Computers and Electronics in Agriculture 154 (2018) 10–17

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Prediction of soybean price in China using QR-RBF neural network model

Dongqing Zhang[a,*], Guangming Zang[b], Jing Li[a], Kaiping Ma[a], Huan Liu[a]

[a] College of Engineering, Nanjing Agricultural University, Nanjing 210031, China
[b] Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

ARTICLE INFO

Keywords:
Forecast
Quantile regression-radial basis function (QR-RBF) neural network
Gradient descent
Genetic algorithm

ABSTRACT

As the price of soybean affects the soybean market development and food security in China, its forecasting is essential. A quantile regression-radial basis function (QR-RBF) neural network model is introduced in this paper. The model has two characteristics: (1) using quantile regression models to describe the distribution of the soybean price range; and (2) using RBF neural networks to approximate the nonlinear component of the soybean price. In order to optimize the QR-RBF neural network model parameters, a hybrid algorithm known as GDGA, based on a combination of the genetic algorithm (performing a global search) and a gradient descent method (performing a local search), is proposed in this paper. Data regarding the monthly domestic soybean price in China were analyzed and the results indicate that the proposed hybrid GDGA is effective. Furthermore, the results suggest that the influencing factors of soybean price vary at different price levels. Money supply and port distribution price of imported soybean were found to be important across a range of quantiles; output of domestic soybean and consumer confidence index were important only for low quantiles; and import volume of soybean and consumer price index were important only for high quantiles.

1. Introduction

In China, soybean is an important commodity with well-developed spot and future markets. The importance of soybean and its various derivatives, such as soy meal and soy oil, cannot be underestimated. Price fluctuations in these product markets have far-reaching effects on consumers, farmers, and soybean processors. Understanding the price trend becomes a prerequisite for policymakers to implement any price-control policy in agricultural product markets and subsidy policies in consumer consumption. Therefore, it is essential to forecast the domestic soybean price in China.

The soybean price has been the subject of numerous previous papers. Malone (1968) discussed the role of the spectator and placed emphasis on the study of forecasting soybean future prices in terms of the fundamental market, technical market analysis, and basic trading rules. Wiles and Enke (2015) optimized the moving average convergence divergence parameter values from traditionally used integers to values that optimize the soybean market profile (soybean future contracts) based on the genetic algorithm (GA), which is dependent on the soybean prices of the entering and exiting time. Berwald and Havenner (1997) used a multivariate state space-time series model to forecast the soybean, meal, and oil prices in both the spot and future markets in the United States (US), and the results indicated that the model fits highly effectively. Cao et al. (2016) examined the causalities of the soybean price movement among the US future market, Chinese domestic future market, and Chinese spot markets, and found that the soybean price movement originated from the US future market, then passed through the Chinese future market, and finally reached the Chinese spot market. Ahumada and Cornejo (2016) forecasted food (corn, soybean, and wheat) prices, and analyzed whether the forecasting accuracies of individual food price models could be improved by considering their cross-dependence. Pal and Mitra (2017) employed a quantile autoregressive distributed lag model to explore the possible relationship between diesel and soybean prices in the US, and the results indicated that the soybean price movement was tail-dependent and varied over quantiles, and the soybean prices responded strongly to diesel price fluctuations in the upper quantiles compared to the lower quantiles. Adrangi et al. (2006) investigated the price discovery process between soybean futures and the Crush constituents, soy meal, and soy oil in the US, and determined a strong bi-directional causality in the Crush futures prices. Moreover, while soybean contracts bore the burden of convergence when the spread between soybean and soymeal contract prices widened, this was not true for the soybean and soy oil contracts. Li et al. (2012) employed a linear quantile regression approach to analyze the movement of agricultural product prices in China.

From the literature mentioned above, it can be determined that the

* Corresponding author.
E-mail address: zhangdq@njau.edu.cn (D. Zhang).

https://doi.org/10.1016/j.compag.2018.08.016
Received 25 February 2018; Received in revised form 12 June 2018; Accepted 7 August 2018
Available online 06 September 2018
0168-1699/

study of soybean price mainly focuses on the prediction of soybean future prices and the relationship between prices for soybean and its derivative products. This paper investigates the soybean spot prices in China. The methods applied in the previous studies mainly focus on the mean soybean price; however, the results are not highly satisfactory when extreme conditions occur; for example, the soybean price is very high or low. Therefore, in this paper, we address the problem of the domestic soybean spot price in China using the quantile regression model, which overcomes certain limitations of the mean model mentioned above.

Quantile regression was developed by Koenker and Bassett (1978) for estimating the quantiles of a variable assumed to be a linear function of other variables. Quantile regression has received considerable attention in the predictive literature, such as precipitation downscaling (Tareghian and Rasmussen, 2013; Cannon, 2011), thermal load prediction (Kapetanakis et al., 2017), inflation (Korobilis, 2017), financial returns (Taylor, 2000), and electricity spot prices (Maciejowska et al., 2016), but less so in the soybean price literature.

The majority of quantile regression applications in prediction tasks involving multiple predictors have relied on linear or simple parametric nonlinear models (Koenker, 2005). A practical implementation of the quantile regression neural network (QRNN), which is a more flexible type of model, was introduced by Taylor (2000). In this study, a radial basis function (RBF) neural network is selected owing its simple topological structure and ability to reveal learning in an explicit manner, and a quantile regression-radial basis function (QR-RBF) neural network model is introduced to predict the domestic soybean price in China.

This study differs from the related work by Li et al. (2012) in terms of the methodology and research question. Li et al. (2012) employed a linear quantile regression approach to analyze the movement of agricultural product prices (pork, chicken, and egg) in China, while this study uses a QR-RBF neural network model to predict the domestic soybean price in China.

The proposed QR-RBF model includes two features: first, the quantile regression models are applied to describe the distribution of the soybean price range, which can be used under extreme conditions and extract more useful information from the data; and secondly, the RBF neural networks are adopted to approximate the nonlinearity of variables without requiring a user-specified problem-solving algorithm, and possesses inherent generalization abilities.

The remainder of this paper is organized as follows. Section 2 presents the architecture of the QR-RBF neural network prediction model, and proposes a hybrid algorithm combining the gradient descent method and GA to estimate the QR-RBF model parameters. A description of a sample of domestic soybean prices and potential influential factors as well as experimental results are provided in Section 3. Finally, Section 4 offers final remarks.

2. Materials and methods

2.1. Materials

This study aimed to predict the soybean price in China, and the data were obtained from the “Food China” journal. The monthly soybean prices from January 2010 to December 2015 are illustrated in Fig. 1. The lowest soybean price was 3595 RMB·ton^{-1} in July 2010, while the highest was 4720 RMB·ton^{-1} in October 2012, and the fluctuation between the highest and lowest prices exceeded 30%. Conventional regression models have been used in numerous price-forecasting studies (Lessmann and Voß, 2017; Rounaghi et al. 2015; Du et al. 2009; Koulouriotis et al. 2002). Fig. 1 indicates that the soybean price fluctuated frequently, and the traditional linear regression model results were not satisfactory. Similar to Statnik and Verstraete (2015), we also assume it was nonlinear. However, the traditional regression model describes the conditional mean of the response variable given certain predictor variable values, which is not effective under extreme conditions (significantly higher or lower prices). Quantile regression can deal with either the conditional median or other quantiles of the response variable, and different measures of central tendency and statistical dispersion may be useful for obtaining a more comprehensive analysis of the relationships among variables, which means that the quantile regression model is available when extreme conditions occur (Koenker, 2005). Therefore, the quantile regression neural network (QRNN) was introduced to forecast the soybean price.

2.2. Linear quantile regression

We first discuss the linear quantile regression introduced by Koenker and Bassett (1978), with the following form:

$$Q_y(\tau\,|\mathbf{X}) = \mathbf{X}^T\beta(\tau) \tag{1}$$

where scale y and vector $\mathbf{X} = (x_1, \cdots, x_n)$ are the dependent and independent variables, respectively, and $\beta(\tau)$ is a vector of parameters dependent on $\tau(0 \leqslant \tau \leqslant 1)$. In a linear quantile regression, each conditional distribution quantile is represented by an individual hyperplane. For a given set of observations $(\mathbf{X}_t, y_t)$, $t = 1, \cdots, T$, $\beta(\tau)$ is defined as the solution to the minimization problem

$$\min_{\beta}\left\{\sum_{y_t \geqslant \mathbf{X}_t^T\beta} \tau\,|y_t - \mathbf{X}_t^T\beta| + \sum_{y_t < \mathbf{X}_t^T\beta} (1-\tau)|y_t - \mathbf{X}_t^T\beta|\right\} \tag{2}$$

Fig. 1. The domestic soybean price in China from 2010 to 2015.

For quantile regression details, please refer to the study of Koenker and Bassett (1978).

2.3. QR-RBF neural network

Eq. (2) describes the linearity of the dependent and independent variables; however, non-linearity is more popular in numerous cases. In order to deal with non-linearity, Taylor (2000) proposed the QRNN. Suppose that the QRNN consists of a set of n inputs, connected to each of m units in a single hidden layer, which, in turn, are connected to an output. The resultant model can be written as

$$O(\mathbf{X}_t, \Theta(\tau)) = g_2\left(\sum_{j=1}^{m} w_j(\tau) g_1\left(\sum_{i=1}^{n} v_{ji}(\tau) x_{ti}\right)\right) \tag{3}$$

where $\mathbf{X}_t$ is the network input with elements of x_{ti}, $\Theta(\tau) = \{\mathbf{V}(\tau), \mathbf{W}(\tau)\}$ represents the network weights with elements of $v_{ji}(\tau)$ and $w_j(\tau)$. Furthermore, $O(\mathbf{X}_t, \Theta(\tau))$ is the output of the τ th quantile QRNN, while $g_1(\cdot)$ and $g_2(\cdot)$ are activation functions, which are frequently selected as sigmoidal and linear, respectively.

In this paper, a RBF neural network with a variable structure is selected to replace the QRNN owing to its simple topological structure and ability to reveal how learning proceeds in an explicit manner. In practice, the most common RBF is a Gaussian kernel (De Freitas et al. 2001), given by $\phi(\tau|X) = \exp(-\|\mathbf{X}_t - \mathbf{C}(\tau)\|^2/2\sigma(\tau)^2)$, where $\|\cdot\|$ denotes the norm, $\mathbf{C}(\tau)$ are the RBF centers, and $\sigma(\tau)$ is the kernel width factor. The function approximation may be expressed as a linear combination of the RBFs, and the QR-RBF neural network can be represented by

$$O(\mathbf{X}_t, \Theta(\tau)) = \sum_{j=1}^{m} w_j(\tau)\phi_j(\tau\,|\mathbf{X}_t) \tag{4}$$

where $\mathbf{X}_t$ and $O(\mathbf{X}_t, \Theta(\tau))$ are the input and output of the τ th quantile QR-RBF neural network. The network parameters are $\Theta(\tau) = \{\mathbf{W}(\tau), \mathbf{C}(\tau), \sigma(\tau)\}$, and $\mathbf{W}(\tau)$ are the output weights of the RBF network.

Similar to fitting a linear quantile function using the expression in Eq. (2), the parameters $\Theta(\tau)$ of the QR-RBF neural network model can be estimated by minimizing the following problem:

$$\min_{\Theta(\tau)} E(\Theta(\tau)) = \min_{\Theta(\tau)} \left\{ \sum_{y_t \geq O(\mathbf{X}_t, \Theta(\tau))} \tau\,|y_t - O(\mathbf{X}_t, \Theta(\tau))| + \sum_{y_t < O(\mathbf{X}_t, \Theta(\tau))} (1-\tau)|y_t - O(\mathbf{X}_t, \Theta(\tau))| \right\} \tag{5}$$

2.4. Predictive density from quantiles

The QR-RBF neural network model is used to predict a finite number of conditional quantiles and the entire predictive distribution is required. For notational convenience, we let

$$Q(\tau\,|\mathbf{X}_t) = O(\mathbf{X}_t, \Theta(\tau)) \tag{6}$$

For a given τ, $Q(\tau\,|\mathbf{X}_t)$ is a network output. When τ varies in the interval [0,1], $Q(\tau\,|\mathbf{X}_t)$ is the quantile function. In order to compute the probability density function (pdf) of $Q(\tau\,|\mathbf{X}_t)$, we assume that the cumulative distribution function (cdf) of $Q(\tau\,|\mathbf{X}_t)$ is $F(Q(\tau\,|\mathbf{X}_t))$, because

$$F(Q(\tau\,|\mathbf{X}_t)) = F(F^{-1}(\tau)) = \tau \tag{7}$$

Taking the derivative of this quantity with respect to τ results in

$$\frac{dF(Q(\tau|\mathbf{X}_t))}{d\tau} = \frac{dF(F^{-1}(\tau|\mathbf{X}_t))}{d\tau} = 1 \tag{8}$$

Using the derivative chain rule and $\frac{dF(Q(\tau\,|\,\mathbf{X}_t))}{dQ(\tau\,|\,\mathbf{X}_t)} = f(Q(\tau|\mathbf{X}_t))$, we have

$$\frac{dF(Q(\tau|\mathbf{X}_t))}{dQ(\tau|\mathbf{X}_t)}\frac{dQ(\tau|\mathbf{X}_t)}{d\tau} = f(Q(\tau|\mathbf{X}_t))\frac{dF^{-1}(\tau|\mathbf{X}_t)}{d\tau} = 1 \tag{9}$$

Therefore, we can obtain the following (Koenker and Machado, 1999):

$$f(Q(\tau|\mathbf{X}_t)) = \frac{d\tau}{dF^{-1}(\tau|\mathbf{X}_t)} = \frac{d\tau}{dQ(\tau|\mathbf{X}_t)} = 1/\{\frac{dQ(\tau|\mathbf{X}_t)}{d\tau}\} \tag{10}$$

The computation of $\frac{dQ(\tau\,|\,\mathbf{X}_t)}{d\tau}$ can be achieved using the difference method:

$$\frac{dQ(\tau|\mathbf{X}_t)}{d\tau} = \frac{Q(\tau + \Delta\tau|\mathbf{X}_t) - Q(\tau|\mathbf{X}_t)}{\Delta\tau} \tag{11}$$

or

$$\frac{dQ(\tau|\mathbf{X}_t)}{d\tau} = \frac{Q(\tau + \Delta\tau|\mathbf{X}_t) - Q(\tau - \Delta\tau|\mathbf{X}_t)}{2\Delta\tau} \tag{12}$$

2.5. Model architecture

The primary goal of this paper is to use the QR-RBF neural network model to predict the soybean price, considering possible influential factors and its past lagged observations; therefore, it is a predictive problem with multi-input and single output. Given a quantile τ, the QR-RBF neural network forecast model is the RBF neural network, which can adaptively select the input influential factors and has a variable number of hidden nodes.

The architecture of the QR-RBF neural network forecast model is presented in Fig. 2. Here, $\tau_1, \cdots, \tau_p$ are the p values, τ of which are distributed uniformly in the interval [0, 1]. Given a quantile τ, the proposed forecast model is the RBF neural network with $n = n_1 + q$ input variables, $m(\tau)$ hidden neurons, and one output node. It is noted that the input $\mathbf{X}_t$ includes two components: $\{s_i(\tau)\cdot x_{ti}\}$, $i = 1, \cdots, n_1$ and $\{y_{t-1}, \cdots, y_{t-q}\}$. Here, $\{s_i(\tau)\cdot x_{ti}\}$, $i = 1, \cdots, n_1$ denotes the potential selected influential factors; $\mathbf{S}(\tau)$ is a binary vector and its element $s_i(\tau)$ takes the value of 1 or 0, where $s_i(\tau) = 1$ if x_{ti} is selected and equal to zero if not; and $\{y_{t-1}, \cdots, y_{t-q}\}$ are the past lagged observations of the soybean price, where q is fixed and can be determined by a trial approach or autocorrelation coefficient. Therefore, the network parameters for this QR-RBF neural network are $\Theta(\tau) = [\mathbf{S}(\tau), m(\tau), \mathbf{C}(\tau), \sigma(\tau), \mathbf{W}(\tau)]$.

The RBF neural network with quantile τ_p is illustrated in the dashed frame in Fig. 2. When τ varies, the network parameters can be optimized to obtain the different outputs $Q(\tau\,|\mathbf{X}_t)$. Then, Eq. (10) is used to obtain the probability density distribution $f(Q(\tau\,|\mathbf{X}_t))$, and forecast value $\widehat{Q}(\tau\,|\mathbf{X}_t)$ can be estimated by

$$\widehat{Q}(\tau\,|\mathbf{X}_t) = \arg\max_{Q(\tau|\mathbf{X}_t)} \{f(Q(\tau\,|\mathbf{X}_t)\} \tag{13}$$

2.6. Hybrid algorithm for QR-RBF neural network model

Prior to forecasting, the architectures and parameters of the QR-RBF neural network prediction model must be estimated. However, few studies have focused on the QR-RBF neural network prediction model, except for Cannon (2011), who adopted the gradient decent method. Although gradient descent is very well suited to suboptimal local searching, it cannot perform global optimization. The GA is capable of global optimization, but exhibits slow convergence properties near local optima. Therefore, a hybrid algorithm known as GDGA is proposed, which combines the gradient descent method and GA. In the GDGA algorithm, we iteratively use a gradient descent algorithm to accelerate the seeking of local optima, and the GA to guarantee global optimization. The GA contains a population of individuals competing against one another in relation to a fitness measure, which means that the GA is a parallel algorithm and may avoid local optimization (Li and Gao, 2016). This proposed approach, together with its corresponding algorithm, is explained in detail in the remainder of the paper.

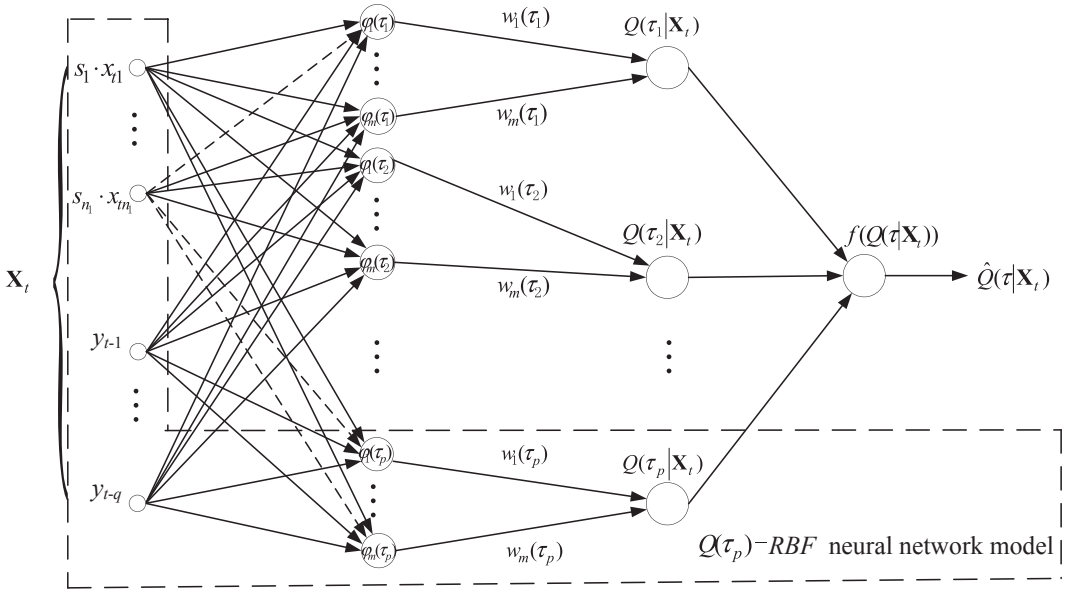

Fig. 2. Architecture of QR-RBF neural network prediction model.

For the QR-RBF neural network prediction model, the number of quantiles p is determined by computational complexity and prediction accuracy, while q is fixed and determined by a trial approach or auto-correlation coefficient. Thus, for a given τ, the entire set of parameters $\Theta(\tau) = \{\mathbf{S}(\tau), m(\tau), \mathbf{C}(\tau), \sigma(\tau), \mathbf{W}(\tau)\}$ in the QR-RBF neural network prediction model is simultaneously optimized by means of the hybrid algorithm. In this study, the possible influential factors are identified by the binary vector variable $\mathbf{S}(\tau)$ with elements 0 or 1. Prior to training the QR-RBF neural network, the data need to be transformed to values within the interval of 0 and 1.

The corresponding process for the hybrid algorithm GDGA is as follows:

Step 1: Parameter initialization

Parameter initialization mainly involves the determination of the parameter range of the QR-RBF neural network model.

Step 2: Coding

Coding can be defined as the feasible solution to the problem being converted from the solution space to the search space, and being able to be processed by the GA. The network parameters $\Theta(\tau) = \{\mathbf{S}(\tau), m(\tau), \mathbf{C}(\tau), \sigma(\tau), \mathbf{W}(\tau)\}$ must be coded in chromosomes in a string representation. In this paper, binary coding is adopted to represent $\mathbf{S}(\tau)$, $m(\tau)$, while real coding is adopted for $\mathbf{C}(\tau)$, $\sigma(\tau)$ and $\mathbf{W}(\tau)$. Fig. 3 illustrates the chromosome structure of the QR-RBF neural network model.

Step 3: Initial population generation

The initial population generation mainly involves determining the population size and randomly creating an initial population of individuals.

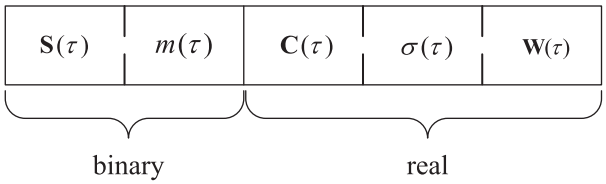

Fig.3. Chromosome representation.

Step 4: Gradient descent optimization

The gradient descent method is used to optimize the parameters $\mathbf{C}(\tau)$, $\sigma(\tau)$, $\mathbf{W}(\tau)$ of the QR-RBF model first, and then these parameters are passed on to the following GA search method.

As noted in Eq. (5), the parameters $\mathbf{C}(\tau)$, $\sigma(\tau)$, $\mathbf{W}(\tau)$ of the QR-RBF neural network model can be optimized by minimizing $E(\Theta(\tau))$. Suppose that e_t, $t = 1, \cdots, T$ is the error of the t th actual and target output values; then,

$$e_t = \begin{cases} \tau\,|y_t - O(\mathbf{X}_t, \Theta(\tau))|, & y_t \geqslant O(\mathbf{X}_t, \Theta(\tau)) \\ (1-\tau)|y_t - O(\mathbf{X}_t, \Theta(\tau))|, & y_t < O(\mathbf{X}_t, \Theta(\tau)) \end{cases} \tag{14}$$

where y_t is the actual observation at time t, and $O(\mathbf{X}_t, \Theta(\tau))$ is the QR-RBF neural network output. Our goal is to determine optimum QR-RBF parameters to minimize this error.

In order to use the gradient-based optimization algorithm of GDGA, the objective function must be differentiable. However, the objective function in the absolute form in Eq. (5) is not differentiable. Therefore, as in Schwenker et al. (2001), we relax and modify the objective function in the square form, because the setting is convenient for taking the derivative, such that

$$E(\Theta(\tau)) = \frac{1}{2}\sum_{t=1}^{T} e_t^2 \tag{15}$$

Our goal is to determine the optimum QR-RBF parameters to minimize the objective function in Eq. (15). Considering the above error function, a necessary condition for the minimal error $E(\Theta(\tau))$ is that its partial derivatives with respect to the parameter center locations $\mathbf{C}(\tau)$, kernel widths $\sigma(\tau)$, and output weights $\mathbf{W}(\tau)$ vanish.

Gradient descent is an iterative procedure for identifying the optimal parameters. Here, the parameters $\mathbf{C}(\tau)$, $\sigma(\tau)$, $\mathbf{W}(\tau)$ are moved by a small distance in the direction in which $E(\Theta(\tau))$ decreases most rapidly; that is, in the direction of the negative gradient of $E(\Theta(\tau))$ with respect to these parameters. For a given τ, we have the following iterative scheme:

$$\mathbf{C}_j^{(k+1)}(\tau) = \mathbf{C}_j^{(k)}(\tau) + \Delta\mathbf{C}_j(\tau) \tag{16}$$

$$\sigma_j^{(k+1)}(\tau) = \sigma_j^{(k)}(\tau) + \Delta\sigma_j(\tau) \tag{17}$$

$$w_j^{(k+1)}(\tau) = w_j^{(k)}(\tau) + \Delta w_j(\tau) \tag{18}$$

where $j = 1, \cdots, m(\tau)$ and k is the iteration time. Moreover, $\Delta\mathbf{C}_j(\tau)$, $\Delta\sigma_j(\tau)$, $\Delta w_j(\tau)$ are the iteration increments, which can be expressed by

$$\Delta\mathbf{C}_j(\tau) = -\eta_C \frac{\partial E(\tau)}{\partial \mathbf{C}_j(\tau)} \tag{19}$$

$$\Delta\sigma_j(\tau) = -\eta_\sigma \frac{\partial E(\tau)}{\partial \sigma_j(\tau)} \tag{20}$$

$$\Delta w_j(\tau) = -\eta_w \frac{\partial E(\tau)}{\partial w_j(\tau)} \tag{21}$$

where $\eta_C, \eta_\sigma, \eta_w \in [0, 1]$ are the learning rates or step sizes of $\mathbf{C}_j(\tau)$, $\sigma_j(\tau)$, $w_j(\tau)$, respectively, and selecting these parameters is at times a critical issue in neural network training. If the value is excessively low, convergence to a minimum is slow; conversely, if it is selected to be excessively high, the successive steps in the parameter space overshoot the error surface minimum.

For the three partial derivatives in Eqs. (19), (20), and (21), we have

$$\frac{\partial E(\Theta(\tau))}{\partial \mathbf{C}_j(\tau)} = \frac{w_j(\tau)}{\sigma_j(\tau)^2}\left(\begin{array}{l}\sum\limits_{y_t \geq O(\mathbf{X}_t,\Theta(\tau))} \tau e_t \phi(\|\mathbf{X}_t - \mathbf{C}_j(\tau)\|)(\mathbf{C}_j(\tau) - \mathbf{X}_t) - \\ \sum\limits_{y_t < O(\mathbf{X}_t,\Theta(\tau))} (1-\tau) e_t \phi(\|\mathbf{X}_t - \mathbf{C}_j(\tau)\|)(\mathbf{C}_j(\tau) - \mathbf{X}_t)\end{array}\right) \tag{22}$$

$$\frac{\partial E(\Theta(\tau))}{\partial \sigma_j(\tau)} = \frac{w_j(\tau)}{\sigma_j(\tau)^3}\left(\begin{array}{l}\sum\limits_{y_t < O(\mathbf{X}_t,\Theta(\tau))} (1-\tau) e_t \phi(\|\mathbf{X}_t - \mathbf{C}_j(\tau)\|)\|\mathbf{X}_t - \mathbf{C}_j(\tau)\|^2 \\ - \sum\limits_{y_t \geq O(\mathbf{X}_t,\Theta(\tau))} \tau e_t \phi(\|\mathbf{X}_t - \mathbf{C}_j(\tau)\|)\|\mathbf{X}_t - \mathbf{C}_j(\tau)\|^2\end{array}\right) \tag{23}$$

$$\begin{aligned}\frac{\partial E(\Theta(\tau))}{\partial w_j(\tau)} = & \sum_{y_t < O(\mathbf{X}_t,\Theta(\tau))} (1-\tau) e_t \phi(\|\mathbf{X}_t - \mathbf{C}_j(\tau)\|) \\ & - \sum_{y_t \geq O(\mathbf{X}_t,\Theta(\tau))} \tau e_t \phi(\|\mathbf{X}_t - \mathbf{C}_j(\tau)\|)\end{aligned} \tag{24}$$

Step 5: Fitness function calculation

In the evolutionary searching of the GA, the fitness function value is used to evaluate the individual merits, and is also an important basis for subsequent genetic manipulation. In this paper, the fitness function is defined as $1/E(\Theta(\tau))$.

Step 6: Selection operator

Among the available chromosomes in a population, some will be selected for reproduction; however, the more graceful chromosomes stand a higher chance of being selected for breeding. In this paper, random ergodic sampling is used to select the parent individual.

Step 7: Crossover operator

This operator works on a pair of production chromosomes, and a new pair of chromosomes will be produced in the process. A one-point crossover is used for $\mathbf{S}(\tau)$ and $m(\tau)$, and a discrete crossover operator is adopted for $\mathbf{W}(\tau)$, $\mathbf{C}(\tau)$ and $\sigma(\tau)$ (Beyer and Schwefel, 2002).

Step 8: Mutation operator

Following the crossover process, the mutation operator will take effect on the chromosomes. This operator accidentally selects a gene from a chromosome, following which it will change the gene content. If the gene consists of binary numbers, it will be converted into the inverse. While the gene is a real, it can be mutated into a value that is randomly sampled according to a Gaussian distribution, with a mean of this gene and fixed standard deviation.

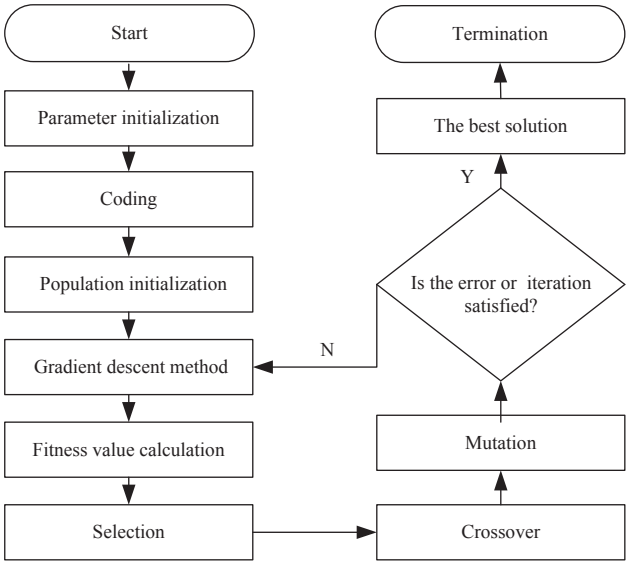

Fig. 4. Process of the hybrid GDGA algorithm.

Step 9: Termination

Has the convergence criteria (error or iteration) been reached? If not, go to step 4; if so, the optimal architectures and parameters of the QR-RBF neural network prediction model are provided.

The specific algorithm process is illustrated in Fig. 4.

3. Results and discussion

3.1. Data

According to Baffes and Haniotis (2016) and Balcombe (2009), supply and demand, macroeconomic factors, and other aspects play a role in determining agricultural commodity prices. Owing to data constraints, we were unable to include all factors in the model. Therefore, the potential quantifiable factors that may have an influence on the domestic soybean price in China were considered, as follows: output of domestic soybean x_1, import volume of soybean x_2, output of global soybean x_3, demand of domestic soybean x_4, consumer price index x_5, consumer confidence index x_6, money supply x_7, and port distribution price of imported soybean x_8. A total of 8 candidate factors influencing the domestic soybean price in China were selected and are listed in Table 1. Moreover, the past lagged observations affected the agricultural product price (Xiong et al. 2015), so the past lagged observations of soybean price y_{t-i}, $i = 1, \cdots, q$ (RMB·ton^{-1}) were selected as the input for the QR-RBF neural network forecast model in this study.

The monthly data for soybean price and candidate influential factors were used in our study. The data used here covered the period spanning from January 2010 to December 2015. The data were split into two sub-samples: 2010.1–2015.4 was used for model training, and

Table 1
Potential factors influencing on domestic soybean price in China.

Influential factors	Abbreviation
Output of domestic soybean (million ton)	DSO
Import volume of soybean (million ton)	SIV
Output of global soybean (million ton)	GSO
Demand of domestic soybean (million ton)	DSD
Consumer price index	CPI
Consumer confidence index	CCI
Money supply (100 million RMB)	MS
Port distribution price of imported soybean (RMB·ton^{-1})	ISPDP

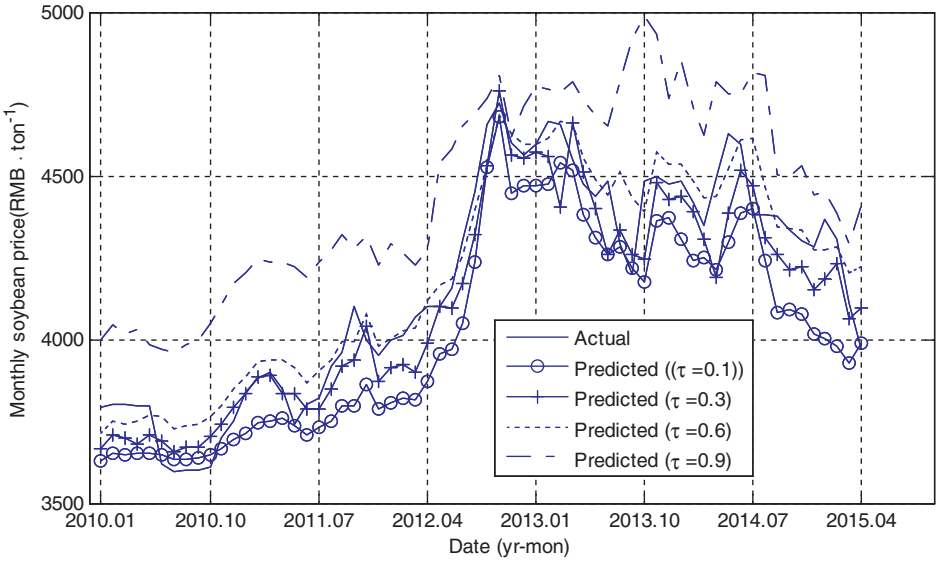

Fig. 5. The QR-RBF fitting curves with quantiles of 0.1, 0.3, 0.6 and 0.9.

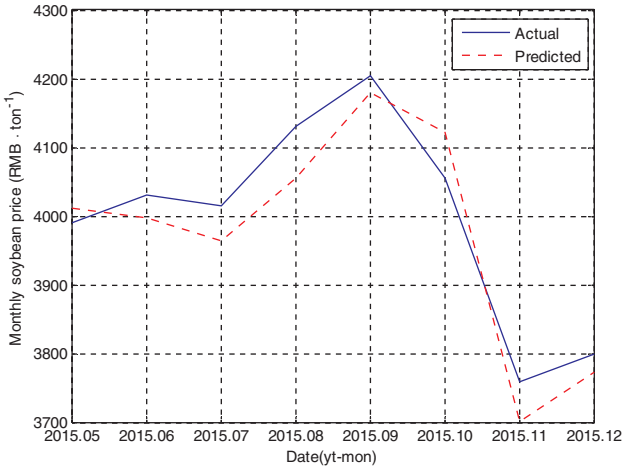

Fig. 6. Predicted soybean prices based on the QR-RBF neural network model.

2015.5–2015.12 was used for testing. The domestic soybean price data were obtained from the "Food China" journal; the output of domestic soybean, import volume of soybean, output of global soybean, and demand of domestic soybean data were obtained from the U.S. Department of Agriculture (USDA, http://www.usda.gov/); the consumer price index, consumer confidence index, and money supply data were obtained from http://www.eastmoney.com/; and the port distribution price of imported soybean was obtained from http://www.feedtrade.com.cn/.

3.2. Parameters setting

Prior to forecasting, several parameters need to be set in the proposed hybrid GDGA. From Sections 2.5 and 3.1, the input variables of the QR-RBF neural network model are $\mathbf{X}_t = (x_{1t}\cdot s_1(\tau), x_{2t}\cdot s_2(\tau), x_{3t}\cdot s_3(\tau), x_{4t}\cdot s_4(\tau), x_{5t}\cdot s_5(\tau), x_{6t}\cdot s_6(\tau), x_{7t}\cdot s_7(\tau), x_{8t}\cdot s_8(\tau), y_{t-1},\cdots,y_{t-q})^T$. For any quantile τ, $s_i(\tau)$, $i = 1, \cdots, 8$ values of 0 or 1 are randomly obtained. According to the autocorrelation and partial correlation function of the domestic soybean price in China, q equals one; that is, $q = 1$. For the quantile τ, we set $(\tau_1, \tau_2, \cdots, \tau_p) = (0.01, 0.02, \cdots, 0.99)$, $p = 99$.

The QR-RBF network experiences a common problem in terms of the number of hidden nodes to allocate. With too few, a network fails to learn, and with too many, its ability to generalize is poor (often referred to as overtraining); thus, in this paper, $m(\tau)$ is drawn as a random integer in the interval of [5, 13] by the trial method. The initial values of $\mathbf{C}(\tau)$, $\sigma(\tau)$, $\mathbf{W}(\tau)$ are randomly generated in the intervals [0, 1.5], [0, 2], and [0, 1], respectively. Then, all parameters $\{\mathbf{S}(\tau), m(\tau), \mathbf{C}(\tau), \sigma(\tau), \mathbf{W}(\tau)\}$ are simultaneously optimized by the proposed hybrid algorithm.

Moreover, we set $\eta_C = \eta_\sigma = \eta_w = 0.0015$, the number of inner iterations is 3 for the gradient descent method, the population size is 80, and the crossover and mutation probabilities are selected as 0.9 and 0.1, respectively. It is worth noting that we adopt the early-stop strategy to avoid overfitting during the training process (Yao et al. 2007), which means that the training procedure will be stopped once the performance on the test data has not improved following a fixed number of training iterations. The training process is terminated when the error of the best fitness function value in two neighboring iterations is smaller than 0.03 or the iteration number is more than 300.

3.3. Model performance

In this study, the relative error (RE) and mean absolute percentage error (MAPE), which are computed from the following equations, are employed as indicators to measure the forecasting performance.

Table 2
Predicted values and errors based on the QR-RBF neural network model (RMB·ton^{-1}).

Date	2015.5	2015.6	2015.7	2015.8	2015.9	2015.10	2015.11	2015.12
Actual price y_t	3990	4030	4015	4130	4204	4056	3759	3800
Predicted price $\hat{y}_t$	4011	3998	3964	4056	4180	4122	3701	3772
RE (%)	0.53	−0.79	−1.27	−1.79	−0.57	1.63	−1.54	−0.74
MAPE (%)	1.11							

$$RE_t = \frac{\hat{y}_t - y_t}{y_t} \tag{25}$$

$$MAPE = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{\hat{y}_t - y_t}{y_t}\right| \tag{26}$$

where y_t, $\hat{y}_t$ are the actual and predicted values at time t, respectively.

The parameters $\{\mathbf{S}(\tau), m(\tau), \mathbf{C}(\tau), \sigma(\tau), \mathbf{W}(\tau)\}$ are simultaneously optimized with the data of 2010.1–2015.4 using the proposed GDGA algorithm. Please refer to Section 2 for details. The predicted (fitting) values of the monthly soybean price for the quantiles $\tau = 0.1, 0.3, 0.6, 0.9$ are provided in Fig. 5.

Once the parameters have been optimized, the QR-RBF neural network model is used to predict the monthly soybean price from May 2015 to December 2015 (test data). The predicted soybean prices and errors against time are illustrated in Fig. 6 and Table 2. From Table 2, the largest absolute RE value was 1.79% and the MAPE was 1.11%, which suggests that the prediction accuracy was satisfactory.

3.4. Influential factors on soybean price at different quantiles

Traditional regression models focus on the mean soybean price, and the influential factor candidates are determined. In the QR-RBF neural network forecast model, the influential factors may differ when the soybean is at different price levels; that is, certain factors influence a high soybean price, while others may influence a low soybean price. The potential influential factors $(x_1(\tau), \cdots, x_8(\tau))$ are identified by the binary variables $s_i(\tau)$, $i = 1, \cdots, 8$, where $s_i(\tau) = 1$ if the factor i is selected, and is equal to zero if not.

The potential influential factors $(x_1(\tau), \cdots, x_8(\tau))$ were selected by means of the binary variables $s_i(\tau)$, $i = 1, \cdots, 8$ simultaneously with $m(\tau)$, $\mathbf{C}(\tau)$, $\sigma(\tau)$, $\mathbf{W}(\tau)$ using the hybrid algorithm for each quantile level (0.01, 0.02, …, 0.99). As an example of the results, Table 3 reports the influential factors that were selected at each decile as well as the 97th percentile. Table 3 indicates that certain influential factors were important across a range of quantiles (for example, MS and ISPDP), certain variables were important only for low quantiles (for example, DSO and CCI), and some were important only for high quantiles (for example, SIV and CPI). This finding suggests that the influential factors were diverse at different price levels, and soybean price prediction with a fixed set of predictors, as in the traditional statistical method, may not make optimal use of the available information. Similar findings were reported by Tareghian and Rasmussen (2013) in the downscaling of precipitation.

3.5. Comparison with other models

In this section, the predictive capabilities of the proposed hybrid GDGA algorithm are compared with the GA and multivariate linear regression model using the domestic soybean price data. Table 4 provides a comparison of the prediction accuracy provided by the three models on the test data.

As illustrated in Table 4, the MAPEs of the hybrid GDGA algorithm, GA, and multivariate linear regression model were 1.11%, 1.77%, and 5.86%, respectively. Obviously, the prediction from the QR-RBF neural network models with the hybrid GDGA algorithm for the soybean price exhibited optimal results, with the lowest MAPE (1.11%). In contrast, the most inferior prediction was found in the multivariate linear regression model (MAPE = 5.86%). These results indicate that the QR-RBF (ANN) exhibited superior performance in predicting the soybean price than multivariate linear regression. Similar findings were reported by Ebrahimi et al. (2017) in soil Azotobacter population prediction. Moreover, the predictive accuracy by means of the hybrid GDGA was obviously significantly improved compared to that obtained using the GA, and the deficiencies of the standard GA in solving nonlinear problems, such as prematurity and poor local searching ability, were effectively overcome. The results of this work prove that the performance of the QR-RBF model can be improved by the hybrid algorithm known as GDGA, which combines the gradient descent method and GA. This finding was also in agreement with the results in the study of Asgari et al. (2017), in which optimization of ultrasound-assisted bleaching of olive oil was accomplished by a hybrid multilayer perceptron (MLP) and GA method.

Table 3
Selected predictor variables for different quantiles for soybean price.

τ	s_1(DSO)	s_2(SIV)	s_3(GSO)	s_4(DSD)	s_5(CPI)	s_6(CCI)	s_7(MS)	s_8(ISPDP)
0.1	1	0	0	1	0	1	1	1
0.2	0	1	0	0	0	0	1	1
0.3	0	0	0	0	0	0	1	1
0.4	0	1	0	1	0	0	1	0
0.5	0	0	1	1	0	0	1	0
0.6	0	1	0	0	0	0	0	1
0.7	0	1	1	1	1	0	1	1
0.8	0	1	0	1	1	0	1	1
0.9	0	1	0	1	1	0	1	1
0.97	0	1	0	1	1	0	1	1

Table 4
Comparison of the performance between GA and the GDGA algorithm (RMB·ton^{-1}).

Date		2015.5	2015.6	2015.7	2015.8	2015.9	2015.10	2015.11	2015.12
Actual price y_t		3990	4030	4015	4130	4204	4056	3759	3800
Multivariate linear regression	Predicted price $\hat{y}_t$	3708	3725	3727	3804	3896	3896	3800	3956
	RE (%)	−7.07	−7.57	−7.17	−7.90	−7.33	−3.95	1.10	−4.79
	MAPE (%)	5.86							
GA	Predicted price $\hat{y}_t$	3952	3994	3968	3999	4109	4090	3890	3746
	RE (%)	−0.95	−0.89	−1.17	−3.17	−2.26	0.84	3.48	−1.42
	MAPE (%)	1.77							
GDGA	Predicted price $\hat{y}_t$	4011	3998	3964	4056	4180	4122	3701	3772
	RE (%)	0.53	−0.79	−1.27	−1.79	−0.57	1.63	−1.54	−0.74
	MAPE (%)	1.11							

Table 5
Comparison of convergence efficiency obtained by GA and hybrid GDGA.

Test	τ	The number of iterations to meet the error	
		GA	GDGA
1st	0.1	21	16*
	0.3	250	128*
	0.6	282	260*
	0.9	25	18*
2nd	0.1	13	24
	0.3	277	131*
	0.6	N	95*
	0.9	9	13
3rd	0.1	13	16
	0.3	N	113*
	0.6	297	81*
	0.9	6	9
4th	0.1	17	15*
	0.3	288	153*
	0.6	153	80*
	0.9	18	15*
5th	0.1	24	20*
	0.3	103	N
	0.6	238	35*
	0.9	19	16*

* means that GDGA is superior GA on the indicator, N means the model does not meet the error limit following 300 iterations.

In addition to the accuracy of the predictive models, the numbers of iterations of the QR-RBF neural network models, based on the proposed hybrid GDGA and GA, were recorded as an indicator of their complexity, and these values are listed in Table 5.

It can be observed from Table 5 that the number of iterations for meeting the same error (0.025) for the hybrid GDGA was less than that of the GA in most cases (15 trials, total 20 trials), which means that the convergence efficiency by the hybrid GDGA algorithm was obviously significantly improved. Moreover, the deficiencies of the standard GA, such as slow convergence properties near the local optima, were effectively overcome.

4. Conclusions

In this paper, a QR-RBF neural network model was proposed to forecast the soybean price in China. In the QR-RBF neural network model, quantile regression models were used to describe the distribution over the soybean price range, and RBF neural networks were used to approximate the nonlinear component of the soybean price. For a given quantile, the forecast model was a nonlinear RBF neural network, and the input variables (the variables determining the price) included certain potential influential factors and past lagged observations. In order to optimize the QR-RBF neural network model parameters, a hybrid algorithm known as GDGA was proposed, which takes advantage of the best characteristics of the global and local search approaches. The results demonstrated that: (1) the proposed hybrid GDGA algorithm provides superior forecasting performance to the multivariate linear regression and pure GA methods, and faster convergence than the pure GA method; and (2) the influential factors of soybean price are unstable at different price levels. The money supply (MS) and port distribution price of imported soybean (ISPDP) were important across a range of quantiles, the output of domestic soybean (DSO) and consumer confidence index (CCI) were important only for low quantiles, and the import volume of soybean (SIV) and consumer price index (CPI) were important only for high quantiles.

References

Adrangi, B., Chatrath, A., Raffiee, K., 2006. Price discovery in the soybean futures market. J. Bus. Econ. Res. 4 (6), 77–88.

Ahumada, H., Cornejo, M., 2016. Forecasting food prices: the case of corn, soybeans and wheat. Int. J. Forecast. 32, 838–848.

Asgari, S., Sahari, M.A., Barzegar, M., 2017. Practical modeling and optimization of ultrasound-assisted bleaching of olive oil using hybrid artificial neural network-genetic algorithm technique. Comput. Electron. Agric. 140, 422–432.

Baffes, J., Haniotis, T., 2016. What explains agricultural price movements? J. Agric. Econ. 67 (3), 706–721.

Balcombe, K., 2009. The Nature and Determinants of Volatility in Agricultural Prices. MPRA Paper No.24819. < http://mpra.ub.uni-muenchen.de/24819/ > (accessed Sep. 7, 2010).

Berwald, D., Havenner, A., 1997. Evaluating state space forecasts of soybean complex prices. Applications of Computer Aided Time Series Modeling 75–89.

Beyer, H.G., Schwefel, H.P., 2002. Evolution strategies – a comprehensive introduction. Nat. Comput. 1, 3–52.

Cannon, A.J., 2011. Quantile regression neural networks: implementation in R and application to precipitation downscaling. Comput. Geosci. 37, 1277–1284.

Cao, Z.W., Gu, H.Y., Zhou, W.M., Yan, S.Q., Ito, S., Isoda, H., 2016. Causality of future and spot grain prices between China and the US: evidence from soybean and corn markets against the surging import pressure. J. Shanghai Jiaotong Univ. (Sci.) 21 (3), 374–384.

De Freitas, N., Andrieu, C., Hojen-Sorensen, P., Niranjan, M., Gee, A., 2001. Sequential monte carlo methods for neural networks. In: Doucet, A., De Freitas, N., Gordon, N. (Eds.), Sequential Monte Carlo Methods in Practice. Springer-Verlag, New York, pp. 361–379.

Du, J., Xie, L., Schroeder, S., 2009. PIN optimal distribution of auction vehicles system: applying price forecasting, elasticity estimation, and genetic algorithms to used-vehicle distribution. Mark. Sci. 28, 637–644.

Ebrahimi, M., Sinegani, A.A.S., Sarikhani, M.R., Mohammadi, S.A., 2017. Comparison of artificial neural network and multivariate regression models for prediction of Azotobacteria population in soil under different land uses. Comput. Electron. Agric. 140, 409–421.

Kapetanakis, D.S., Manginam, E., Finn, D.P., 2017. Input variable selection for thermal load predictive models of commercial buildings. Energy Build. 137, 13–26.

Koenker, R., 2005. Quantile Regression. Cambridge University Press, New York.

Koenker, R., Bassett, G.W., 1978. Regression quantiles. Econometrica 46, 33–50.

Koenker, R., Machado, J., 1999. Goodness of fit and related inference processes for quantile regression. J. Am. Stat. Assoc. 94 (448), 1296–1310.

Korobilis, D., 2017. Quantile regression forecasts of inflation under model uncertainty. Int. J. Forecast. 33, 11–20.

Koulouriotis, D., Emiris, D., Diakoulakis, I., Zopounidis, C., 2002. Behavioristic analysis and comparative evaluation of intelligent methodologies for short-term stock price forecasting. Fuzzy Econ. Rev. 2, 23–57.

Lessmann, S., Voß, S., 2017. Car resale price forecasting: the impact of regression method, private information, and heterogeneity on forecast accuracy. Int. J. Forecast. 33, 864–877.

Li, G.Q., Xu, S.W., Li, Z.M., Sun, Y.G., Dong, X.X., 2012. Using quantile regression approach to analyze price movements of agricultural products in China. J. Integr. Agric. 11 (4), 674–683.

Li, X., Gao, L., 2016. An effective hybrid genetic algorithm and tabu search for flexible job shop scheduling problem. Int. J. Prod. Econ. 174, 93–110.

Maciejowska, K., Nowotarski, J., Weron, R., 2016. Probabilistic forecasting of electricity spot prices using factor quantile regression averaging. Int. J. Forecast. 32, 957–965.

Malone, P.J., 1968. A guide for forecasting soybean futures prices. J. Am. Oil Chem. Soc. 45 (3), A150–A152.

Pal, D., Mitra, S.K., 2017. Diesel and soybean price relationship in the USA: evidence from a quantile autoregressive distributed lag model. Empirical Econ. 52, 1–18.

Rounaghi, M.M., Abbaszadeh, M.R., Arashi, M., 2015. Stock price forecasting for companies listed on Tehran stock exchange using multivariate adaptive regression splines model and semi-parametric splines technique. Phys. A: Stat. Mech. Appl. 438, 625–633.

Schwenker, F., Kestler, H.A., Palm, G., 2001. Three learning phases for radial-basis-function networks. Neural Netw. 14, 439–458.

Statnik, J.C., Verstraete, D., 2015. Price dynamics in agricultural commodity markets: a comparison of European and US markets. Empirical Econ. 48 (3), 1103–1117.

Tareghian, R., Rasmussen, P.F., 2013. Statistical downscaling of precipitation using quantile regression. J. Hydrol. 487, 122–135.

Taylor, J.W., 2000. A quantile regression neural network approach to estimating the conditional density of multiperiod returns. J. Forecast. 19 (4), 299–311.

Wiles, P.S., Enke, D., 2015. Optimizing MACD parameters via genetic algorithms for soybean futures. Procedia Comput. Sci. 61, 85–91.

Xiong, T., Li, C.G., Bao, Y.K., Hu, Z.Y., Zhang, L., 2015. A combination method for interval forecasting of agricultural commodity futures prices. Knowl.-Based Syst. 77, 92–102.

Yao, Y., Rosasco, L., Caponnetto, A., 2007. On early stopping in gradient descent learning. Constr. Approx. 26 (2), 289–315.