

Caspian Sea Level Prediction by Auto Fuzzy Regression

Farhad Ramezani Moziraji, Mehdi Yaghobi and Jasem Zargari Kordkolaii

Computer Engineering Department,
Islamic Azad University, Ayatollah Amoli Branch, Amol, Iran

Abstract: we apply fuzzy techniques for system identification and apply statistical techniques to modelling system. This paper proposes an adaptive technique to predict the Caspian Sea level by combining fuzzy concept with statistical logistic regression. The methodology framework is creation regression modelling by fuzzy techniques. Identification is performed through learning from examples method introduced by Wang and Mendel algorithm. Delta test residual noise estimation is used in order to select the best subset of inputs as well as the number of linguistic labels for the inputs. Experimental results for Caspian Sea level prediction are compared with statistical model and showing the advantages of the proposed methodology in terms of approximation accuracy, generalization capability and linguistic interpretability.

Key words: Identification % Fuzzy techniques % Statistical techniques % Caspian Sea Level % Prediction

INTRODUCTION

The Caspian Sea is the largest lake on our planet [1]. Caspian Sea level changes are one of the most important natural impacts on the biodiversity of this huge water body. In 1977 to 1995 sever increase in the level of the water Caused worry for the people who live there so that it caused destruction of homes, fishing, commercial and administrative [2]. Farm lands went under water, it threaten too many cities, especially coastal city of sewage and threading of health and development in cities and villages. The increase of sea level during the past two decades caused various reasons was presented in this field consist of Hydroclimatology characters ties and the factor of human activities. The views and opinions about the causes of fluctuation of Caspian Sea is generally divided into two categories: 1. a series of geological factors that consists activating of salt domes on the sea floor, tectonic movement and slippage of the Volga. 2. A series of hydrological factors such as reaching changes of Volga water, warming of the earth and the emergence of solar glow, which are in this category [1]. These factors create a chain of causes and effects of communication unit which is finally ends whit sea level.

After the reduction of water level in 1930s, an especial attention was paid to predict the water level. Prediction methods were developed in several ways:

that can be named indirect methods, possible methods, climatology prediction calculation using water level. Bolgov stated that the only correct way to predict the water level is making physical model. It seams the correct method to predict sea water level is modelling the system and receive the result and out put of the system [3].

Generally forecasts done for the Caspian is in a category in Table 1.

Since now none of the predictions performed in the Caspian Sea water level has a correct answer. In perwater period it was predicted about 20 types from the level which in 19 type of it the Caspian Sea was shown with increasing water while only one of them shows the retreating.

Since it has been suggested by Box-Jenkins [4] that the time-series ARIMA model has enjoyed fruitful applications in forecasting social, economic, engineering, foreign exchange and stock problems.

For more than half-century, the Box-Jenkins methodology using autoregressive moving average model (ARMA) linear models have dominated many areas of time series forecasting [5].

Regression analysis is one of the basic tools of scientific investigation, enabling identification of functional relationship between independent and dependent variables [6]. In the classical regression

Table 1: Forecasts of Caspian water level between 1990 and 2007

Rank	Authors	Year	Method	Year	Level	Year	Level	Year	Level
1	Bodyko etal	1998	Palaeoanalogues	2000	-28.1	2020	-27.1	2050	-22.4
2	Labanov	1999	Palaeoanalogues	2000	-28.6	2020	-27.8	2050	-24.2
3	Frolow	1995	Probability/ Palaeo	2010	-27 to -25 by 90% probabillity				
4	Golitsyn	1998	Sea Balance	"Most Probable level -27 to -26 in 2000					
5	Klige	1994	Sea Balance	2005	-26	2010	-25.5 by50%probability		
6	Golubtsovetal	1995	Atmospheric circul	1995	-26.5	2000	-26.3	2010	-26.2
7	Rodionov	1991	Atmospheric circul	Notescecd -26					
8	Krenk & popove	1992	Atmospheric circul	Upper Limit -25 to -26					
9	Droydov	1990	Lapse rates	Contnue rising until 2005					
10	Sidorenkov	1990	Atmospheric circul	Contnue rising until 2010					
11	Kim & Nikulina	1994	North Atantic Circul	9 cm rise until 2040(ie 4.0 metres)					
12	Malinin	1994	Tele water air temp	2000	-26.5	2010	-25.5		
13	Naidenov etal	1993	Statistical	2 stable level -25.5 & -27.7 unstable -26.4					
14	Shlyamin	1993	Solar cyclicity	Rise to 2000 Fall to 2020 rise to 2060					
15	Alishaev etal	1993	Solar cyclicity	2025	-25	2050	-25.3	2100	-27
16	Sheko	1993	Solar cyclicity	12 cm rise to 1999 max, falling at 2005					
17	Duvanenin	1995	Wolf Number	2025	-25				
18	Getman	2001	Wolf / Belinski	2002	-25.3	2006	-25.5	2012	-24.5
19	National Centre for Caspian Studies and Research	1996	Sea Balance	2000	-26.3	2005	-25.5	2010	-24
20	Gorky	2001	Sea Balance	2005	-27.3	2015	-27.9	2030	-28.4

analysis both the independent and dependent variables are given as real numbers. It assumes that the future values of a time series have a clear and definite functional relationship with current, past values and white noise. This model has the advantage of accurate forecasting in a short time period [7].

These techniques have limited capabilities for modelling time series data and more advanced nonlinear methods including neural networks have been frequently applied. Fuzzy inference systems, despite its good performance in terms of accuracy and interpretability, have showed little application in the field of time series prediction as compared to other nonlinear modeling techniques such as neural networks and support vector machines.

In this paper, we propose an adaptive combining technique to prediction of Caspian Sea level. This method combined fuzzy concept and statistical regression. The methodology proposed here is intended to apply to crisp time series.

In the next section we propose a methodology framework. Section III illustrates the methodology through a case study for prediction Caspian Sea Level. The conclusion is presented in Section IV.

Proposed Model: Suppose that discrete time series as a vector $\bar{x} = x_1, x_2, \dots, x_t$ that represents an ordered set of values, where t is the number of values in the series.

The problem of predicting one future value, x_{t+1} using a statistical model with no exogenous inputs can be stated as follows:

$$\hat{x}_{t+1} = f_r(x_t, x_{t-1}, \dots, x_{t-M+1}) \tag{1}$$

Where \hat{x}_{t+1} is the prediction of model, f_r and M is the number of inputs to the regressors [8].

Predicting the first unknown value requires building a model, f_r , that maps regressor inputs (known values) into regressor outputs (predictions). When a prediction horizon higher than 1 is considered, the unknown values can be predicted following two main strategies: recursive and direct prediction.

Direct prediction requires that the process of building a f_r model be applied for each unknown future value. Thus, for a maximum prediction horizon H, H direct models are built, one for each prediction horizon h [9]:

$$\hat{x}_{t+h} = f_h(x_t, x_{t-1}, \dots, x_{t-M+1}) \text{with } 1 \leq h \leq H \tag{2}$$

Direct prediction does not suffer from accumulation of prediction errors.

In order to build each model, a fuzzy inference system is defined as a mapping between a vector of crisp inputs and a crisp output. In principle, any combination of membership functions, operators and inference model can be employed, but the selection has a significant impact

on practical results. As a concrete implementation, we use the minimum for conjunctions and implications, Gaussian membership functions for inputs, singleton outputs and fuzzy mean as defuzzification method following the Mamdani defuzzification model. In this particular case a fuzzy model with M inputs for prediction horizon h is formulated as:

$$F_h(\bar{x}) = \frac{\sum_{l=1}^{N_h} \min(m_{R_l^h}, \min m_{L_i^h}(x_v))}{\sum_{l=1}^{N_h} \min m_{L_i^h}(x_v)} \quad (3)$$

Where N_h is the number of rules in the rule-base for horizon h, $m_{L_i^h}$ are Gaussian membership functions for the input linguistic labels and $m_{R_l^h}$ are singleton membership functions.

The problem of building a model can be precisely stated as defining a proper number and configuration of membership functions and building a fuzzy rule-base from a data set of t sample data from a time series such that the fuzzy systems $F_h(\bar{x})$ closely predict the h-th next values of the sea level.

We propose a methodology framework in which a fuzzy inference system is defined for each prediction horizon has three stages. These stages are detailed in the following subsections.

Variable Selection: The first step in this method is choosing the optimal subset of inputs from the initial set of M inputs, with the maximum model size of M.

Delta Test is a nonparametric noise estimation method for estimating the lowest mean square error (MSE) that can be achieved by a model without over fitting the training set [10].

We use the result of the Delta Test applied to a particular variable selection as a measure of the goodness of the selection. The input selection that minimizes the Delta Test estimate is chosen for the next stages.

System Identification and Tuning: This stage comprises three substages that are performed iteratively and in a coordinated manner.

System Identification: In this substage, the structure of the inference system (linguistic labels and rule base) is defined. For the concrete implementation analyzed in this paper, identification is performed using the W&M algorithm driven by the Delta Test estimate.

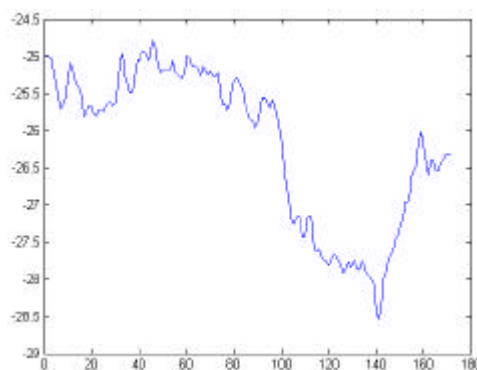


Fig. 1: Caspian Sea Level time series

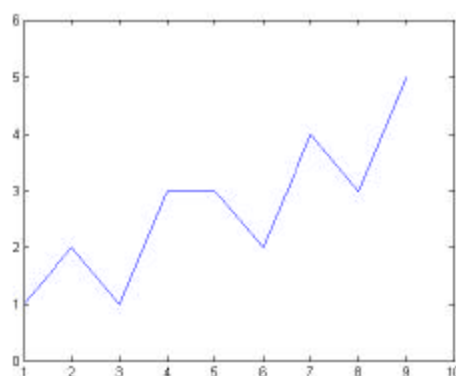


Fig. 2: Number of selected variables

This iterative identification process for increasing grid partitions of the universe of discourse stops when a system is built such that the training error is lower than the Delta Test estimate or a threshold based on the Delta Test estimate. The selection is made by comparing the error after the next (tuning) stage.

System Tuning: We consider an additional tuning step in the methodology as a substage separated from the identification substage. As concrete implementation for this paper we apply the supervised learning algorithm driven by the normalized MSE (NMSE).

Complexity Selection: As last step, the complexity of the fuzzy model is selected depending on the DT estimate. The first (simplest) system that falls within the error range defined by the Delta Test is selected.

Case Study (Predicting Caspian Sea Level): For the purposes of validating and illustrating the proposed methodology framework and concrete algorithms and criteria, we analyze the Caspian Sea level time series. This data set (Figure 1) consists of 171 samples of MEAN annual Caspian Sea water level [11].

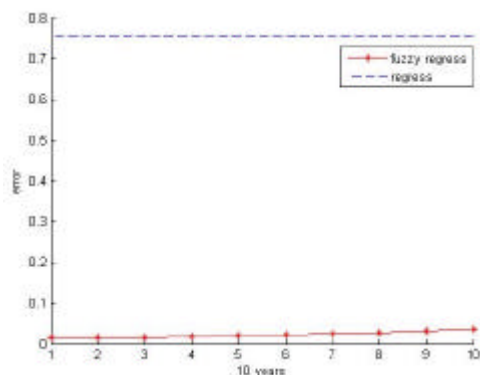


Fig. 3: Comparison of our methodology against statistical regressor. Generalization errors of regress models (- -). Generalization errors of fuzzy models (*-).

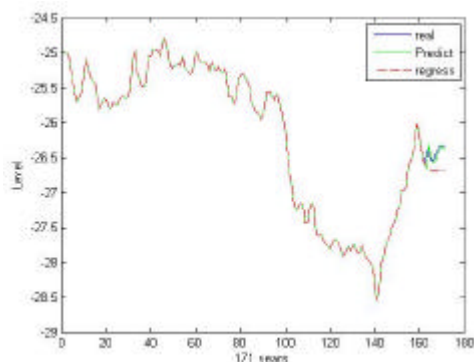


Fig. 4: Prediction of 10 values after the training set by regress model and fuzzy model

The original Caspian Sea water series (CSWS) is split into two subsets: a training set (first 161 samples) and a second set (last 10 samples) that will be used for validation.

A maximum model size of 8 parameters and a prediction horizon of 10 are considered.

As first stage within our methodology, Delta Test is performed on the training set for all the possible variable selections and the one with lowest Delta Test estimate is chosen. This process is performed independently for each prediction horizon. The number of selected variables is shown in Figure 2.

Second stage, is applying to the training set in order to identify fuzzy inference systems. These models are then tuned through supervised learning, to over the training set. As last step, the first (simplest) system that falls within the error range defined by the Delta Test is selected.

Figure 3 shows the training and validation errors of the fuzzy regressor model. Training and test errors of statistical regressor models are also shown. Statistical regressor models were built with the same fuzzy model size.

Figure 4 shows the predictions for the first 10 values after the training set together with a fragment of the actual Caspian Sea Level time series.

CONCLUSION

We have developed an automatic methodology framework for long-term Caspian Sea Level prediction by means of fuzzy inference systems. Experimental results for a concrete implementation of the methodology confirm good approximation accuracy and generalization capability.

Linguistic interpretability for both short-term and long-term prediction as well as low computational cost is two remarkable advantages over common time series prediction methods. Also, the proposed methodology has been shown to outperform based statistical regressor model predictions in terms of approximation accuracy.

REFERENCES

1. Bolgov, M.V., M.D. Trubetskova and M.K. Filimonova, 2004. On the Problem of the Caspian Sea Level Forecasting, Water Problems Institute, Russian Academy of Sciences, Russia.
2. Nicolai, A. and P. Igor, 2004. The Caspian Sea, Lake Basin Management Initiative, 2004.
3. Bolgov, M.V., 2005. Stochastic models in a problem of the Caspian sea level forecasting, Stochastic Hydraulics, Russia, 2005.
4. Box, G.P. and G.M. Jenkins, Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco, CA, 1976.
5. Man, K.S., 2003. Long memory time series and short term forecasts, International Journal of Forecasting, pp: 477-491.
6. Jan, G. and J. Rob, 2006. 25 years of time series forecasting, International Journal of Forecasting 22, 2006, pp: 443-473.
7. Lii, K.S. and M. Rosenblatt, 1993. Non-Gaussian autoregressive moving average processes. Applied Mathematics, pp: 9168-9170.
8. Wang, H. and N.F. Pan, 2008. A Fuzzy Regression Model for Predicting Non-Crisp Variable, Fifth International Conference on Fuzzy Systems and Knowledge Discovery, IEEE, pp: 104-106.

9. Pouzols, M., A. Lendasse and A. Barriga, 2008. Fuzzy Inference Based Autoregressors for Time Series Prediction Using Nonparametric Residual Variance Estimation. International Conference on Fuzzy Systems, pp: 613-618.
10. Jones, A.J. 2004. New Tools in Non-linear Modeling and Prediction, Computational Management Science, pp: 109-149, Sep. 2004.
11. Caspian Sea Level Database, National Centre for Caspian Studies and Research.