Applied Soft Computing 12 (2012) 2614–2630

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

# An integrated fuzzy regression-data envelopment analysis algorithm for optimum oil consumption estimation with ambiguous data

A. Azadeh[a,*], O. Seraj[a], S.M. Asadzadeh[a], M. Saberi[b]

[a] Department of Industrial Engineering and Center of Excellence for Intelligent Based Experimental Mechanics, Collage of Engineering, University of Tehran, P.O. Box 11365-4563, Iran
[b] Islamic Azad University, Tafresh Branch, Young Researchers Club, Tafresh, Iran

ARTICLE INFO

*Article history:*
Received 15 April 2009
Received in revised form 26 May 2011
Accepted 18 March 2012
Available online 7 April 2012

*Keywords:*
Fuzzy regression
Fuzzy mathematical programming
Data envelopment analysis
Optimization
Oil consumption estimation

ABSTRACT

This study introduces an integrated fuzzy regression (FR) data envelopment analysis (DEA) algorithm for oil consumption estimation and optimization with uncertain and ambiguous data. This is quite important as oil consumption estimations deals with several uncertainties due to social, economic factors. Furthermore, DEA is integrated with FR because there is no clear cut as to which FR approach is superior for oil consumption estimation. The standard indicators used in this paper are population, cost of crude oil, gross domestic production (GDP) and annual oil production. Fifteen popular and most cited FR models are considered in the algorithm. Each FR model has different approach and advantages. The input data is divided into train and test data. The FR models have been tuned for all their parameters according to the train data, and the best coefficients are identified. Center of Average Method for defuzzification output process is applied. For determining the rate of error of FR models estimations, the rate of defuzzified output of each model is compared with its actual rate consumption in test data. The efficiency of 15 FR models is examined by the output-oriented Data Envelopment Analysis (DEA) model without inputs by considering three types of relative error: RMSE, MAE and MAPE. The applicability and superiority of the proposed algorithm is shown for monthly oil consumption of Canada, United States, Japan and Australia from 1990 to 2005.



## 1. Introduction

Regression analysis refers to a set of methods by which estimates are made for the model parameters from the knowledge of the values of a given input–output data set. The goals of the regression analysis are finding an appropriate mathematical model, and determining the best fitting coefficients of the model from the given data. The use of statistical regression is bounded by some strict assumptions about the given data. Overcoming such limitations, fuzzy regression (FR) is introduced which is an extension of the classical regression and is used in estimating the relationships among variables where the available data are very limited and imprecise and variables are interacting in an uncertain, qualitative and fuzzy way. Fundamental differences between FR and classical regression are as follows [1,2]:

- FR can be used to fit fuzzy data and crisp data into a regression model, whereas ordinary regression can only fit crisp data.
- Statistical regression analysis is based on some assumptions. The unobserved error term should mutually be independent and identically distributed. Lack of such assumptions affects the effectiveness of the method. In this case FR can be replaced.
- In contrast to the ordinary regression that is based on probability theory, FR is based on possibility theory and fuzzy set theory.
- Ordinary regression modeling data with randomness type of uncertainty but FR modeling data with fuzziness type of uncertainty.
- In ordinary regression, the unfitted errors between a regression model and observed data are assumed as observation error that is a random variable. In FR, the same unfitted errors are viewed as the fuzziness of the model structure [3].

FR models have been successfully applied to various problems such as forecasting [4] and engineering [5,6]. In energy modeling applications for example Shakouri and Nadimi [7] proposed a novel fuzzy linear regression approach to model the total energy consumption of the Residential-Commercial sector in Iran. Also, in another work by Shakouri et al. [8] the fuzzy rule-based Takagi–Sugeno–Kang (TSK) model is combined with a set of fuzzy regressions (FR) to forecast the short-term electricity demand. It can also be applied for oil forecasting problems. The goal of FR analysis is to find a regression model that fits all observed fuzzy data within a specified fitting criterion. Different FR models are obtained depending on the fitting criterion used. In general, there

* Corresponding author. Tel.: +98 21 82084162; fax: +98 21 82084162.
*E-mail addresses:* aazadeh@ut.ac.ir, ali@azadeh.com (A. Azadeh).

1568-4946/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2012.03.026

are two approaches of FR due to different fitting criterions [9]. The first approach is based on minimizing fuzziness as an optimal criterion which first proposed by Tanaka et al. [20]. Different researchers used Tanaka's approach to minimize the total spread of the output [10]. As pointed out by Wang and Tsaur [11], the advantage of this approach is its simplicity in programming and computation, but it has been criticized to provide too wide ranges in estimation which could not give much help in application [11] and not to utilize the concept of least-squares [12]. The second approach uses least squares of errors as a fitting criterion to minimize the total square error of the output. Different aspects of this approach were investigated by Celmins [13], Diamond [14], Savic and Pedrycz [15] and Chang and Ayyub [32]. Celmins [16] defines a compatibility measure between fuzzy data and a model and uses this measure as a model-fitting criterion. Diamond [14] developed a fuzzy least square method by using the compact α-level sets. Savic and Pedrycz [15] proposed a combined approach for FLSRA by integrating minimum fuzziness criterion into the ordinary least-squares regression. Chang and Ayyub [32] discussed reliability issues of FLSRA, such as standard error and correlation coefficient. This approach, though providing narrower range, costs too much of computation time (2000). Hojati et al. [17] introduced a goal programming-like approach to minimize the total deviation of upper values of H-certain estimated and corresponded observed intervals and deviation of lower values of H-certain estimated and related observed intervals. FR models have been successfully applied to various problems such as forecasting [4] and engineering [6]. It can also be applied for energy forecasting problems. Conventional regression along with intelligent approaches has been used for energy consumption estimation [18,19].

The remainder of this paper is organized as follows. In the next section, different types of FR and their shortcomings are brought. Error estimation methods are presented in Section 3.1. Some popular defuzzification methods are introduced in Section 3.2. Comparing the FR models for each case study is shown in Section 4 followed by the obtained results of the case studies in Section 4.1. At last the summary of this research is presented in Section 5.

## 2. Fuzzy regression models

Fuzzy linear regression was introduced by Tanaka et al. [20], to decide a fuzzy linear relationship by, $Y = A_0X_0 + A_1X_1 + \cdots + A_kX_k$; where regression coefficients $A_j, j = 0, \ldots, K$, were supposed to be a symmetric triangular fuzzy number, with center $\alpha_j$, having membership function equal to one, and spreads $c_j, c_j \geq 0$. The dependent variable ($y$) is a fuzzy number. The independent variables ($x$) can be taken into consideration as crisp or fuzzy numbers.

The input information are $n$ sets of variables $y_i, x_{i0}, x_{i1}, \ldots, x_{ij}$, $i = 1, 2, \ldots, n$; $n \geq j+1$, where $x_{i0} = 1$. The response variable $y_i$ is assumed to be a symmetric triangular fuzzy number with central value $\bar{y}_i$ and spreads $\bar{e}_i$, where $\bar{e}_i \geq 0$. Independent variables values $x_{ij}$, $i = 1, 2, \ldots, n$ $j = 1, 2, \ldots, k$, is also supposing to be a symmetric triangular fuzzy number with a center $\bar{x}_{ij}$ and spreads $f_{ij}$ ($f_{ij} \geq 0$). The assigned membership functions of both dependent and independent variables are linear. If we are just interested in that membership function value of $y_i$ has at least $H$, where $0 \leq H \leq 1$, we should consider the interval $[\bar{y}_i - (1-H) \times \bar{e}_i \bar{y}_i + (1-H) \times \bar{e}_i]$. This interval is illustrated in Fig. 1.

Here, $H$ shows the minimum acceptable degree of precision, and we will make reference to this interval as $H$-certain observed interval. Similarly, suppose that the independent variables $x_j, j = 1, 2, \ldots k$, have certain values and regression coefficient $A_j, j = 1, 2, \ldots k$, are assume to be symmetric triangular fuzzy numbers, the estimated interval corresponding to a input set of independent variables

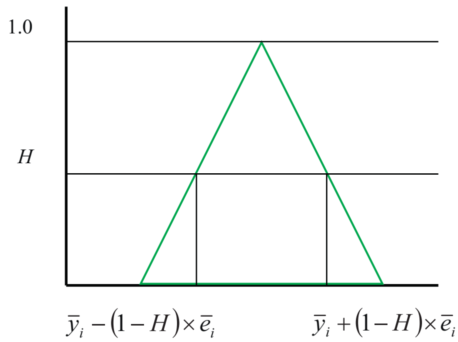


**Fig. 1.** An $H$-certain observed interval.

$X(x_{i0}, x_{i1}, \ldots, x_{ik})$ having membership function value of at least $H$ is: $\left[\sum_{j=0}^{k}(\alpha_j - (1-H) \times c_j) \times x_{ij} \quad \sum_{j=0}^{k}(\alpha_j + (1-H) \times c_j) \times x_{ij}\right]$, We will refer to this distance as $H$-certain estimated interval.

The membership function of the fuzzy parameter $A_j$ is represented by:

$$\mu_{A_j}(a_j) = \begin{cases} 1 - \dfrac{|\alpha_j - a_j|}{c_j} & \text{for} \quad \alpha_j - c_j \leq a_j \leq \alpha_j + c_j \\ 0 & \text{otherwise} \end{cases}$$

For Case 1 [20] introduced the following linear programming formulation to predict $A_j, j = 1, 2, \ldots, k$:

$$\begin{aligned} &\text{Minimize} && c_0 + c_1 + c_2 + \cdots + c_k \\ &\text{subject to}: && \sum_{j=0}^{k}(\alpha_j + (1-H) \times c_j) \times x_{ij} \geq \bar{y}_i + (1-H) \times \bar{e}_i \quad i = 1, \ldots, n, \\ & && \sum_{j=0}^{k}(\alpha_j - (1-H) \times c_j) \times x_{ij} \leq \bar{y}_i - (1-H) \times \bar{e}_i \quad i = 1, \ldots, n, \\ & && \alpha_j = \text{free}, \quad c_j \geq 0, \quad j = 0, \ldots, k. \end{aligned} \tag{1}$$

Note that in the above model $c_j$s are supposed to be non-negative, because the fuzziness in estimated intervals usually increases for larger values of independent variables $x_j$ [17]. Fig. 2 presents degree of imprecision of $\tilde{Y}_i$ to the collected data $\tilde{\tilde{Y}}_i$.

There are some criticisms on Tanaka et al.'s [20] FR model. One of them is that the results are $x_j$-scale dependent and many $c_j$s might equal to zero [26]. To repair this problem, replacement for sum of spreads of FR model's coefficients, sum of spreads of the estimated intervals can be used as an objective function. Tanaka

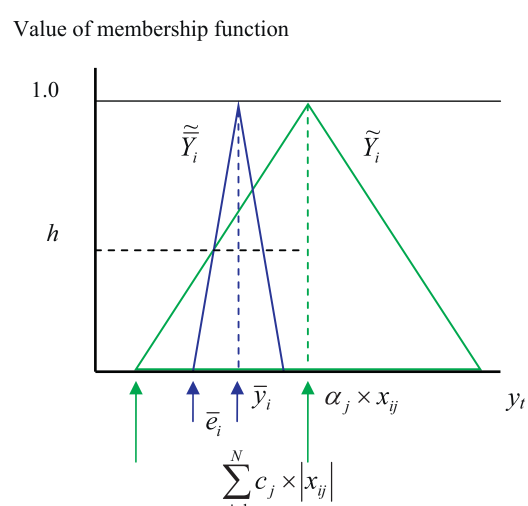


**Fig. 2.** Degree of imprecision of $\tilde{Y}_i$ to the collected data $\tilde{\tilde{Y}}_i$.