doi:10.5194/piahs-368-204-2015

Annual runoff prediction using a nearest-neighbour method based on cosine angle distance for similarity estimation

GUANGHUA QIN[1,2], HONGXIA LI[1,2] XIN WANG[1,2] QINGYAN HE[2] & SHENQI LI[2]

1 State Key Laboratory of Hydraulics and Mountain River Engineering, Sichuan University, Chengdu, 610065, China
hx_li406@126.com

2 College of Water Resource & Hydropower, Sichuan University, Chengdu, 610065, China

Abstract The Nearest Neighbour Method (NNM) is a data-driven and non-parametric scheme established on the similarity characteristics of hydrological phenomena. One of the important parts of NNM is to choose a proper distance measure. The Euclidean distance (EUD) is a commonly used distance measure, which represents the absolute distance of a spatial point and is directly related to the coordinate of the point, but is not sensitive to the direction of the feature vector. This paper used the cosine angle distance (CAD) for the similarity measure, which reflects more differences in the direction, and compared it to EUD. This technique is applied to annual runoff at YiChang station on the Yangtze River. The results show the NNM with CAD has a better performance than that of EUD.

Key words nearest neighbour; similarity estimation; Euclidean distance; cosine angle distance; annual runoff

1 INTRODUCTION

Annual runoff is influenced by climate, land cover and human activities, and its prediction is quite complex because it has a longer lead time than daily or monthly prediction. The commonly used prediction methods for annual runoff can be divided into traditional and modern methods. The traditional method is carried out according to the variation characteristics of runoff, such as hydrological analysis and hydrological statistics (Chen *et al.*, 1985; Chen, 1997; Kenea and Thian, 2009; Xu *et al.*, 2010). The modern methods, which developed with computing technology, include artificial neural networks (Seckin *et al.*, 2013), the fuzzy method (Zhu *et al.*, 2009), chaos method (Sivakumar, 2000), grey method (Liu, 2009), etc., have achieved some good results. However, most of them are developed based on the prediction pattern “assumption–calibration–verification”, which needs parameter calibration before prediction.

The Nearest-Neighbour Method (NNM) is data driven and non-parametric, with potential priority, and needs no assumption about the form of the dependence and probability distribution, or estimation of many parameters. Using NNM to model hydrologic process and dynamics in rivers and streams has been well documented (e.g. Lall and Sharma, 1996; Yuan *et al.*, 2000; Wang *et al.*, 2001; Mehrotra and Sharma, 2006; Lee *et al.*, 2011; Liu *et al.*, 2012), since Karlsson and Yakowtz (1987) used NNM for rainfall–runoff forecasting. One of the important parts of NNM is to choose a proper distance measure, as different distance measures may behave quite differently (Qian *et al.*, 2004). Euclidean distance (EUD) is a commonly used distance measure, which represents the absolute distance of a spatial point and is directly related to the coordinate of the point. The cosine angle distance (CAD) is another popular distance measure, which is sensitive to the direction of the feature vector, but has not been used in hydrological time series.

This paper used CAD for the similarity measure in annual runoff prediction. The annual runoff prediction of YiChang station is used as a case study. Section 2 presents the nearest neighbour method for hydrological time series; Section 3 is a theoretical analysis of CAD for runoff; Section 4 is the prediction of YiChang runoff at Yangtze River to assess the NNM with CAD; and the conclusions are summarized in Section 5.

2 NNM FOR HYDROLOGICAL TIME SERIES

Generally, correlation exists between hydrology phenomena through time. Thus, the extent, X_t depends on the historical runoff $Q_{t\text{-}1}$, $Q_{t\text{-}2}$, …, $Q_{t\text{-}P}$. Given $D_t = (Q_{t\text{-}1}, Q_{t\text{-}2}, …, Q_{t\text{-}P})$, this is called

The feature vector of the runoff series. Then, X_t = (Q_t, Q_{t+1}, …, Q_{t+m-1} (t = P + 1, P + 2, …, n–m+1) and can be defined as the succeeding value of D_t。

Among D_t(t = P + 1, P + 2, …, n) which are constituted by $\{Q_t\}_n$, there must be some feature vectors that are nearest neighbours to the current feature vector D_i. Suppose the number of nearest neighbour feature vectors is K, and it is represented by $D_{1(i)}$, $D_{2(i)}$, …, $D_{K(i)}$, then $X_{1(i)}$, $X_{2(i)}$, …, $X_{K(i)}$ must be the succeeding values of each corresponding feature vector. The nearest neighbour is judged by the difference between D_i and D_t, which is usually calculated by Euclidean distance:

$$r_{t(i)} = (\sum_{j=1}^{p} (d_{ij} - d_{tj})^2)^{\frac{1}{2}} \tag{1}$$

where $r_{t(i)}$ represents the difference between D_i and D_t, d_{ij} and d_{tj} are number j variable of D_i and D_t respectively, and P is the dimension of the feature vector. Then, $r_{j(i)}$(j = 1,2, …, K) is denoted as the difference between $D_{j(i)}$ and D_i, and it should be mentioned that $r_{1(i)} < r_{2(i)} <\ldots< r_{K(i)}$ (the number j is ordered according to the value of $r_{j(i)}$). The smaller $r_{j(i)}$ is, the nearer D_i and $D_{j(i)}$ will be and X_i is more similar to $X_{j(i)}$. Let $G_{j(i)}$ be the nearest neighbour bootstrapping weight of $X_{j(i)}$, which shows similarity between X_i and $X_{j(i)}$. Obviously, $G_{j(i)}$ is related to $r_{j(i)}$。

As discussed above, the relative value of number l variables of number j nearest neighbour succeeding vector $X_{j(i)}$, is known. The succeeding vector X_i can be obtained through multiplying predicted runoff $G_{j(i)}$. Thus, the ultimate formula of the NNM model can be given as:

$$X_i = \sum_{j=1}^{k} G_{j(i)} X_{j(i)} \tag{2}$$

The NNM model is confirmed when the number of nearest neighbours, K, the dimension of feature vector P, and the nearest neighbour bootstrapping weight $G_{j(i)}$ are estimated.

Generally, K = $\text{int}\sqrt{n - P}$ is given. If $P \geq 2$, the dimension of feature vector P can be estimated by a runoff auto-correlation graph or the trial and error method.

There are a number of methods to estimate bootstrapping weight $G_{j(i)}$. When estimating, first of all, its restraint condition must be satisfied, and then the bootstrapping weight $G_{j(I)}$ should be related to $r_{j(i)}$, and the bootstrapping weight function should be equal to one (equation (3)). As the number j is ordered according to the value of $r_{j(i)}$, in this paper, the following formula is adopted:

$$\sum_{j=1}^{k} G_{j(i)} = 1 \tag{3}$$

$$G_{j(i)} = \frac{(1/j)}{\sum_{L=1}^{k} 1/L} \quad (j = 1,2,\ldots,K) \tag{4}$$

When K is confirmed, we can only calculate $G_{j(i)}$ once.

3 COSINE ANGLE DISTANCE FOR NNM

The value of angle (D_t, D_i.) is defined as follows:

$$cos(D_t , D_i) = \frac{D_t \cdot D_i}{|D_t||D_i|} \tag{5}$$

where D_t is the feature vector and D_i is the current vector. The smaller cos (D_t, D_i) is, the nearer D_t and D_i.

We illustrate our approach to comparing EUD and CAD using a 2-dimensional space. Figure 1 shows a 2-dimensional space where A is a query point. Suppose that NN(A) is the nearest neighbour of A by EUD, and the EUD between query point A and NN(A) is r. B and C are two points that are on the ssp(A,r). B and C have the same distance to A for EUD, as dist(A,B) = dist(A,C), so it is possible to judge which is better. But for CAD, angle (A,C) < angle (A,B), so C is nearer to A.

In the annual runoff prediction, the angle of the feature vector is very important. So we propose to predict annual runoff using NNM based on CAD for similarity estimation. NNM based on CAD can also be used in daily or monthly runoff prediction, as the hydrological variation trend can be reflected by the angle of the feature vector.

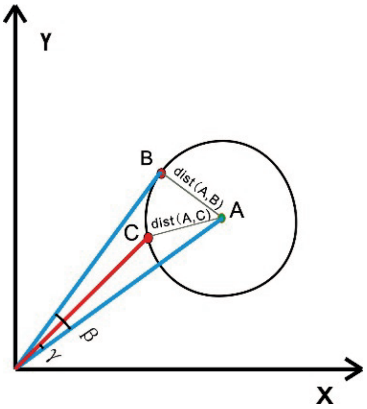

Fig. 1 Difference between Euclidean distance and cosine angle distance.

4 CASE STUDY

4.1 Data

The data used in this study are annual runoff (1890–2010) from Yichang station on the Yangtze River (Fig. 2). Yichang station (10 005 501 km^2) is the controlling station for the Three Gorges Dam. The data from 1890 to 1989 are used to develop the model, and data from 1990 to 2010 are used for prediction and model assessment.

First, the annual runoff time series of Yichang station was examined to determine any trends during the past 121 years. Figure 3 shows that it has a decreasing trend.

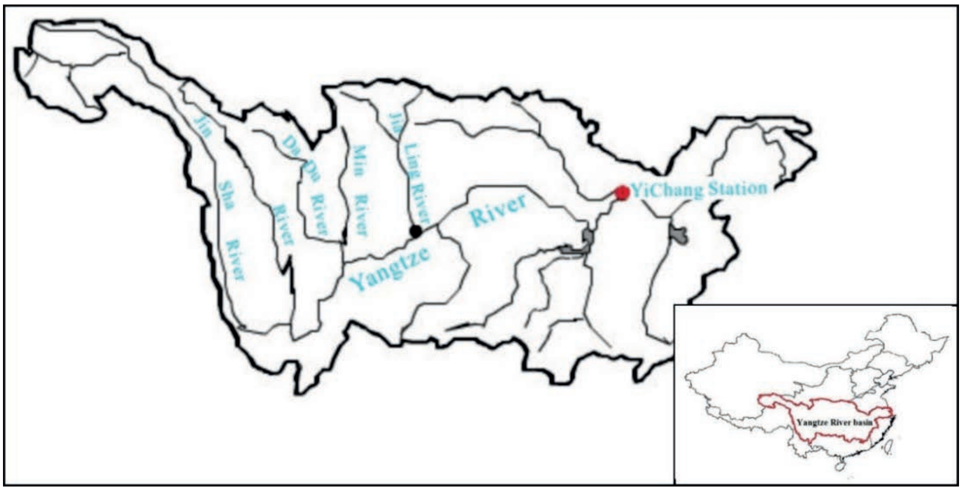

Fig. 2 The location of Yichang station on the Yangtze River, China.

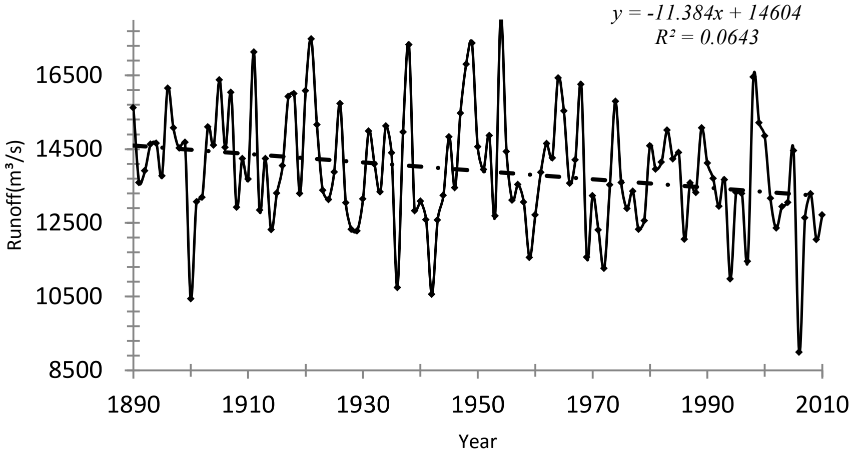

Fig. 3 Annual runoff time series of Yichang station.

4.2 Prediction results and discussion

Through primary selection of the model parameters, by trial and error, it determines that p = 3, nearest neighbour number = 6. Then the annual runoff of the year 1980–1989 is used to constitute the feature vector D_t (t = 1, 2, 3... 88), a total of 97, which is used to predict annual runoff of the years 1990–2010. The mean relative error (MRE) is 5.10%, the qualified rates (QR, error less than 10%, 20% and 30%) are all 100.0% and r is 0.897 (Table 1, Figs 4 and 5). Compared with EUD, CAD is obviously better, with a lower MRE and higher QR.

CAD and EUD consider one aspect of the similarity measure; developing a better distance measure which considers both the direction and absolute distance would further improve the runoff prediction using NNM. Also, NNM, the same as using other models, has a low accuracy for annual runoff prediction and would make no sense when the future motion trail of the series is out of the law obtained from the historical data.

Table 1 Prediction performance based on CAD and EUD.

Similarity estimation	e < 10%	e < 20%	e < 30%	MRE	*r*
EUD	71.4%	90.5%	95.2%	8.84%	0.048
CAD	100.0%	100.0%	100.0%	5.10%	0.897

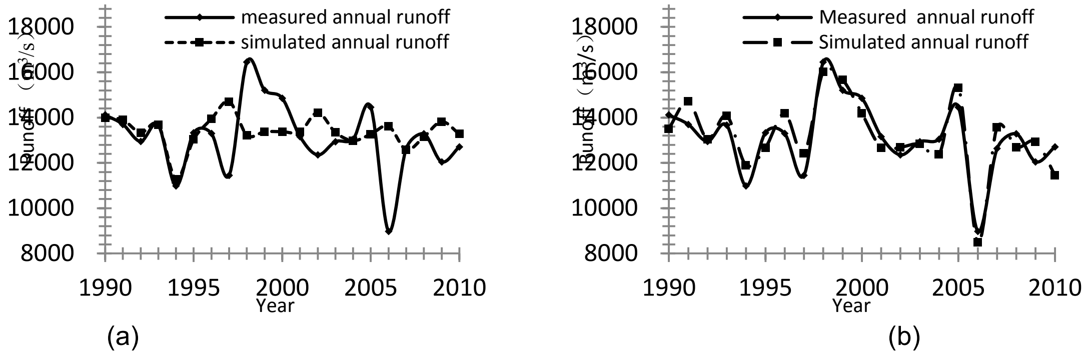

Fig. 4 Comparisons between measured and simulated annual runoff during the period of 1990–2010: (a) model performances with EUD, and (b) model performances with CAD.

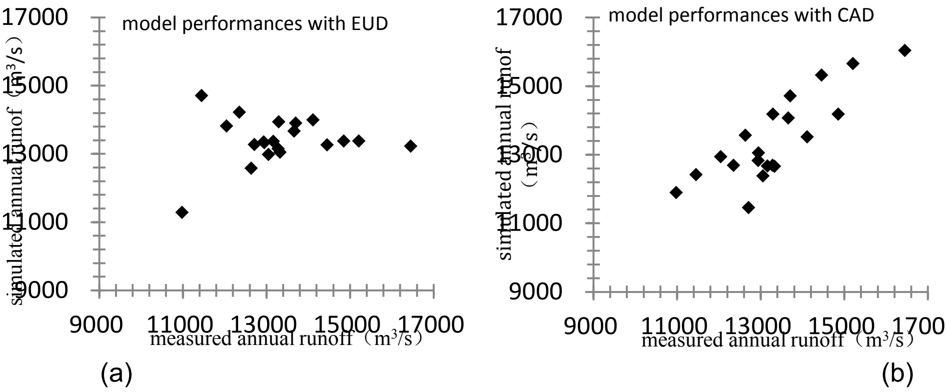

Fig. 5 Comparisons between measured and simulated annual runoff during the period of 1990–2010: (a) model performances with EUD, and (b) model performances with CAD.

5 CONCLUSION

This paper predicted annual runoff prediction using NNM, with two distance measures, CAD and EUD. The results of annual runoff prediction at YiChang station showed that the results of CAD are significantly better than EUD as the former reflects more differences in the direction. Developing a better distance measure which considers both the direction and absolute distance would further improve the runoff prediction using NNM.

Acknowledgements The present research is supported by the National Key Basic Research Program of China (973 Program, No. 2013CB036401), National Natural Science Foundation of China (No. 51209152) and Doctoral Program of Higher Education of China (No. 20110181120100).

REFERENCES

Chen, J.R., *et al.* (1985) The situation and development trend of medium-and-long term hydrological forecasting in China. In: *Selected papers of hydrological forecasting*. Water Power Press, Beijing.

Chen, S.Y. (1997) Theoretical pattern of comprehensive analysis and method for mid and long term hydrology forecasts. *Journal of Hydraulic Engineering* 8, 15–21.

Karlsson, M. and Yakowitz, S. (1987) Nearest-neighbor method for nonparametric rainfall–runoff forecasting. *Water Resources Research* 23(7), 1300– 1308.

Kenea, G. A. and Thian, K. G. (2009) Statistical Ensemble Seasonal Streamflow Forecasting in the South Saskatchewan River Basin by a Modified Nearest Neighbors Resampling. *Journal of Hydrology* 14(6), 628–639.

Lall, U. and Sharma, A. (1996) A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research* 32(3), 679– 693.

Lee, T., *et al.* (2011) Identification of model order and number of neighbors for k-nearest neighbor resampling. *Journal of Hydrology* 404, 136–145.

Liu, Y., *et al.* (2012) Application of nearest neighbor nootstrapping regressive model in the dry season monthly runoff forecast. *Water Sciences and Engineering Technology* 6, 14–16.

Liu, C.Y., *et al.* (2009) Application research on hydrological forecasting based on grey prediction model. *Proceedings of the Third International Conference on Information and Computing Science* 1, 290–293.

Mehrotra, R. and Sharma, A. (2006) Conditional resampling of hydrologic time series using multiple predictor variables: A K-nearest neighbour approach. *Advances in Water Resources* 29(7), 987–999.

Qian, *et al.*, (2004) Similarity between Euclidean and cosine angle distance for nearest neighbor queries. *Proceedings of the 2004 ACM Symposium on Applied Computing (SAC)*, 14–17.

Seckin, N, *et al.* (2013) Comparison of artificial neural network methods with L-moments for estimating flood flow at ungauged sites: the case of East Mediterranean River Basin, Turkey. W*ater Resources Management* 27, 2103–2124.

Sivakumar, B. (2000) Chaos theory in hydrology: Important issues and interpretations. *Journal of Hydrology* 227, 1–20.

Wang, W.S., *et al.* (2001) Predication of hydrology and water resources with nearest neighbor bootstrapping regressive model. *Hydroelectric Energy* 19(2), 8–10.

Xu, D.M., *et al.* (2010) Review on study of mid-and long–term hydrological forecasting technique. *Water Conservancy Science and Technology and Economy* 1, 1–7.

Yuan, P., *et al.* (2000) Nonparametric perturbing nearest neighbor bootstrapping model for simulation of flood time series. *Journal of Sichuan University (Engineering Science Edition)* 32(1), 82–86.

Zhu, Y.Y., *et al.* (2009) Rough fuzzy inference model and its application in multi-factor medium and long-term hydrological forecast. *Water Resources Management* 23, 493–507.