FUZZY NEAREST NEIGHBOUR METHOD FOR MONTHLY INFLOWS FORECASTING INTO "IRON GATES I" RESERVOIR

Prof.dr.ing. **Radu POPA**[1] and Ing. **Cristian BOCSE**[2]
[1] University POLITEHNICA Bucharest
[2]S.C. Hidroelectrica S.A. –S.H. Portile de Fier

Abstract**:** This paper details a nearest neighbour pattern recognition method for hydrological time-series forecasting. The aim of the forecasting algorithm is to make single point forecasts into the future on the basis of the past nearest neighbours. The nearest neighbours are selected using some membership criteria.

The monthly inflow series at the "Iron Gates I" reservoir was used to illustrate the forecasting method. This time –series covers a period on one hundred and sixty years (1840 – 1999) and contains a great variety of patterns.

The periodic components of the mean and standard deviation are first eliminated using the Fourier approach, and then the pattern modeling was applied to the resulted stochastic components time-series. This pattern recognition based tool is a different way to analyze the time dependence which exists into the stochastic components time-series of many hydrologic variables.

Two experimentations are developed: the first one having a basic subseries of 144 years data and a test subseries of 16 years (10% of total), and the second one with 120 and respectively 40 years (25% of total). The forecasts are compared to the actual values over the two test periods. The results include the mean and maximum absolute percentage errors, the mean and maximum absolute errors and other error measures. An ARMA (4,0) classical model obtained with PEST program was used as a reference case.

The results are very encouraging for operational forecasting purposes, considering the small amount of field data used in analysis. The proposed method is simple to understand and to put into practice. Further work should be prompted to compare the ability of such tools against well established statistical and neural network methods.

Keywords**:** time – series forecasting, fuzzy nearest neighbour method, monthly inflows prediction in reservoir.

1. Introduction

Time series prediction is used to support the decision – making in many applications areas, including power generation. The most typical estimations are concerned with power load forecasting, but in mixed system (hydro and non – hydroelectric plants) some accurate forecasts of reservoir inflows allow to the decision – maker to allocate economically the fuel resources, production strategies etc.

In the past, conventional statistical techniques based on the Box – Jenkins approach (ARIMA models) have been extensively used for modeling and predicting time series. This approach, although widespread, is only capable to construct linear models. Subsequently, other variants of this model have been developed such as ARARMA and delta – NARMA, a non – linear extension of the preceding type. Apart from statistical models, a number of advanced methods has been applied to the problems of forecasting, including artificial neural networks, evolutionary neural networks, neuro – fuzzy systems, fuzzy techniques and pattern imitation methods (Singh and Stuart, 1998).

In this paper a pattern matching technique is discussed in connection with monthly inflows forecasting into "Iron Gates I" reservoir. Thanks to its basic philosophy, the method can be named as well a fuzzy nearest neighbour technique. The main characteristic of the nearest neighbour pattern recognition method lies in the fact that it is a powerful tool for identifying some relationships between current and past data sequences of an univariate time – series. After the

identification of several historic neighbours, these could be then used for prediction by either averaging their values or using an extrapolation procedure. The accuracy of the method is strongly dependent on the pattern matching algorithm.

2. Fuzzy pattern matching

A pattern matching approach is based on the realistic premise that current structures of a given process can be matched with old structures to perform a future forecast. This approach is a fuzzy type one, because there is only a partial match between a current and any possible past structure.

Pattern modeling refers to the process of describing the time – series as a series of patterns. To obtain such patterns, a mathematical formalism must be imposed in time – series analysis. This formalism depends on particular origin of the recorded data. Many of hydrological processes present a cyclic component, induced by some natural circumstances. Assuming that any detectable influences have been removed, the remained time – series data can be considered as a vector $\boldsymbol{s} = \{s_1, s_2, \ldots, s_t\}$, where s_t is the latest value in the series. A pattern can be specified in terms of the gradient at a given time (i.e. an upward or downward change) and its size (number of included segments). If a segment is defined as the difference $d_i = s_{i+1} - s_i$, a new time –series can be then attached to the original one, i.e. the difference vector $d = \{d_1, d_2, ..., d_{t-1}\}$. And this new time –series can be encoded mathematically as a vector of binary values b_i = 0 or 1. For $d_i < 0$ (i.e. $s_{i+1} < s_i$) we can choose b_i = 0 and if $d_i > 0$ (i.e. $s_{i+1} > s_i$), then b_i = 1. In cases when the series doesn't change ($d_i = 0$) we can use b_i = 2, but such a situation appears only by hazard in hydrological processes. The complete time – series is therefore encoded as a binary vector $b = \{b_1, b_2, ..., b_{t-1}\}$.

Formally, a pattern in the time –series contains one or more segments: $p_d = \{d_i, d_{i+1}, ..., d_j\}$ and can be represented as a binary string: $p_b = \{b_i, b_{i+1}, ..., b_j\}$. For a total of *k* segments in a pattern, $p_d = \{d_i, d_{i+1}, ..., d_{i+k-1}\}$, $p_b = \{b_i, b_{i+1}, ..., b_{i+k-1}\}$ and, excepting encoded value b_i=2, the total number of pattern shapes possible is $2^k + 1$.

The basic procedure involving pattern matching for forecasting is summarized as follows:

1. Suppose that we are at time moment *t*, having at disposal the known values $s_1, s_2, ..., s_t$ and trying to predict s_{t+1} value.
2. A pattern size, *k*, is selected and the latest known pattern of this size is considered, i.e. the pattern $c = \{b_{t-k}, b_{t-k+1}, ..., b_{t-1}\}$.
3. Search the time –series $\{b_1, b_2, ..., b_{t-k-1}\}$ in order to find the closest match for *c*. Suppose that closest match is found as $c' = \{b_{j-k}, b_{j-k+1}, ..., b_{j-1}\}$, where *j* is an index to mark the pattern position. Corresponding segments lengths for *c* and *c'* are $(d_{t-k}, d_{t-k+1}, ..., d_{t-1})$ and $(d_{j-k}, d_{j-k+1}, ..., d_{j-1})$, and corresponding time – series data are $(s_{t-k}, s_{t-k+1}, ..., s_{t-1}, s_t)$ and $(s_{j-k}, s_{j-k+1}, ..., s_{j-1}, s_j)$.
4. Use the past pattern to predict the future value s_{t+1} by a certain method. The simplest one is by averaging the *k* nearest neighbours, i.e.:

$$s_{t+1} = (s_{j-k+1} + s_{j-k+2} + ... + s_j + s_{j+1})/k \quad (1)$$

but some other more complicated schema may be adopted.

A lot of details should be given for each step.

1. Our time –series covers a period of 160 years (1840 – 1999), including monthly inflow data i.e. 1920 monthly values. The seasonal component having a period of 12 was first removed by using Fourier series for periodic mean and standard deviation. If P_{mj} and $P_{\sigma j}$, j =1,2, …, 12 denote these values, then the time –series of stochastic process s was derived with:

$$s_{i,j} = \frac{Q_{i,j} - Pmj}{P_{\sigma j}}, \quad I = 1, \ldots, 160;\ j = 1,2, \ldots, 12 \tag{2}$$

where $Q_{i,j}$ is the monthly inflow in the i^{th} year and j^{th} month. The other terms are as follows:

$$P_{mj} = Q_m + \sum_{n=1}^{N}\left[A_n \cos\left(n\frac{2\pi j}{12}\right) + B_n \sin\left(n\frac{2\pi j}{12}\right)\right]$$

$$P_{\sigma j} = Q_\sigma + \sum_{n=1}^{N}\left[A_n' \cos\left(n\frac{2\pi j}{12}\right) + B_n' \sin\left(n\frac{2\pi j}{12}\right)\right]$$

$$A_n = \frac{2}{12}\sum_{j=1}^{12}\left(Q_{mj} - Q_m\right)\cos\left(n\frac{2\pi j}{12}\right),\ \ldots,\ B_n' = \frac{2}{12}\sum_{j=1}^{12}\left(Q_{\sigma j} - Q_\sigma\right)\sin\left(n\frac{2\pi j}{12}\right) \tag{3}$$

$$Q_{mj} = \frac{1}{160}\sum_{i=1}^{160} Q_{i,j}\ ,\ j = 1, 2, \ldots, 12;$$

$$Q_{\sigma j} = \sqrt{\frac{1}{160}\sum_{i=1}^{160}\left(Q_{i,j} - Qmj\right)^2}\ ,\ j = 1, 2, \ldots, 12;$$

$$Q_m = \frac{1}{12}\sum_{j=1}^{12} Q_{mj}\ ;\ \ Q_\sigma = \frac{1}{12}\sum_{j=1}^{12} Q_{\sigma j}$$

A total number of N = 5 harmonics was selected and the above relations had been used to yield the chronologically ordered time –series $s = \{s_1, s_2, ..., s_t\}$, t=1920, where the stochastic process ***s*** shows (in many cases) a certain time – dependence. For hydrological time –series, this is a typical case, and some conventional statistical techniques describe this dependence by various linear models as AR, MA, ARMA, ARIMA etc. In this paper, a such dependence will be identified by pattern matching.

2. Pattern size, k, could be accepted as an optimization parameter, and the optimal model should be selected according to some standard error measures (the mean absolute percentage error, the mean square error, etc). For our time –series ***s***, a value of k = 3 provides very good results.

3. The closest match for c implies to meet a number of conditions. Taking the current time –series value s_t, the all past values satisfying a proximity criterion are searched using:

$$\frac{1}{1 + \frac{\left|s_j - s_t\right|}{2}} > \alpha \tag{4}$$

or $\left|s_j - s_t\right| \le \frac{2}{\alpha} - 2$, $k \le j \le t - k$,

where $\alpha \in (0;1)$ is a threshold value set by the experimenter. As α increases, a smaller numbers of candidate neighbours are selected.

Among all selected candidates s_j, *j*=1,2,… only those having a similar pattern $(b_{j-k},...b_{j-1})$ with the current pattern $(b_{t-k},...b_{t-1})$ are analyzed on.

The third criterion uses the difference vector to evaluate the estimator

$$\delta_j = \sum_{i=1}^{k} \left| d_{j-i} - d_{t-i} \right|, \tag{5}$$

for all similar patterns with the current one.

Finally, the closest match is accepted for index pattern j^* having

$$\delta_{j^*} = \min\{\delta_j\} \tag{6}$$

and the corresponding past pattern $\left(s_{j^*-k}, s_{j^*-k+1},...,s_{j^*}\right)$ and / or $\left(d_{j^*-k}, d_{j^*-k+1},...,d_{j^*-1}\right)$ will be used to predict s_{t+1} value.

4. Forecasting method can be selected by a trial and errors procedure, using the recorded data. A way is by relation (1), where $j = j^*$. This method has been used in our paper, but in a more complex manner: the steps 3 – 4 were resumed for five threshold decreasing values α (i.e. α=0.96; 0.94; … etc) and for each j^* founded index, a forecast value s_{t+1} was obtained. The final forecast has been accepted as an averaged value among these five individual predictions.

An alternative solution makes use of the latest known value, s_t, and a change (up or down), derived with the closest past and current patterns as –for example:

$$s_{t+1} = s_t \pm \beta \cdot d_{j^*} \tag{7}$$

where the weighting factor β is obtained with:

$$\beta = \frac{1}{k} \sum_{i=1}^{k} \frac{d_{t-i}}{d_{j^*-i}}$$

and the sign (+) or (-) dictated by the binary value b_{j^*} (- for 0, + for 1).

Certainly, other alternatives may be imagined.

3. Numerical results

In this paper we compare the fuzzy nearest neighbour method with an AR (4) model obtained using the ITSM package – in particular the program PEST (Brockwell and Davis, 1994).

Numerical experiment has been conducted as follows:

- Firstly – the conventional Box –Jenkins procedure has been reproduced and parameters of the AR (4) model were derived with PEST, using the complete time –series (1920 monthly values, from 1840 through 1999). This model has been accepted as a reference model for both future predictions (after 1999) and some previous ones (before 2000).
- The overall data before 2000 have been divided into an estimation (training) subseries and a test (forecasting) subseries. Two sizes were used for this decomposition: 90% ÷ 10%, and 75% ÷25% of the complete time –series. The aim is to compare the performances obtained with AR (4) model and fuzzy pattern matching method in terms of some standard error measures.

- Finally, the recorded data collected from January 2000 through March 2002, were used to give an up-to-day image on these performances.

As error measures were selected:

- maximum and mean absolute errors, mean error for $\left(Q_i - Q_i^p\right)$
- maximum and mean absolute percentage errors, mean percentage error
- correlation coefficient $R = \sqrt{\sum \left(Q_i^p - Q_m\right)^2 / \sum (Q_i - Q_m)^2}$
- total number of predictions within 5%, 10% and 20% absolute percentage errors

where Q_i is the recorded inflow, Q_i^p is the predicted inflow for i^{th} month and Q_m is the mean inflow during the test period.

In both methods, for current prediction at time *i*, the necessary (four in PEST, all in fuzzy) past recorded data were used.

Table 1 shows the comparative performances of the two forecasting models for the test periods of 16 and respectively 40 years before 2000.

Table 1 –Ccomparative results on two test periods before 2000

		PEST		Fuzzy nearest neighbour	
Test period		16 years	40 years	16 years	40 years
MxAE ($m^3 \cdot s^{-1}$)		3165	4776	4099	4995
MnAE ($m^3 \cdot s^{-1}$)		965	1013	1031	1124
MnE ($m^3 \cdot s^{-1}$)		29	142	-16	36
MxAPE (%)		92.01	92.01	91.00	107.25
MnAPE (%)		20.20	19.50	21.75	21.68
MnPE (%)		-4.32	-2.41	-4.62	-3.66
R		0.851	0.818	0.955	0.946
Number of predictions with errors	≤ 5%	30	79	34	78
	≤ 10%	62	150	60	149
	≤ 20%	108	281	102	254
	from	192	480	192	480

It should be observed that on all mean error measures, excepting correlation coefficient, the results are very closed. Nevertheless, the pattern matching algorithm has the advantage of his great simplicity over the more complex ARIMA modeling.

Table 2 shows the monthly inflows from January 2000 through March 2002, together with the PEST and fuzzy predictions. The absolute percentage errors are also presented. In 17 months from 27 (63%) the pattern matching procedure yields better results. The number of within 1% APE situations is 4 for fuzzy method and one for PEST, and within 3% APE situations is 8 for fuzzy method and 3 for PEST. Honestly, there are 6 predictions with fuzzy method exceeding 30% APE and 3 cases only for PEST.

Table 2 – Actual data and predictions for January 2000 through March 2002 period

	Recorded Q ($m^3 \cdot s^{-1}$)	PEST		Fuzzy nearest neighbour	
		Predicted Q^p ($m^3 \cdot s^{-1}$)	Absolute percentage error (%)	Predicted Q^p ($m^3 \cdot s^{-1}$)	Absolute percentage error (%)
01.00	6017	5411	10.07	5620	6.60
02.00	7475	5565	25.55	5536	25.94
03.00	7892	8263	4.70	8003	1.41
04.00	10558	8619	18.37	9332	11.61
05.00	7353	9345	27.09	9681	31.66
06.00	4130	6364	54.09	7158	73.32
07.00	3593	4115	14.53	4072	13.33
08.00	3676	3478	5.39	2954	19.64
09.00	2696	3458	28.26	3038	12.69
10.00	3872	2853	23.73	2574	33.52
11.00	4264	4564	7.09	4282	0.47
12.00	3974	4513	13.56	4812	21.09
01.01	5084	3822	24.82	4284	15.74
02.01	5425	5266	2.93	5382	0.79
03.01	6854	6666	2.74	7016	2.36
04.01	8512	7879	7.44	8536	0.28
05.01	6280	8051	28.20	7452	18.66
06.01	6230	5748	7.74	6190	0.64
07.01	5116	5369	4.95	5611	9.68
08.01	3811	4154	9.00	4132	8.42
09.01	5690	3479	38.86	3377	40.65
10.01	4144	4942	19.23	4569	10.26
11.01	3993	4528	13.40	5407	35.41
12.01	4445	4463	0.40	4575	2.92
01.02	4017	4264	6.15	4134	2.91
02.02	6776	4470	34.03	4624	31.76
03.02	6595	7891	19.65	7780	17.97

In the figure 1 are plotted the forecasts made with both models and the actual data.

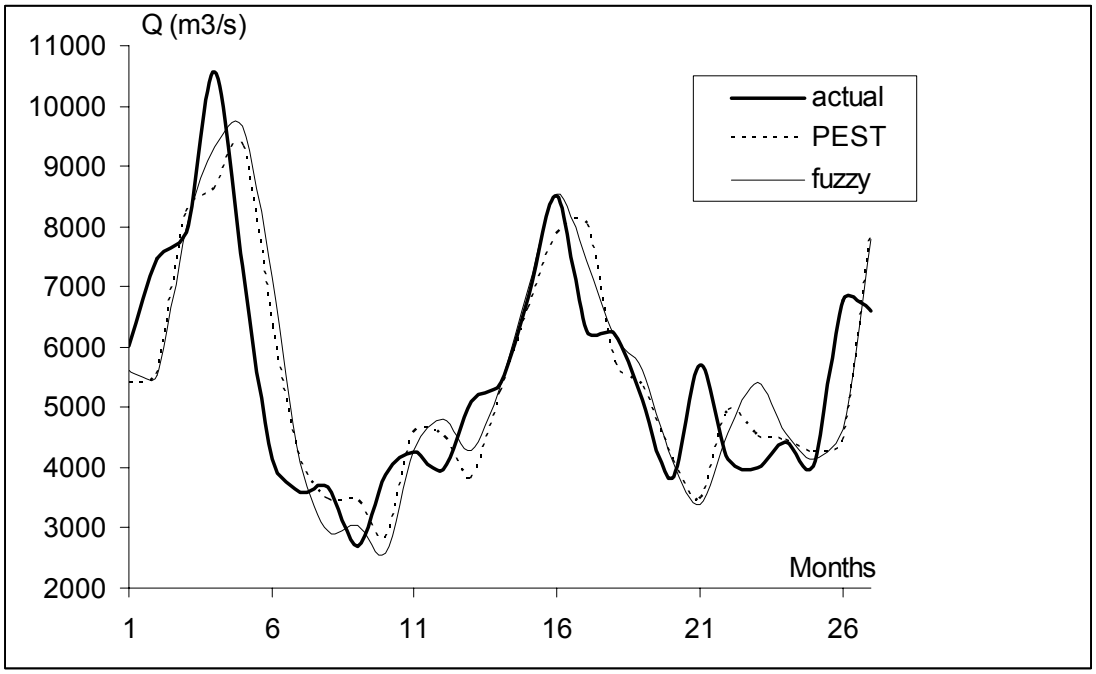

Figure 1: Actual data and predictions for 01.2000 – 03.2002 period

4. Conclusion

The proposed method was evaluated against the classical ARMA model used for forecasting. The results on monthly inflow series of Danube river at "Iron Gates I" are greatly similar for both methods. By its simplicity, the fuzzy nearest neighbour method could be preferable for operational proposes, when is no time to do laborious analysis.

The forecasting accuracy of this method is data dependent (theoretically, work better with large time –series) and can be improved by optimization of its parameters.

Further work should be devoted to another way on treating the non – stationary time –series, as for example to predict the difference or the difference of difference time –series, followed by the translation of forecasts to original data.

5. References:

1. Brockwell, P.J., Davis, R.A. (1994) – *ITSM for Windows; A user's guide to time series modeling and forecasting*; Springer – Verlag, New York.
2. Singh, S., Stuart, E (1998) – *A pattern matching tool for forecasting*; Proc. 14th Intern. Conf. on Pattern Recognition; Brisbane, IEEE Press, vol. 1, pp. 103 -105