Prediction of Oil Production with: Data Mining, Neuro-Fuzzy and Linear Regression

Zahra Mahdavi , Maryam Khademi

***Abstract*—According to importance and usage of researches of petroleum production, our major goal in this paperis to forecast the oil production by using Data Mining Technique for improving estimation of oil consumption base of History of data. The implement of auto regression, Data cleaning and concept of pre-processing to viewpoint of Time series analysis are traditional concept in intelligent format. We use data cleaning for integration data and auto regression to determine input of model and for pre-processing for upgrade operation of ANFIS. ANFIS algorithm is developed by different data pre-processing methods and the efficiency of ANFIS is examined against auto regression (AR) in Canada. For this purpose, mean absolute percentage error (MAPE) is used to show the efficiency of ANFIS. However, it is concluded that ANFIS provides better results than AR in Canada. This is unlike previous expectations that ANFIS always provides better estimation than conventional approaches.**

***Index Terms*—Time series, neuro-fuzzy system, ANFIS, data mining, oil production forecasting.**

I. Introduction

Due to abrupt and uncontrolled changes in demand of countries, it is so critical for economic programmers to be able to estimate the required oil production appropriately for developing their future programs, finding a data structure that can estimate the future oil production with the least possible error is considered [1].

Adaptive neural network is inspired by neurobiology to perform brain-like computation. Network structure consists of a number of nodes connected through directional links. Each node represents a process unit, and the links between nodes specify the causal relationship between the connected nodes In addition, the Neuro-fuzzy approach combines the advantages of fuzzy logic and neural network models to design an architecture that uses a fuzzy logic to represent knowledge [2] in an interpretable manner and the learning ability of a neural network to optimize its parameters The Neuro-fuzzy approach has the advantage of reduced training time not only due to its smaller dimensions but also because the network can be initialized with parameters relating to the problem domain A specific approach in Neuro-fuzzy development is the adaptive Neuro-fuzzy inference system (ANFIS), which has shown significant results in the model nonlinear functions. The ANFIS learns features in the data set and adjusts the system parameters according to a given error criterion [3], [4].

All methods, except MAPE have scaled output. MAPE method is the most suitable method to estimate the relative error because input data used for the model estimation, pre-processed data and raw data have different scales, so we used MAPE for select best Neuro-fuzzy network [5].

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{x_t - x'}{x_t}\right|$$

This paper organized as follows: in section 2 we present the methods used for knowledge discovery.In section 3, we discuss why we choose Canada for model in our methodology and the paper will be ended by conclusions and future works in section 4.

II. Methods Used For Knowledge Discovery

Generally, methods for knowledge discovery can be divided into 5 categories:

- Classification: in this method, one sample is classified to one of predefined categories.
- Regression: predict one variable amount based on other variables.
- Clustering: one category of data, mapped out to one of clusters.
- Association roll: this method explains, dependency relationship between different features state.
- Analysis of sequence: this method models, sequence pattern, like time series.

In our methodology that will be explained in next section, we incorporate regression method and analysis of sequence method by using ANFIS, and we considering some points like: just write a computer program is not suffice for make a model, model will be useful when practical result achieve. And our model fault is relative, and must be proportional to aim and cost. So we choose Canada for compare our result to real result and evaluate our model fault better [6], [7].

III. Methodology

An algorithm is developed to model the time series process and analysis with analysis of sequence and ANFIS networks. Neuro-fuzzy model is considered to determine the impact of preprocessing on ANFIS model for estimation. So, there is one basic model in our algorithm. The proposed algorithm is applied to 4 data sets related to Canada We run all ANFIS network on Matlab; there are 91 rows of data (monthly oil consumption)from January 2003 to July 2010 for Canada [8]-[10].The steps of this algorithm are:

Step 1: Divide data into two sets, one for estimating the model called train data set and the other one for evaluating

Manuscript received March 15, 2012; revised May 10, 2012.

the validity of the estimated model called test data set. Usually, train data set contains 80% of all data and remained data (20%) are used in test data set. The data is divided into training data (63) and test data (6). Also, the preprocessing data is divided into training data and test data.

Step 2:for integration data and clean data, it is necessary to remove invalid data from train data, and make some dome for this, next we have to make static data, it means average and variance must be equal in time duration [11].

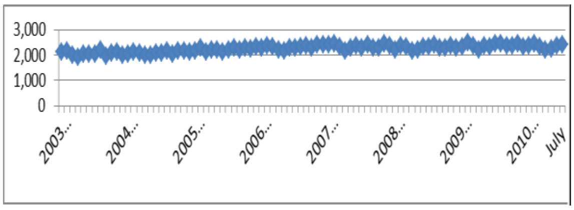


Fig. 1. Cleaning and static data

For the ANFIS model, input variables are selected by using autocorrelation function (ACF).

oD t=f(oDt-1,oDt-2)(2)

In Canada *oDt= 2.* It means, demand of oil in current month is depends on demand of oil on 2 months ago.

Step 3: make different model of Neuro-fuzzy network and evaluate each network operation. So our Neuro-fuzzy needs four follows arrays of data matrix to do this operation:

1) Test input
2) Test output
3) Train input
4) Train output

This model is run regarding to two main parameters in ANFIS. Type of membership function and number of membership functions associated with each input are mentioned parameters of ANFIS network.

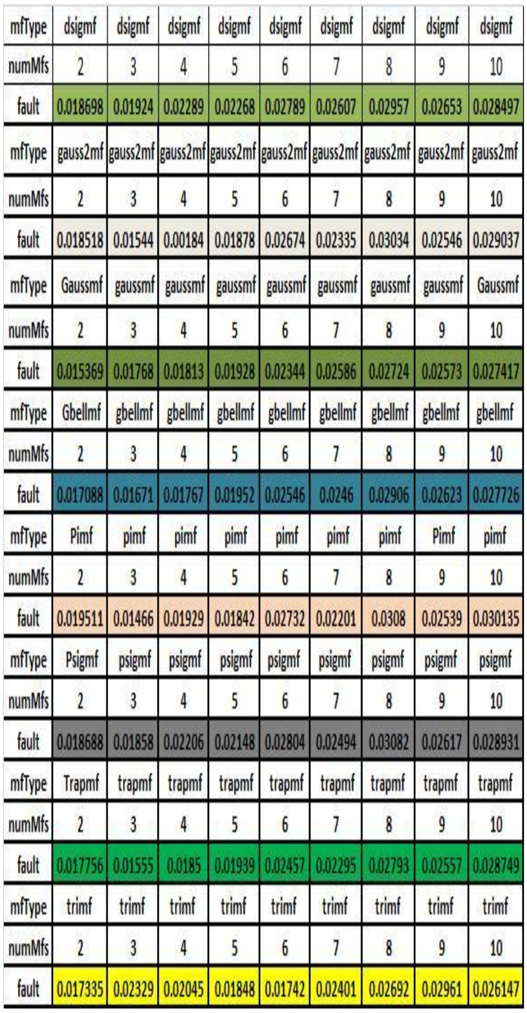
Step 4: ANFIS have been run for Canada. The value of MAPE for each network that has different at type of membership function and number of membership function associated with each input.

IV. Conclusion and Future Works

ANFIS algorithms can be considered in terms of simplicity, generality and applicable. In this paper an accurate method to forecast oil production is presented. The proposed efficient, simple and fast method works based on ANFIS algorithm that we clean and integrate data and then turn to static all data, then divide data to two categories (train data and test data) after this cerate different ANFIS network and train networks with train data. Finally, we evaluate test data to all train networks and choose the network that has lower error among all networks, and call it Best-network. In our case study, the Best-network has less than %5 Error. The proposed algorithms were implemented on 91 rows of data (monthly oil consumption) from January 2003 to July 2010 for Canada and the rate of accuracy is 95%.

In future, we can develop and design this algorithm for logistic Regression for many countries, and so can add new model and parameter like Economics parameter to improve our network for estimating and get closer to reality

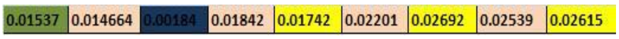
TABLE I: Different At Type of Membership Function and Number of Membership Function Associated with Each Input and Each Fault



mfType	dsigmf	dsigmf	dsigmf	dsigmf	dsigmf	dsigmf	dsigmf	dsigmf	dsigmf
numMfs	2	3	4	5	6	7	8	9	10
fault	0.018698	0.01924	0.02289	0.02268	0.02789	0.02607	0.02957	0.02653	0.028497
mfType	gauss2mf	gauss2mf	gauss2mf	gauss2mf	gauss2mf	gauss2mf	gauss2mf	gauss2mf	gauss2mf
numMfs	2	3	4	5	6	7	8	9	10
fault	0.018518	0.01544	0.00184	0.01878	0.02674	0.02335	0.03034	0.02546	0.029037
mfType	Gaussmf	gaussmf	gaussmf	gaussmf	gaussmf	gaussmf	gaussmf	gaussmf	Gaussmf
numMfs	2	3	4	5	6	7	8	9	10
fault	0.015369	0.01768	0.01813	0.01928	0.02344	0.02586	0.02724	0.02573	0.027417
mfType	Gbellmf	gbellmf	gbellmf	gbellmf	gbellmf	gbellmf	gbellmf	gbellmf	gbellmf
numMfs	2	3	4	5	6	7	8	9	10
fault	0.017088	0.01671	0.01767	0.01952	0.02546	0.0246	0.02906	0.02623	0.027726
mfType	Pimf	pimf	pimf	pimf	pimf	pimf	pimf	Pimf	pimf
numMfs	2	3	4	5	6	7	8	9	10
fault	0.019511	0.01466	0.01929	0.01842	0.02732	0.02201	0.0308	0.02539	0.030135
mfType	Psigmf	psigmf	psigmf	psigmf	psigmf	psigmf	psigmf	psigmf	psigmf
numMfs	2	3	4	5	6	7	8	9	10
fault	0.018688	0.01858	0.02206	0.02148	0.02804	0.02494	0.03082	0.02617	0.028931
mfType	Trapmf	trapmf	trapmf	trapmf	trapmf	trapmf	trapmf	trapmf	trapmf
numMfs	2	3	4	5	6	7	8	9	10
fault	0.017756	0.01555	0.0185	0.01939	0.02457	0.02295	0.02793	0.02557	0.028749
mfType	trimf	trimf	trimf	trimf	trimf	trimf	trimf	trimf	trimf
numMfs	2	3	4	5	6	7	8	9	10
fault	0.017335	0.02329	0.02045	0.01848	0.01742	0.02401	0.02692	0.02961	0.026147

Now we chose the column have less fault in Table 2.

TABLE II: Minimum Fault for Each Network



0.01537	0.014664	0.00184	0.01842	0.01742	0.02201	0.02692	0.02539	0.02615

As you see in Table 2, minimum fault is for gauss2mf with 4 numMfs, and now we can find the best function.

References

[1] A. SEU, Ministry *of Energy and Mineral Resources (MEMR)*, Amman, Jordan, 2008 ch.3.

[2] L. A. Zadeh, *Fuzzy Sets, Information and Control 8* pp.338–353, 1965.

[3] S. Subasi, "Prediction of Mechanical Properties of Cement Containing Class C Fly Ash by Using Artificial Neural Networkand Regression Technique," *Sci. Res. Essays*, vol. 4, no. 4, pp. 289-297, 2009.

[4] S. Terzi, "Modeling the payment serviceability ratio of flexible highway payments by artificial neural networks," *Construction Building Mater.*, vol. 21, pp. 590-593, 2007.

[5] H. Altun, A. Bilgil, and C. F. Fidan, "Treatment of Multi-dimensional data to enhance neural network estimators in regression problems," *Expert Syst. Appl.*, vol. 32, no. 2, pp.599-605, 2006.

[6] A. Yarar, M. Onucyıldız, and N. K. Copty, "Modeling level change in lakes using Neuro-fuzzy and artificial neural networks," *J. Hydrol*, vol. 365, no.3-4, pp. 329-334, 2009.

[7] R. Wieland and W. Mirschel, "Adaptive fuzzy modeling versus artificial neural networks," *Environmental Modeling and Software* vol. 23, pp. 215-224, 2008.

[8] Energy. Gov. [Online]. Available: http://www.doe.gov/

[9] I. Dincer and S. Dost, "Energy intensities for Canada," *Applied Energy*, vol. 53, 1996, 283-298. [10]Canada Oil- Consumption. [Online]. Available: http://www.indexmundi.com/canada/oil_consumption.html

[10] *Oil and Gas Journal*, July 13, 2003.