A Novel Approach to Gasoline Price Forecasting Based on Karhunen-Loève Transform and Network for Vector Quantization with Voronoid Polyhedral

Haruna Chiroma, Sameem Abdulkareem, Adamu I. Abubakar, Eka Novita Sari, Tutut Herawan

A Novel Approach to Gasoline Price Forecasting based on Karhunen-Loève Transform and Network for Vector Quantization with Voronoid Polyhedral

Haruna Chiroma[1], Sameem Abdulkareem[1], Adamu I. Abubakar[2], Eka Novita Sari[3] and Tutut Herawan[4]

[1]Department of Artificial Intelligence
[4]Department of Information systems
University of Malaya
50603 Pantai Valley, Kuala Lumpur, Malaysia
[2]Department of Information system
International Islamic University
Gombak, Kuala Lumpur, Malaysia
[3]AMCS Research Center
Yogyakarta, Indonesia

freedonchi@yahoo.com, 100adamu@gmail.com
{sameem,tutut}@um.edu.my, eka@amcs.co

Abstract. We propose an intelligent approach to gasoline price forecasting as an alternative to the statistical and econometric approaches typically applied in the literature. The linear nature of the statistics and Econometrics models assume normal distribution for input data which makes it unsuitable for forecasting nonlinear, and volatile gasoline price. Karhunen-Loève Transform and Network for Vector Quantization (KLNVQ) is proposed to build a model for the forecasting of gasoline prices. Experimental findings indicated that the proposed KLNVQ outperforms Autoregressive Integrated Moving Average, multiple linear regression, and vector autoregression model. The KLNVQ model constitutes an alternative to the forecasting of gasoline prices and the method has added to methods propose in the literature. Accurate forecasting of gasoline price has implication for the formulation of policies that can help deviate from the hardship of gasoline shortage.

Keywords: Vector quantization; Gasoline price; Karhunen-Loève Transform.

1 Introduction

Gasoline is obtained from the refining of crude oil through a process called fractional distillation. The gasoline is highly utilized by the general public for daily activities. Shortage of gasoline can inflict pain on communities, especially in US where long queues of vehicles were typically experienced when the gasoline is scarce. Similarly, the shortage of gasoline causes cuts in motoring, reduction of work weeks, and threats of job cuts in automobile industries [1]. Forecasting that is of low quality contributes to poor or inadequate investment which in turn might result in losses in welfare.

When forecasting is bias, it makes investors not to make a cost effective investment in the energy efficiency [2].

There are studies in the literature that forecast gasoline price in order to provide advance knowledge of the price so that its negative impact can be reduced successfully. For example, asymmetric and symmetric models were applied to build a model for the forecasting of the gasoline price. It was found that the regression results emanated from the study suggested asymmetric model performs better than the symmetric in out of sample forecast accuracy [3]. The linear trend correlation error was used to build a model for the forecasting of gasoline price [4]. Similarly, Autoregressive Integrated Moving Average (ARIMA) was used for the forecasting of gasoline price by [5]. Anderson *et al*. [2] used the Michigan Consumers Survey data to forecast the price of Gasoline. The results indicated that the forecast accuracy outperform the Econometrics Auto Regressive Moving Average (ARMA). The studies in the literature mainly focus on statistics and econometric models for the forecasting of the gasoline price. However, those models assume linear distribution for input data, whereas gasoline price is nonlinear and volatile. Therefore, those models cannot provide effective solution to the problem of gasoline price forecasting. In addition, experimental evidence documented in the literature shows that artificial intelligence techniques such as neural network, fuzzy logic, expert systems, Genetic algorithms provide better solution for forecasting than statistical and econometric models. Though, there are very few instances in which statistical tools perform better the artificial intelligence methodologies [6]. Despite the significance of gasoline price, we have not found a reference in the literature that investigates the forecast of gasoline prices using artificial intelligence techniques.

In this paper, we propose to forecast the price of gasoline based on the network for vector quantization with Voronoid Polyhedral, and Karhunen-Loève Transform (KL) to select the most relevant inputs and eliminate the irrelevant attributes to improve forecast accuracy.

The rest of this paper is organized as follows. Section 2 describes the theoretical background of the study. Section 3 describes the experimentations. Section 4 describes results and discussion. Finally the conclusion of this work is described in Section 5.

2 Theoretical Background of the Study

2.1 Karhunen-Loève Transform

Suppose that the vector Y is for random variables V and the point of interest is covariance or correlation of V. The approach here is to find the most valuable subset of V (<< V) that have the most relevant information given by these variances and co-variances or correlations. The KLT mainly concentrates on variants, although correlations and co-variances are not totally ignored. Let x_1, x_2, and x_3 be variables as given by equation (1)

$$\left(X = x_1, x_2, x_3, \cdots, x_V\right). \tag{1}$$

$$\alpha_1' X. \tag{2}$$

The linear function of X is given by Equation (2) in which α_1 is a vector of V constants as given in Equation (3)

$$\alpha_{11}, \alpha_{12}, \alpha_{13}, \cdots, \alpha_{1V}, \tag{3}$$

where α_1' is the transpose of α_1 therefore,

$$\alpha_1' X = \alpha_{11} x_1 + \alpha_{12} x_2 + \alpha_{13} x_3 + \cdots + \alpha_{1V} x_V = \sum_{i=1}^{V} \alpha_{1j} x_j. \tag{4}$$

$$\alpha_2' X = \alpha_{21} x_1 + \alpha_{22} x_2 + \alpha_{23} x_3 + \cdots + \alpha_{2V} x_V = \sum_{i=1}^{V} \alpha_{2j} x_j. \tag{5}$$

$$\alpha_3' X = \alpha_{31} x_1 + \alpha_{32} x_2 + \alpha_{33} x_3 + \cdots + \alpha_{3V} x_V = \sum_{i=1}^{V} \alpha_{3j} x_j. \tag{6}$$

Eqs. (1), (2) and (3) are uncorrelated having maximum variance up to Equation (7) so that in step *kth* a linear function given in Equation (8) with maximum variance subject to being uncorrelated with Equation (9)

$$\alpha_V' X. \tag{7}$$

$$\alpha_k' X. \tag{8}$$

$$\alpha_1' X, \alpha_2' X, \alpha_3', \cdots, \alpha_{k-1}' X. \tag{9}$$

The principal component *kth* is the Equation (8), as such the principal component can be found up to V principal components. In this way it is expected that the highest number of variations in X will account for m principal components, where m << V [7-8].

2.2 Network for Vector Quantization

Vector quantization (VQ) methods encode a manifold such that sub manifold $V \subseteq \Re^D$, use a finite set $w = (w_1, w_2, \cdots, w_n)$ of reference (codebook) or cluster centers. $w_i \in \Re^D$, $i = 1, \cdots, N$ vector of a data $\nu \in V$ can best be defined as the optimal matched or wining reference vector $w_{i(\nu)}$ of w for $d(\nu, w_{i(\nu)})$ which is a distortion error, for instance, square error $\left\| \nu - w_{i(\nu)} \right\|^2$ is minimum as possible. The procedure further divides the V into smaller units of regions

$$V_i = \left\{ \nu \in V \middle| \left\| \nu - w_i \right\| \leq \left\| \nu - w_j \right\| \forall_j \right\} \tag{10}$$

Equation (10) is referred to as voronoid polyhedral through which every data vector v corresponds to reference vector w_i if probability of the data vectors over V is defined by $p(v)$, then Equation (11) is the average of reconstruction error (distortion error)

$$E = \int d^D v p(v)(v - w_i)^2. \tag{11}$$

Equation (11) is optimized to the minimal by the optimum selection of the w_i (reference vector).

Theorem 1. *For a set of reference vectors* $w = (w_1,...w_N)$, $w_i \in \Re^D$, *and a density distribution* $p(v)$ *of data points* $v \in \Re^D$ *over the input space* $V \subseteq \Re^D$, *then*

$$\int_v d^D v p(v) h\lambda(k_i(v,w))(v - w_i) = -\frac{\partial E}{\partial w_i}, \tag{12}$$

$$E = \frac{1}{2}\sum_{j=1}^{N} d^D v p(v) h\lambda(k_j(v,w))(v - w_j)^2, \tag{13}$$

where $k_j(v,w)$ *represent the number of reference vectors* w_i *with* $\|v - w_i\| < \|v - w_j\|$.

By substituting $d_i(v) = v - w_i$ for ease we obtain

$$-\frac{\partial E}{\partial w_i} = R_i + \int_v d^D v p(v) h\lambda(k_i(v,w))(v - w_i). \tag{14}$$

$$R_i = -\frac{1}{2}\sum_{j=1}^{N}\int d^D v p(v) h'\lambda(k_j(v,w)) d_j^2 \frac{\partial k_j(v,w)}{\partial w_i}. \tag{15}$$

The derivative of $h'\lambda(\bullet)$ is $h\lambda(\bullet)$ $\forall_i = 1,...,N$ R_i vanishes for $k_j(v,w)$ then

$$k_j(v,w) = \sum_{i=1}^{N} \theta(d_j^2 - d_i^2). \tag{16}$$

Equation (16) is valid with the heavy-side step function $\theta(\bullet)$

$$\theta(x) = \begin{cases} 1, & for \quad x>0 \\ 0, & for \quad x\leq 0 \end{cases}. \tag{17}$$

$$\theta(x) = 0 \quad for \quad x \neq 0. \tag{18}$$

$$R_i = \int_v d^D v p(x) h' \lambda(k_i(v,w)) d_i^2 d_i \sum_{l=1}^{N} \partial(d_j^2 - d_l^2). \tag{19}$$

$$-\sum_{j=1}^{N} \int_v d^D v p(x) h' \lambda(k_j(v,w)) d_j^2 d_i \partial(d_j^2 - d_i^2). \tag{20}$$

The integral of the N integrands in the 2nd term of Equation (20) is non-vanishing only for those in which $d_j^2 \approx d_i^2$. Those $v's$ can be defined as

$$k_j(v,w) = \sum_{l=1}^{N} \theta(d_j^2 - d_l^2) = \sum_{l=1}^{N} \theta(d_i^2 - d_l^2) = k_i(v,w). \tag{21}$$

Hence, $R_i = \int_v d^D v p(v) h' \lambda(k_i(v,w)) d_i^2 d_i \sum_{l=1}^{N} \partial(d_i^2 - d_l^2)$. (22)

$$-\int_v d^D v p(v) h' \lambda(k_i(v,w)) d_i^2 d_i \sum_{j=1}^{N} \partial(d_j^2 - d_i^2). \tag{23}$$

R_i vanishes $\forall i = 1, \cdots, N$ because $\partial(x) = \partial(-x)$. If the priori of the data point distribution $(p(v))$ is not given whereas stochastic sequence of input data points $v(t=1)$, $v(t=2)$, … govern by *P(x)* drives the adaptation procedure, adjusting steps for the reference vectors or cluster centers w_i is defined by

$$\Delta w_i = \in . \partial_{ii}(v(t)).(v(t) - w_i). \tag{24}$$

where $\in$ and ∂_{ij} are the step size and Kronecker delta, respectively.

Due to Equation (11) having many local minima, soft max is introduced into the learning process to adjust the winning reference vector *i(x)* as shown in Equation (25) to prevent the network for vector quantization from being trapped in local minima.

$$\Delta w_i = \in . \frac{\ell^{-\beta(v-w_i)}}{\sum_{j=1}^{N} \ell^{-\beta(v-w_i)^2}} . (v - w_i). \tag{25}$$

$$E_{\min} = -\frac{1}{\beta}\int d^{D}v p(v) \ln \sum_{i=1}^{N} \ell^{-\beta(v-w_i)^2} . \quad (26)$$

Equation (25) corresponds to stochastic gradient decent on the cost function. The cost function $(E_{\min})$ as given in Equation (26) is equivalent to E of Equation (11) [9].

3 Experiments

3.1 Dataset

The data of New York Harbor Conventional Gasoline Regular Sport Price FOB ($/Gallon) were collected from the Energy Information Administration of the US Department of Energy. The data are freely available through the official website of the organization. The data were collected on a daily frequency from 2 of Jun, 1986 to October 15, 12012. Fig. 1 depicted the window of the original gasoline price data clearly showing its nonlinearity and volatile nature. The dataset consist of 6639 observations within the period under study, weekends and other public holidays created empty spaces of which we have used imputation to fill in the empty spaces. The value of 0 was used to fill the empty spaces since the dataset were normalized within the range of -1 to 1. Therefore, adding zero to empty spaces cannot affect the results because they are approximately equal proportion. Future contract prices (price of gasoline agreed between two parties today, but payment to be made on a specific future date) determine the gasoline price as pointed out in [5]. Therefore, fifteen futures contract prices were collected as the independent attributes, whereas gasoline price is the dependent variable. Descriptive statistics of the time series are reported in Table 1 showing the minimum and maximum gasoline price for the period under study. The standard deviation suggests that the gasoline price data are in good agreement with the observations.

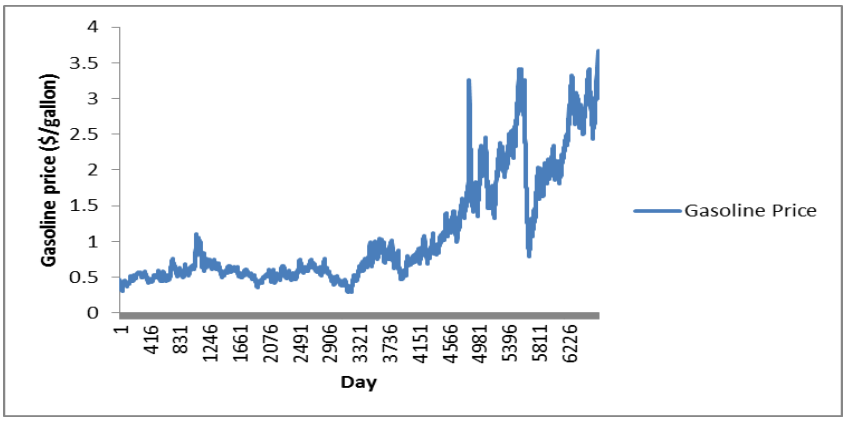

Fig. 1. Window of the gasoline price

Table 1 Descriptive statistics of the gasoline price

				Mean		SD
	Observations	Min	Max	Statistics	Standard error	Statistics
Gasoline price	6637	0.29	3.67	1.0901	0.00959	0.7811

3.2 The Proposed Application of the Network for Vector Quantization

The KLT is applied to reduce the number of input attributes in order to eliminate irrelevant inputs and used only minimum and relevant attributes. Including irrelevant attributes in the modeling process could affect the model performance, accuracy and increase the complexity of the network. The data of the gasoline price are partitioned into training, validation and testing dataset in the ration of 80%: 10%: 10% after several trials, since there is no ideal way for determining the exact percentage ratio. In building an NVQ model, initial parameter selection is critical to the performance of the model. For our study, we conducted several initial experimentation for choosing the optimal values that can yield global or near global solution. Number of adaptation steps, w_i, minimal distribution error E_0, performance measure were all set at the beginning of our experimentations.

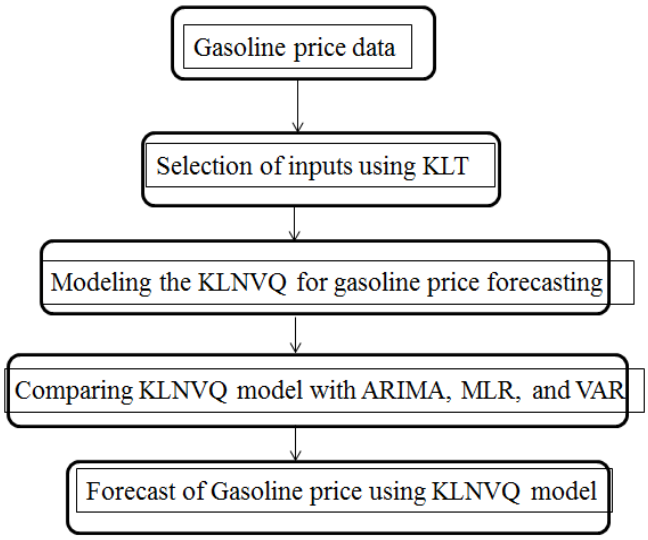

Fig. 2. The propose conceptual framework for the KLNVQ model

The KLNVQ model builds in this study was used to forecast the prices of gasoline. For the purpose of comparison, ARIMA, Multiple Linear Regression (MLR) and Vector Auto-regression (VAR) models were also used to forecast the gasoline price. The entire process of building the KLNVQ model is presented in Fig. 2 and it was implemented in MATLAB (2013a) neural network ToolBox and SPSS version 16 on

a machine (HP L1750 model, 4Gb RAM, 232.4 GB HDD, 32- bit OS, Intel (R) Core (TM)2 Duo CPU @ 3.00 GHz).

4 Results and Discussion

4.1 Analysis of the results

We use KLT for the selection of input attributes in order to reduce its dimension. The attributes selected for the study are future contract 1 = 21.7, future contract 2 = 16.31, future contract 3 = 11.61, futures contract 4 = 9.4, future contract 5 = 8.7, future contract 6 = 7.2, future contract 7 = 6.5, future contract 8 = 6.1, and future contract 8 = 5.55 accounts for 93.07% cumulative variance. Others were rejected for inclusion in the model, therefore, six attributes were not included. The KLNVQ model has nine (9) inputs neurons, four hidden layer neurons, one (1) output neuron, soft max is used during adaptation. Number of adaptation steps = 0.2, w_i = 0.04, minimal distribution error E_0 = 0.0012. These are the optimal parameters for the KLNVQ model. The performance of the model for training, validation, test and complete datasets is presented in Fig. 3. The Mean Square Error (MSE) for training, validation and out of sample test are 0.006241, 0.006134, and 0.001284 respectively.

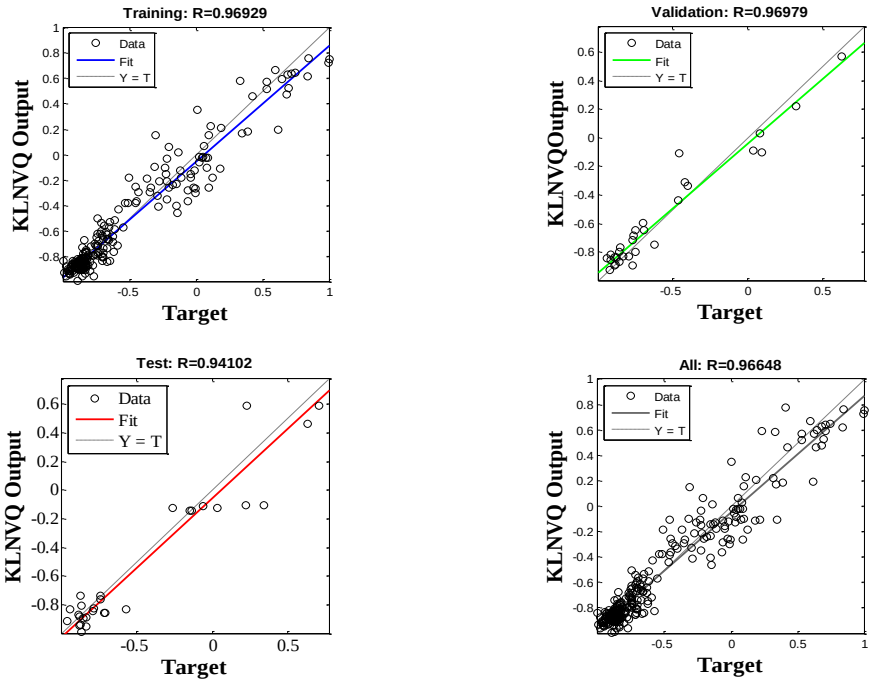

Fig. 3. Regression plots of the KLNVQ model

For comparison purpose as earlier mentioned we also forecast the gasoline price using statistics and econometric models such as ARIMA, MLR, and VAR. The ARIMA

model was first identified several models were tried in order to identify the best fit model. The models that were built and tested are ARIMA (1,0,0), ARIMA (2,1,0), ARIMA (2,2,2), ARIMA (1,0,3), ARIMA (1,3,4), ARIMA (3,0,0) and ARIMA (1,1,1). The model ARIMA (1,0.0) was identified as the best among other comparable models. Ljung-Box was used to validate the ARIMA (1,0,0) model and the results obtained are: $R^2 = 0.74$, statistics = 9.822, Sig. 0.341 and predictors = 5. The MSE of the gasoline price predicted by the ARIMA (1,0,0) and observed prices was completed and it was found to be 0.471101. MLR model was used for the forecasting of gasoline prices. The R^2 measurement of variability generated by the independent variables is 0.61522 and MSE is 0.83522. The R^2 adjusted value of 0.61522 was identical with the original R^2 showing how well the MLR models were able to generalize. This is because shrinkage was not found from the adjusted values. VAR model is also applied for the forecasting of gasoline price and the adjusted R^2 value obtained is 0.79137. The MSE measured for the forecasted and observed gasoline price values is 0.339014.

4.2 Comparing performances of the propose KLNVQ model with ARIMA, MLR, and VAR

From the simulation results of the models presented in section 4.1, it can be summarized that the forecast accuracy of KLNVQ is better than the ARIMA, MLR and VAR in terms of MSE and R^2 (see Fig. 3). Evidence from this research has suggested that the propose KLNVQ can be a substitute for the statistical and econometric models commonly use in the literature for forecasting gasoline price. This performance demonstrated by the propose KLNVQ model can best be attributed to the capability of the model to approximate any nonlinear function with acceptable accuracy as well as the use of Voronoid Polyhedral which likely makes it easier for the model to detect patterns in the historical data. The research will be extended to compare the accuracy of several attribute selection methods such as genetic algorithm, particle swarm optimization, and hybridization of genetic algorithms and wavelet transform.

5 Conclusion

In this paper, we have presented a novel approach for the forecasting of the gasoline price. The approach is modeled based on KLT and NVQ with Voronoid Polyhedral. The experimental data were collected from the Energy Information Administration of the US Department of the Energy. Our approach is effective, robust, and efficient more than the statistical and econometric models propose in the literature. Comparative analysis suggested that the propose KLNVQ model performs better than the ARIMA, MLR, and VAR. Accurate forecasting of gasoline price has implication for the formulation of policies related to sustainable economic development which might improve the economic standard. In addition, having future knowledge of gasoline price can significantly assist policy makers in taking decisions that might successfully deviate from hardship typically cause by shortage of gasoline.

Acknowledgments. This work is supported by University of Malaya High Impact Research Grant no vote UM.C/625/HIR/MOHE/SC/13/2 from Ministry of Higher Education Malaysia.

References

1. Hamilton, J.D.: Historical oil shocks. In: Handbook of Major Events in Economic History, forthcoming (2011)
2. Anderson, S.R., Kellogg, R., Sallee, J.M., Curtin, R.T.: Forecasting Gasoline Prices Using Consumer Surveys. Am. Econ. Rev. 101(3), 110--114 (2011)
3. Deltas, G.: Retail gasoline price dynamics and local market power. J. Ind. Econ. 56(3), 613--628 (2008)
4. Borenstein, S., Cameron, C.A., Gilbert, R.: Do gasoline prices respond asymmetrically to crude oil price change?. Q. J. Econ. 102, 305--339 (1997)
5. Chinn, M., LeBlanc, M., Coibion, O.: The predictive of energy futures: An update on petroleum, natural gas, heating oil and gasoline. National Bureau of Economics research Massachusetts Cambridge, working paper 11033 (2005)
6. Bahrammirzaee, A. A.: comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. Neural Comput. Appl. 19, 1165 --1195 (2010)
7. Jolliffe, I.T.: Principal component analysis (2nd Edn). Springer: New York (2002)
8. Karhunen, J.: Robust PCA methods for complete and missing data. Neural Netw. World 5,357—392 (2011)
9. Martinetz, T.M.**,** Berkovich, S.G., Klaus J. Schulten, K.J.: "Neural-Gas" Network for Vector Quantization and its Application to Time-Series Prediction. IEEE T. Neural Networ. 4(4), 558--569 (1993)