ISSN: 2229-6956(ONLINE)
DOI: 10.21917/ijsc.2014.0113

FORECASTING PETROLEUM PRODUCTION USING CHAOS TIME SERIES ANALYSIS AND FUZZY CLUSTERING

K.I. Jabbarova[1] and O.H. Huseynov[2]
Department of Computer-Aided Control Systems, Azerbaijan State Oil Academy, Azerbaijan
E-mail: [1]konul.jabbarova@mail.ru, [2]oleg_huseynov@yahoo.com

Abstract

Forecasting of petroleum production time series is a key task underlying scheduling of oil refinery production. In turn, forecasting requires analysing whether time series exhibits chaotic behavior. In this paper we consider chaos analysis based forecasting of time series of gasoline and diesel production. Chaos analysis is based on Lyapunov exponents and includes determination of optimal values of embedding dimension and time lag by using differential entropy approach. For forecasting of petroleum production, fuzzy "IF-THEN" rules constructed on the base of fuzzy clustering of the time series are used. The obtained prediction results show adequacy of the used methodology.

Keywords:

Chaos, Lyapunov Exponents, Embedding Dimension, Petroleum Production, Fuzzy "IF-THEN" Rules

1. INTRODUCTION

Scheduling of gasoline and diesel production in oil refining industry is subject to account for future demand. Therefore, prediction of these petroleum products becomes an important problem. In this paper we consider time series of gasoline and diesel production given in Fig.1 and Fig.2.

In order to forecast future values for gasoline and diesel production on the base of historical data, analysis of properties of the considered time series is needed. The latter may exhibit nonlinear chaotic behavior and in this case the use of linear and traditional nonlinear models for prediction is not suitable. In [1] they consider forecasting of oil prices dynamics of which is very complex and may exhibit chaotic behavior due to political, technological and other influential factors. In order to uncover and analyze the chaotic behavior of the considered dynamics, the formalism of Lyapunov exponents and embedding dimension is used. Given the results indicating chaos in the oil prices time series, the authors use an artificial neural network (ANN) to forecast future oil prices. The obtained results show increased accuracy of the ANN-based prediction as compared with other existing approaches. [2] is devoted to development of a hybrid approach to oil prices forecasting on the basis of systematic analysis of the considered time series. The conducted analysis includes variance ratio test, Brock-Dechert-Scheinkman (BDS) test, chaos analysis and other approaches. The used chaos analysis utilizes Lyapunov exponents. The suggested hybrid approach to forecasting is characterized by an increased performance as compared to the existing approach. In [3] prediction of petroleum prices on the base of non-linear dynamical theory is considered. The suggested research is based on Lyapunov exponents, phase space reconstruction and other approaches. The obtained results of prediction show adequacy of the constructed nonlinear forecasting model.

In [4] they apply the differential entropy based approach to simultaneously determine optimal embedding dimension and time lag in chaotic time series. The used approach utilizes so-called "entropy ratio" between the phase space representation of time series and an ensemble of its surrogates. The suggested approach was tested on several benchmark real-world time series. The obtained forecasting results show better performance of the suggested approach as compared to the existing methods.

[5] is devoted to forecasting of oil and gas spot prices. The authors use Lyapunov exponents to uncover chaotic dynamics in the considered time series. Given chaotic dynamics uncovered, an optimal embedding dimension, time delay and predictability are obtained with the aid of minimization of the root mean square error. Next, embedding dimension and time delay are considered as inputs to fuzzy neural network used as a nonlinear forecasting model. The forecasting results obtained by using fuzzy neural network provide low value of forecasting error.

We can conclude that chaotic time series analysis is investigated fundamentally in the existing literature. The overview of literature devoted to chaotic time series shows that there are different methods to identify whether considered time series exhibit chaotic behavior. If chaotic behavior is present, various methods can be applied to investigate its characteristics and use them for forecasting of chaotic time series. Unfortunately, the existing literature on application and practical validation of the existing theoretical approaches is scarce.

In this paper we consider forecasting of the time series of gasoline and diesel production on the basis of analysis of chaotic behavior. Given the chaotic behavior uncovered, optimal values of embedding dimension and time lag should be found which characterize the pattern of chaotic dynamics. We use the differential entropy to find optimal values of embedding dimension and time lag simultaneously. Given the results of the analysis, we use fuzzy clustering to construct fuzzy "IF-THEN" rules fuzzy "IF-THEN" rules which are able to provide an intuitive description of time series dynamics. Then the prediction is made on the base of fuzzy reasoning within the considered fuzzy "IF-THEN" rules.

2. STATEMENT OF PROBLEM

The considered time series of gasoline and diesel production are given in Fig.1 and Fig.2 respectively. The graphs represent per day dynamics of the considered products in tons during a month.

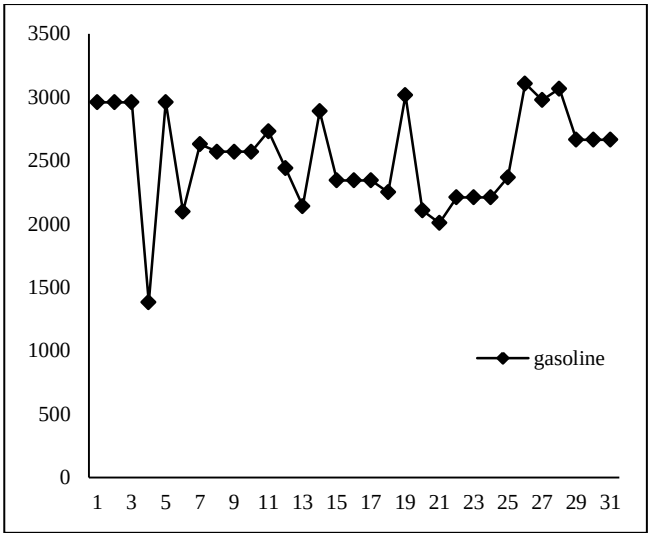

Fig.1. Gasoline production dynamics

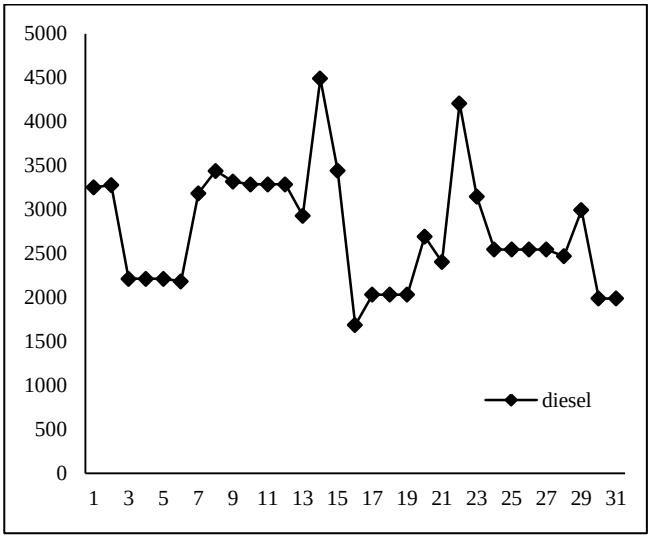

Fig.2. Diesel fuel production dynamics

For the time series given in Fig.1 and Fig.2, it is needed to determine an optimal value of *k*, and then to find such *f* that the computation of future amount of gasoline production denoted y_{n+1},

$$y_{n+1} = f(y_n, y_{n-1},\ldots,y_{n-k+1})$$

will minimize the forecasting RMSE error.

3. SOLUTION METHODOLOGY

Let us consider first forecasting of time series for gasoline production. Forecasting of diesel production is conducted analogously. At the first step we need to verify whether the considered time series exhibit chaotic behavior. This analysis will be conducted with the aid of local Lyapunov exponents which show sensitivity of time series dynamics to initial conditions. The Lyapunov exponents are determined as follows [6].

$$\lambda = \lim_{n\to\infty}\frac{1}{n}\lim_{\left|x_0-\bar{x}_0\right|}\log_2\frac{\left|x_0-\bar{x}_0\right|}{\left|x_n-\bar{x}_n\right|},$$

where, x_0 and $\bar{x}_0$ are close initial conditions of a time series, and x_n and $\bar{x}_n$ are values obtained after *n* successive iterations. A negative and zero values of Lyapunov exponent shows absence of chaotic behavior. However, positive value of Lyapunov exponent $\lambda > 0$ indicates chaotic behavior in time series. For $\lambda > 0$, the ratio of the distance $\left|x_n-\bar{x}_n\right|$ between x_n and $\bar{x}_n$ obtained after *n* successive iterations to the distance between initial conditions $\left|x_0-\bar{x}_0\right|$ increases exponentially. In other words, this means that trajectories with even close initial conditions will exhibit exponential divergence.

For the considered time series of gasoline production, the Lyapunov exponent is found as 0.48 which indicates chaotic behavior. Therefore it is now needed to determine optimal values of embedding dimension *m* and time lag τ. In several approaches, optimal values of these two parameters m_{opt} and τ_{opt} are found separately. Often, first they find optimal time lag τ_{opt} and then determine optimal embedding dimension m_{opt}. In this paper, we will find optimal values of embedding dimension and time lag simultaneously. For this purpose we will use differential entropy criterion [4, 7]:

$$H(x)=\int_{-\infty}^{+\infty}p(x)\ln p(x)dx$$

One of the most convenient way to compute differential entropy from practical point of view is the Kozachenko-Leonenko (*K-L*) estimate [4, 7]:

$$H(x)=\sum_{j=1}^{N}\ln(Np_j)+\ln 2+C_{E,} \qquad (1)$$

where, *N* is the number of samples in the data set, ρ_j is the Euclidean distance of the j^{th} delay vector to its nearest neighbor, and $C_E \approx 0.5772$ is the Euler constant. One advantage of *K-L* estimate is its flexibility with respect to the dimensionality of the data set.

Let us now consider $H(x,m,\tau)$ which is the differential entropies estimated for time delay embedded versions of a time series, *x*, and is used as an inverse measure of the structure in the phase space.

The optimal values m_{opt} and τ_{opt} define such representation of phase space which best reflects the dynamics of the chaotic time series. Formally, m_{opt} and τ_{opt} are values at which differential entropy *H* takes its minimal value (minimal disorder). Thus, to find m_{opt} and τ_{opt} we should minimize Eq.(1).

The *K-L* estimates for the considered time series $H(x,m,\tau)$, and the surrogates $H(x_s,m,\tau)$ are computed using Eq.(1). In order to find optimal values m_{opt} and τ_{opt} the following ratio should be minimized:

$$I(m,\tau)=\frac{H(x,m,\tau)}{\langle H(x_s,i,m,\tau)\rangle_i} \qquad (2)$$

where, $\langle\rangle_i$ denotes the average over *i*. To escape dealing with higher embedding dimensions, the minimum description length (MDL) method is used with the "entropy ratio" (ER):

$$R_{ent(m,\tau)}=I(m,\tau)+\frac{m\ln N}{N},$$

where, N is the number of delay vectors, which is kept constant for all values of m and τ under consideration.

The values m_{opt} and τ_{opt} for gasoline production which constitute minimum of $R_{ent}(m, \tau)$ are found as $m_{opt} = 3$ and $\tau_{opt} = 1$. The plot of $R_{ent}(m, \tau)$ is shown in Fig.3.

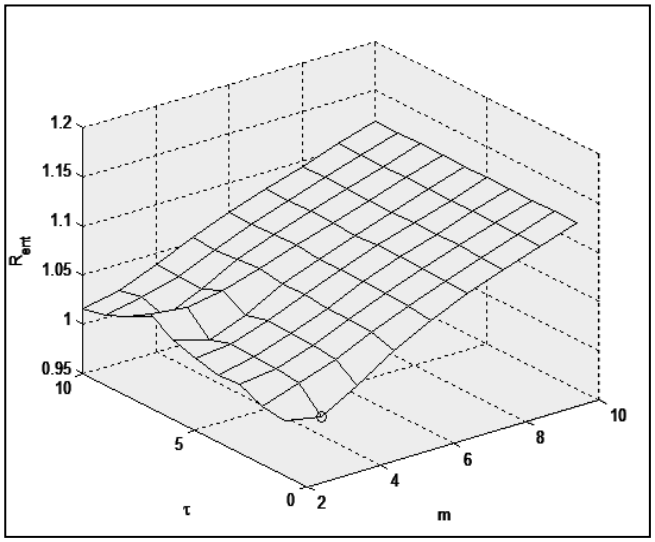

Fig.3. The plot of $R_{ent}(m, \tau)$ with the local minimum at $m_{opt} = 3$ and $\tau_{opt} = 1$

Given m_{opt} and τ_{opt}, let us now proceed to solving the problem of forecasting of the time series of gasoline production stated in section 2. For solving of this problem, we will construct fuzzy "IF-THEN" rules to arrive at an intuitive linguistic description of the considered time series dynamics. Fuzzy "IF-THEN" rules will be constructed on the basis of fuzzy clustering of the considered time series by means of Fuzzy C-means clustering method outlined below.

The problem of fuzzy clustering on the base of Fuzzy C-means approach consists in partition of the considered time series $X = \{x_1, x_2,.., x_n\}$ into c fuzzy clusters such that the following criterion is minimized:

$$J_m = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{m}\left\|x_i - v_j\right\|^2 \to min \qquad (3)$$

subject to

$$0 < \sum_{i=1}^{n} u_{ij} < n\left(j = \overline{1,c}\right) \text{ and } \sum_{i=1}^{c} u_{ij} = 1\left(i = \overline{1,n}\right) \qquad (4)$$

where, $\|*\|$– Euclid distance, c is the number of clusters given in advance, m is the value of fuzzifier which defines curvature of membership functions of obtained clusters, v_j are coordinates of centers of clusters to be found.

As a result of solving problem Eq.(3)-Eq.(4), centers v_j and membership functions μ_j of fuzzy clusters are obtained.

We have solved the considered problem of fuzzy clustering of gasoline production time series for c = 3, 4, 5, 8. The values of v_j and the membership degrees $\mu_j()$ are shown in Table.1 and Table.2 respectively.

Table.1. The centers of the obtained fuzzy clusters

c = 3

cluster_1	cluster_2	cluster_3
2091.4	2785.9	2304.3
2620.9	2550.1	2462.2
2659.5	2103.6	2787.8

c = 4

cluster_1	cluster_2	cluster_3	cluster_4
2650.9	2499.1	2134.8	2832.6
2926.6	1551.6	2918.3	2113.4
2111.7	2780.6	2373.3	2426.6
2609.0	2516.9	2576.6	2424.9

c = 5

cluster_1	cluster _2	cluster _3	cluster _4	cluster _5
2558.0	2656.6	2474.3	2149.3	2864.0
2942.1	1469.3	2937.5	2109.0	2626.6
2761.4	2560.0	2028.6	2832.9	2311.0
2626.4	2457.1	2610.3	2497.1	2405.6
2046.9	2805.1	2366.5	2453.0	2463.7

Table.2. The membership degrees of the obtained fuzzy clusters

c = 3

cluster_1	cluster_2	cluster_3
0.2122	0.4857	0.3021
0.3687	0.4160	0.2153
0.1289	0.2355	0.6355
0.6596	0.2069	0.1335

c = 4

cluster_1	cluster_2	cluster_3	cluster_4
0.1776	0.2178	0.2511	0.3535
0.4868	0.0856	0.2145	0.2131
0.0119	0.9601	0.0097	0.0183
0.1987	0.0697	0.5438	0.1878

c = 5

cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
0.3799	0.1442	0.1149	0.2026	0.1585
0.1267	0.0604	0.5028	0.1563	0.1538
0.0040	0.9854	0.0029	0.0052	0.0025
0.1543	0.0529	0.1372	0.1415	0.5141

Each of obtained fuzzy clusters represents one fuzzy rule, and therefore the value of c represents the number of fuzzy "IF-THEN" rules which have the following form:

- If X_{t-m} is A_{11} and X_{t-m+1} is $A_{12},\ldots,$ and X_{t-1} is A_{1m} then X_t is B_1
- If X_{t-m} is A_{21} and $X_{t-\tau+1}$ is $A_{22},\ldots,$ and X_{t-1} is $A_{2\tau}$ then X_t is B_2

.
.
.

- If $X_{t-\tau}$ is A_{i1} and $X_{t-\tau+1}$ is $A_{i2},\ldots,$ and X_{t-1} is $A_{i\tau}$ then X_t is B_i

.
.
.

- If $X_{t-\tau}$ is A_{c1} and $X_{t-\tau+1}$ is $A_{c2},\ldots,$ and X_{t-1} is $A_{c\tau}$ then X_t is B_c

The forecasting of X_t is then implemented on the basis of Mamdani inference approach within the above fuzzy "IF-THEN" rules base. We have considered forecasting for rule base with c = 3, 4, 5 rules and compared the corresponding values of forecasting error. As forecasting error RMSE is used:

$$RMSE=\sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{N-1}}$$

The results are shown in Table.3. Training and testing sets are chosen as 2/3and 1/3 parts respectively of the considered time series.

Table.3. Forecasting errors for c = 3, 4, 5

Number of clusters	Training error (RMSE)	Testing error (RMSE)
c = 3	224.59	416.05
c = 4	215.75	449.73
c = 5	197.40	446.81

For diesel fuel production time series, Lyapunov exponent is found as 0.046, optimal values of m and τ are m_{opt} = 3 and τ_{opt} = 1. Forecasting errors for c = 3, 4, 5,6, 8 are given in Table.4:

Table.4. Forecasting errors for c = 3, 4, 5, 6, 8

Number of clusters	Training error (RMSE)	Testing error (RMSE)
c = 3	213.08	517.04
c = 4	225.6	462.3
c = 5	216.5	426.9
c = 6	204.4	611.2
c = 8	199.2	953.7

The optimal value of fuzzy "IF-THEN" rules is c = 5.

4. COMPUTER SIMULATION

The best results for gasoline production time series are obtained under c = 3. Thus, the following fuzzy "IF-THEN" rules base should be used for forecasting:

- If X_{t-2} is *LOW* and X_{t-1} is *VERY HIGH* then X_t is *LOW*
- If X_{t-2} is *HIGH* and X_{t-1} is *HIGH* then X_t is *HIGH*
- If X_{t-2} is *HIGH* and X_{t-1} is *LOW* then X_t is *HIGH*

For diesel production time series, the best results are under c = 5. Thus, the following fuzzy "IF-THEN" rules base should be used for forecasting:

- If X_{t-4} is *HIGH* and X_{t-3} is *MEDIUM* X_{t-2} is *HIGH* and X_{t-1} is *VERY LOW* then X_t is *VERY LOW*
- If X_{t-4} is *HIGH* and X_{t-3} is *HIGH* X_{t-2} is *HIGH* and X_{t-1} is *LOW* then X_t is *MEDIUM*
- If X_{t-4} is *LOW* and X_{t-3} is *LOW* X_{t-2} is *LOW* and X_{t-1} is *LOW* then X_t is *HIGH*
- If X_{t-4} is *MEDIUM* and X_{t-3} is *LOW* X_{t-2} is *VERY LOW* and X_{t-1} is *HIGH then* X_t is *VERY LOW*
- If X_{t-4} is *MEDIUM* and X_{t-3} is *LOW* X_{t-2} is *LOW* and X_{t-1} is *VERY LOW* then X_t is *HIGH*

The forecasting is made on the base of reasoning within the above given IF-THEN rules.

The forecasted and actual trends for gasoline production are shown in Fig.4.

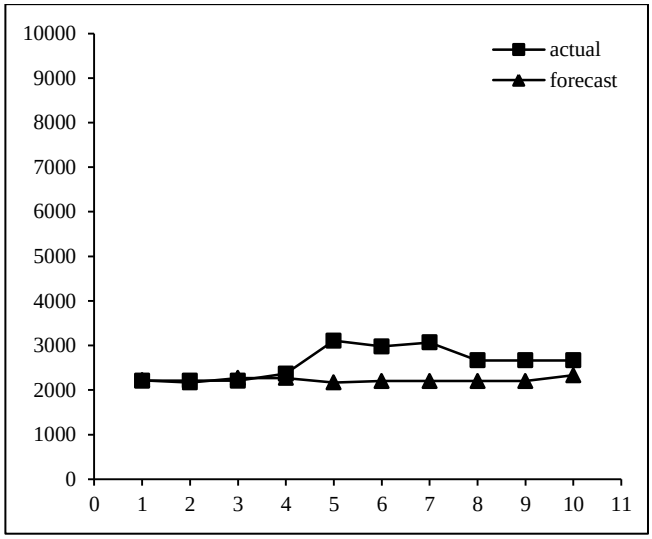

Fig.4. Forecasted and actual trends for gasoline production

The forecasted and actual trends for diesel fuel production are shown in Fig.5.

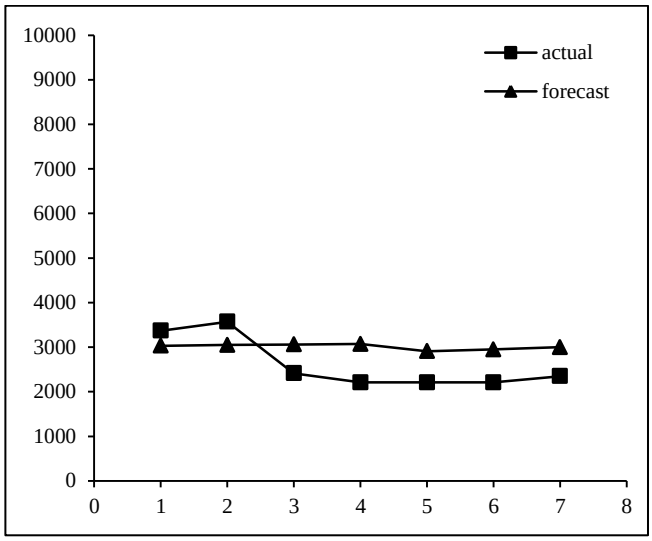

Fig.5. Forecasted and actual trends for diesel production

5. CONCLUSION

Forecasting of gasoline production is one of the necessary problems underlying planning of oil refinery production. An important issue related to this problem is analysis of gasoline production time series to find out whether it exhibits chaotic behavior. In this paper we uncover chaotic behavior in the time series of gasoline production by means of Lyapunov exponents and then compute the optimal values of embedding dimension and time lag which characterize the uncovered chaos. For forecasting of future amount of gasoline production, we use fuzzy "IF-THEN" rules constructed on the base of fuzzy clustering of the considered time series. The optimal number of fuzzy rules for forecasting is found and the obtained prediction results show adequacy of the used methodology.

REFERENCES

[1] Saeed Moshiri and Faezeh Foroutan, "Testing for deterministic chaos in futures crude oil price: Does neural network lead to better forecast?", *38th Annual Meeting of the Canadian Economics Association*, pp. 1-23, Available at: http://economics.ca/2004/papers/0077.pdf, 2004.

[2] Akbar Komijani, Esmaeil Naderi, Nadiya and Gandali Alikhani, "Hybrid Approach for Forecasting of Oil Prices Volatility", *Munich Personal RePEc Archive*, Available at: http://mpra.ub.uni-muenchen.de/44654/, 2013.

[3] Liu Lixia, "Nonlinear Test and Forecasting of Petroleum Futures Prices Time Series", *International Conference on Energy, Environment and Development*, Vol. 5, pp. 754-758, 2011.

[4] J. Beirlant, E. J. Dudewicz, L. Györfi and E. C. van der Meulen, "Nonparametric entropy estimation: An overview", *International Journal on Mathematical and Statistical Sciences*, Vol. 6, pp. 17-39, 1997.

[5] I. S. Agbon and J. C. Araque, "Predicting Oil and Gas Spot Prices Using Chaos Time Series Analysis and Fuzzy Neural Network Model", *Society of Petroleum Engineers Hydrocarbon Economics and Evaluation Symposium*, pp. 1-8, Available at: http: //wenku.baidu.com /view/ 3924ac39376baf1ffc4fadef, 2003.

[6] Alan Wolf, Jack B. Swift, Harry L. Swinney and John A. Vastano, "Determining Lyapunov exponents from a time series", *Physica*, 16 *D*, pp. 285-317, 1985.

[7] T. Gautama, D. P. Mandic, Van Hulle and M. M. Van Hulle, "A differential entropy based method for determining the optimal embedding parameters of a signal", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 6, pp. 29-32, 2003.