E. K., Harun, S., and Ahmadi, A. (2015). Application of SVM and SWAT models for monthly streamflow prediction, a case study in South of Iran. KSCE Journal of Civil Engineering, 19(1), pp. 345-357.

Leavesley, G. H., Markstrom, S. L., Restrepo, P. J., and Viger, R. J. (2002). A modular approach to addressing model design, scale, and parameter estimation issues in distributed hydrological modelling. Hydrological Processes, 16(2), pp.173-187.

Lundin, L.C., Bergstrom, S., Eriksson, E., and Seibert, J. (2015) 11. Hydrological Models and Modeling. Retrieved from: [online]
http://www.balticuniv.uu.se/index.php/component/docman/doc_download/270-the-waterscape-11-hydrological-models-and-modelling> [27, June, 2015]

Moore, I. D., O'Loughlin, E. M., and Burch, G. J. (1998). A contour-based topographic model for hydrological and ecological applications. In EARTH SURF. PROC. LANDFORMS. Vol. 13, No. 4, pp. 305-320.

National River Flow Archive (NRFA) (2015) 27021 – Don at Doncaster. In Centre for Ecology & Hydrology: Natural Environment Research Council. Retrieved from: <http://nrfa.ceh.ac.uk/data/station/info/27021> [6, June, 2015]

Seifi, A., and Riahi-Madvar, H. (2012). Input Variable Selection in expert systems based on hybrid Gamma Test-Least Square Support Vector Machine, ANFIS and ANN models. Provisional chapter. Intech.

Sivapragasam, C., & LIONG, S. Y. (2004). Identifying Optimal Training Data Set-A New Approach. In Proceedings of the 6th International Conference (Vol. 21, p. 24).

Suliman, A., Nazri, N., Othman, M., Abdul, M., & Ku-Mahamud, K. R. (2013). Artificial neural network and support vector machine in flood forecasting: A review. In Proceedings of the 4th International Conference on Computing and Informatics, ICOCI. pp. 28-30.

Vapnik, V.N. (1998). Statistical learning theory (Vol. 1). New York: Wiley.

Yu, P. S., Chen, S. T., & Chang, I. F. (2006). Support vector regression for real-time flood stage forecasting. Journal of Hydrology, 328(3), 704-716.

6. Acknowledgements

This project has received funding from the European Union's Seventh Programme for research,

technological development and demonstration under grant agreement No 603960

7. Biography

Yu Shi is a 2015 MSc GIS graduate from University College London, currently pursuing a career in GI science in China.

Tao Cheng is a Professor in GeoInformatics, and Director of SpaceTimeLab (http://www.ucl.ac.uk/spacetimelab), at University College London. She has broad knowledge and experience in Geographic Information Sciences (GISc), from data acquisition, to information processing, management and analysis, with applications in environmental monitoring, natural resource management, health, transport and crime studies. She has over 140 publications.

Khaled Taalab is a research associate in the SpaceTimeLab at University College London. In 2013 he was awarded a PhD in digital soil mapping. He is currently working on a European FP7 project called InfraRISK which is developing models linking natural hazards, critical infrastructure and the environment. His research interests include environmental mapping, data mining and predictive modelling.