Fuzzy System with Customized Subset Selection for Financial Trading Applications

W.M. Tang[1], K.F.C. Yiu[1], K.Y. Chan[2] and H. Wong[1]

[1]Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong, PR China

[2] School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Australia

Key word: subset selection, fuzzy regression, technical analysis, financial trading

Abstract

In the financial industry, identifying useful information from big data becomes a key research topic. Since the vast number of technical indicators can be captured nowadays, the indicator selection can be used to support investment decision for different financial products concurrently; however, this process is still required experience from investors. In this article, we propose a novel recommendation system which is incorporated with technical indicators. The method of fuzzy subset selection is used to feature relevant indicators which have more impact to the variation in the transaction history. The proposed method enables automatic customization of indicators for different financial products in different markets. In particular, the least absolute distance fuzzy regression with non-symmetric lower and upper bounds is proposed in order to avoid extreme values in dominating the model. Furthermore, in order to reduce computational complexity in the subset selection, the selection algorithm operates in the frequency domain for identifying and matching key patterns and peaks in transacted volumes with the technical indicators. This method performs very effective although the number of factors is much greater than the sample size. The proposed method can benefit participants in the finance markets to customize their own trading dashboard as well as set up their own trading strategies.

1. Introduction

In the financial industry, big data prevail, not only in data size, but also in data scope and variety. However, increasing number of data scopes makes the analytics more challenging. A more efficient and faster computational method is necessary to learn hidden patterns from a large volume of data or extract more useful data from the original big data. Since markets are becoming more volatile, and more capital involves in high frequency trading, significant subsets are essential to be updated from time to time in order to reflect the latest information for different financial products in different markets.

To make an investment decision, recent literature by Lo and Hasanhodzic [1] and Blume et. al. [2] showed that technical analysis can review "information" that cannot be found from the raw data in a quantitative way, since market statistics can only indicate partial information but not the full financial attributes. Technical analysis is one of the major streams to predict the future price based on the historical price patterns. Charting and technical indicators are essential to deploy technical analysis for making the trading decision. Charting method attempts to detect obvious patterns from price and volume charts, and the technical indicators trigger trading decisions when certain conditions are satisfactory for those indicators. They are usually derived from the time series of price, traded volume and trading position. Trading decisions are correlated to the volume transacted, as some of the investors use different technical indicators to make their buying and selling decisions. Hence, the trading volume increases. A similar phenomenon was suggested by Lo et. al. [3]. It suggests that the chance of detecting repetitive patterns is higher when technical indicators are used to feature the data of which the volume increases. In addition, both [4] and [5] concluded a significant relation between price and traded volume, and suggested to analyze both together in order to entail the market attributes.

Frequency domain techniques should be one of the probabilities to explore for matching cyclical patterns. Indeed spectral analysis in term of frequencies has a long history of successful applications in different areas [6]. It can perform well even when the number of factors is large relative to the sample size. Therefore, our proposed methodology uses frequency domain to screen suitable factors with cyclical patterns, which is important to extract key information in big data analytics. To the best of our knowledge, this paper is the first to customize a subset of technical indicators which can be automatically evaluated by the trading volume. In order to handle variations in the data set, fuzzy regression is deployed to model the vagueness in regression coefficients under time domain, where differences between the estimated model and

the observed data are embedded as the fuzziness of the selected explained variables. Fuzzy regression can be used to develop a robust model when data variation exists in the dynamic financial markets. For example, over-fitting may exist when more signals (or more information) are used for modelling the quantity values (the crisp values) [7]. However, the proposed fuzzy models may not need to concern the over-fitting problem, since the proposed fuzzy model not only performs the quantity prediction (the crisp value) and also the tolerance of the prediction is also given by the fuzzy model. The tolerance is indicated by the fuzziness of the prediction.

In the literature, researchers have explored the use of machine learning techniques, such as the genetic algorithm [8-10], for subset selection for fuzzy model. Subset selection by genetic algorithm approach is usually slow in convergence when the problem is complicated [11]. Moreover, selected subset may not capture the suitable cyclical pattern from dependent variable in subset selection. On the contrary, the search algorithms in frequency domain are more effective to ensure screened factors which can be incorporated with useful key patterns for appropriate subset formulation. In the modelling with fuzzy regression, fuzzy approach has been applied in analyzing or modelling large-scale and complex systems [12,13], and forecasting [14]. Tanaka et. al. [15] were the first to use triangular fuzzy numbers as the estimates of the model, where the fuzzy coefficients of the model were determined by minimizing the fuzziness spread which are used to cover the given non-fuzzy samples. Later, Savic and Pedrycz [16] proposed a two-step method to estimate the central fitted line by least square and minimize fuzziness spreads respectively; Ishibuchi and Nii [17] considered asymmetric fuzzy coefficients in regression. There are various methods to determine the fuzzy coefficients, such as neural networks [17,18], support vector regression machine [19], and genetic-algorithm-based learning [20]. Urso and Gastaldi [21] proposed an automatic procedure to fit a polynomial model to a set of data. However, optimizing the central fitted line using least square is rather problematic when extreme values exist in the data [22]; instead, using least absolute deviation (LAD) is more robust by minimizing the sum of absolute residuals. Here, LAD is applied since the LAD is more suitable for modeling traded volume data which is included with sudden increased in transactions.

To summarize, the main contribution of this article attempts to develop a system to automatically customize a useful subset of technical indicators which can aid investment decision. In the system, the fuzzy regression with non-symmetric upper and lower bounds is used for advising the investment decision and the fuzzy coefficients are determined by the LAD method. The performance of the fuzzy regression models is evaluated based on the fuzziness

of the subset which is related to the financial product. Finally, a novel subset selection algorithm is developed to select technical indicators which are correlated to the frequency peaks. The approach attempts to find significant patterns which can be aligned with our evaluation criteria.

The rest of the paper is organized as follows. Section 2 portraits the proposed methodologies to select the best subset. Section 3 describes and discusses the empirical results using five equities and one commodity future, which are traded in Hong Kong and the US. Finally, Section 4 summarizes our key findings, concludes our proposed framework, and suggests the future research.

2. The Model & Methodology

To track the past transaction variations of a financial product, we relate the transacted volume of the product with respect to the movement of technical indicators. This is motivated by the literature with support in analyzing price and volume together for better predictability. We formulate a fuzzy regression model to estimate the transacted volume, when the technical indicators are given as the regressors. When the price movements increase, transacted volume is driven through its direction; Therefore, magnitude of technical indicators, such as absolute value of their first difference ($|g(t)-g(t-1)|$) and squared value of their first difference ($(g(t)-g(t-1))^2$) can be used as the independent variables to determine the potential values when making the trading decision. The following subsections discuss the specification of regressors and estimation of coefficients of fuzzy regression when the time domain is used. We also discuss the evaluation criteria to select significant independent variables are significant to the fuzzy regression models. Finally, we discuss the notion and mechanism of the proposed factor search algorithm, which selects independent variables from their peaks in frequency domain. Based on the selected independent variables, the fuzzy regression model can be generated.

2.1 Multi-factor Fuzzy Regression

The following multi-factor models can be used to represent the relationship of some factors to the dependent variable.

$$y_t = \hat{y}_t + \varepsilon_t = \sum_{i=1}^{r_1} p_i x_{it} + \sum_{i=r_1+1}^{r_2} p_i \left|\Delta x_{it}\right| + \sum_{i=r_2+1}^{r} p_i (\Delta x_{it})^2 + \varepsilon_t, \quad (1)$$

where y_t is the dependent variable; x_{it}, $\left|\Delta x_{it}\right|$ and $(\Delta x_{it})^2$ are the independent variables. ε_t is the error term, p_i are the regression coefficients of the independent variables and r is the total number of independent variables.

When the vagueness property exists in the relation between variables, fuzzy regression model [15] can be deployed to quantify the vagueness, where the fuzziness estimated by the fuzzy regression is used to quantify the vagueness. Although the input data is observable, the regression coefficients can be used to estimate the fuzziness of the input data. The fuzzy regression model can be represented as

$$\tilde{Y}_t = \sum_{i=1}^{r_1} \tilde{A}_i x_{it} + \sum_{i=r_1+1}^{r_2} \tilde{A}_i \left|\Delta x_{it}\right| + \sum_{i=r_2+1}^{r} \tilde{A}_i (\Delta x_{it})^2 . \quad (2)$$

The fuzzy dependent variable ($\tilde{Y}_t$) is estimated based on the triangular membership function as shown in Fig. 1, and $\tilde{A}_i = (c_i^L, p_i, c_i^U)$ is the fuzzy coefficient with i=1,2,..,n, where c_i^U, c_i^L and p_i is formulated as the triangular membership (μ) for the dependent variable. Hence the fuzzy upper and lower spreads for each dependent variable can be different. These spreads can be used to model the fuzziness in transacted volume data, since upwards movement is generally larger than the lower.

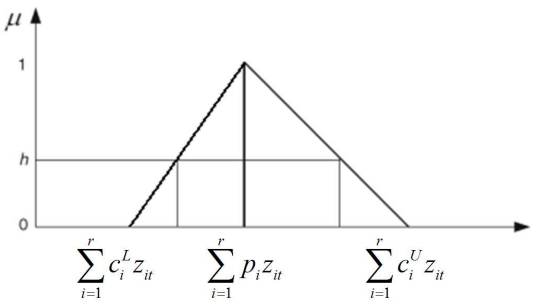

Figure 1. Fuzzy output for our proposed fuzzy model

Thus, the fuzzy regression model can be rewritten:

$$\tilde{Y}_t = \sum_{i=1}^{r_1}(c_i^L, p_i, c_i^U)x_{it} + \sum_{i=r_1+1}^{r_2}(c_i^L, p_i, c_i^U)\left|\Delta x_{it}\right| + \sum_{i=r_2+1}^{r}(c_i^L, p_i, c_i^U)(\Delta x_{it})^2 . \quad (3)$$

The central fitted line in the member function (i.e $h = 1$) is given as:

$$\tilde{Y}_t^{h=1} = \sum_{i=1}^{r_1} p_i x_{it} + \sum_{i=r_1+1}^{r_2} p_i \left|\Delta x_{it}\right| + \sum_{i=r_2+1}^{r} p_i (\Delta x_{it})^2 ;$$

where the upper bound of the fuzzy model is given as,

$$\tilde{Y}_t^U = \sum_{i=1}^{r_1}(p_i + c_i^U)\left|x_{it}\right| + \sum_{i=r_1+1}^{r_2}(p_i + c_i^U)\left|\Delta x_{it}\right| + \sum_{i=r_2+1}^{r}(p_i + c_i^U)(\Delta x_{it})^2 ,$$

and the lower bound of the fuzzy model is given as,

$$\tilde{Y}_t^L = \sum_{i=1}^{r_1}(p_i - c_i^L)\left|x_{it}\right| + \sum_{i=r_1+1}^{r_2}(p_i - c_i^L)\left|\Delta x_{it}\right| + \sum_{i=r_2+1}^{r}(p_i - c_i^L)(\Delta x_{it})^2 .$$

2.2 Fuzzy Model Estimation and Selection

In the estimation process, optimization of the fuzzy regression model is constrained with the bi-criteria requirements. The coefficients in the central fitted line and fuzzy spreads are determined by minimizing the total fuzziness where the linear programming based algorithm is used. When h is set to be zero [15], each data observation is bounded by the lower line and the upper line of the model.

The linear programming based algorithm can be used to compute central line, upper and lower spreads of the fuzzy model, where the objective value attempts to minimize the total fuzziness by the spreads c_i^U and c_i^L. The linear programming problem is formulated as follows:

$$\min \quad \sum_{i=1}^{r}\sum_{t=1}^{T}(c_i^U\left|x_{it}\right| + c_i^L\left|x_{it}\right|) \quad (4)$$

$$s.t. \quad -(1-h)\sum_{i=1}^{r} c_i^L\left|x_{it}\right| \le y_t - \sum_{i=1}^{r} p_i x_{it} \quad \forall t = 1...T$$

$$(1-h)\sum_{i=1}^{r} c_i^U\left|x_{it}\right| \ge y_t - \sum_{i=1}^{r} p_i x_{it} \quad \forall t = 1...T$$

The optimal fuzzy spread of the regression model in (1) can be used to determine the significant independent variables. When the same number of independent variables is used, the best selection of independent variables is the one with the minimum value in total fuzziness.

Based on the best subset with different number of independent variables, the most suitable model to represent the dependent variable can be determined based on their total fuzziness ($\sum_{i=1}^{r}\sum_{t=1}^{T}(c_i^U|x_{it}|+c_i^L|x_{it}|)$) and the penalty function based on Akaike information criterion (AIC) method. The algorithm is incorporated with AIC in (5) to evaluate the model when different numbers of independent variables are used.

$$\text{AIC} = T\ln\left(\frac{1}{T}\sum_{i=1}^{r}\sum_{t=1}^{T}(c_i^U|x_{it}|+c_i^L|x_{it}|)\right)+6r. \qquad (5)$$

2.3 Factor Search Algorithm

Significant independent variables which have impact to the dependent variable can be selected based on the peak patterns in the spectrum, and then the significant independent variables are constructed after evaluating the algorithmic performance which is discussed in Section 2.2. The algorithm consists of four main components namely initialization, factor selection, factor replacement and process termination. The algorithm is discussed as follows.

Step 1: Initialization

The time series samples of all dependent and independent variables are converted into the frequency domain by using the Discrete Fourier Transform (DFT). The DFT is used to convert a time series sample into a frequency domain spectrum, where the Fourier or fundamental frequencies are given as $\omega_n = \frac{n}{T}$ for $n = -\frac{T-1}{2},...,0,...,\frac{T-1}{2},\frac{T}{2}$ in the estimation, and T is the total number of data points in the sample. The linear model is defined by the Fourier frequencies as

$$x_t = \frac{1}{T}\sum_{n=-\frac{T-1}{2}}^{T/2}(a_n\sin(\frac{2\pi nt}{T})+b_n\cos(\frac{2\pi nt}{T})), \qquad for \qquad t=1,...,T,$$

$$S_x(\omega_n) = a_n + ib_n = \sum_{t=1}^{T} x_t \cdot e^{-\frac{i2\pi\omega_n t}{T}} .$$

After the DFT is performed, the time series x_t is transformed into frequency domain with the frequency spectrum $S_x(\omega_n)$ with the amplitude $A_x(\omega_n)$ at frequency ω_n, where ω_n is in the frequency range $(-\frac{1}{2}, \frac{1}{2}]$, and $S_x(\omega_n) = \overline{S_x(\omega_{-n})}$. Then, the spectral amplitudes can be calculated based on the following formulation.

$$A_x(\omega_0) = a_0$$
$$A_x(\omega_{\frac{T}{2}}) = b_{\frac{T}{2}}$$
$$A_x(\omega_n) = 2(a_n^2 + b_n^2)^{\frac{1}{2}} \qquad for \quad n = -\frac{T-1}{2}, ..., 0, ..., \frac{T}{2}.$$

In the regression analysis, the spectrum amplitudes with the frequency range $\omega_n \in (-\frac{1}{2}, \frac{1}{2}]$ are considered.

Step 2: Factor Ranking & Selection

Prioritize independent variables by ranking their total scores. The score evaluates whether the variable can match the peaks of the dependent variable. The selected independent variables attempt to fit the peak of the dependent variable. The score is calculated by using the independent variable when $i = 0$; if $i \geq 1$, the explained spectrum ($S_{\hat{y}_{X_i}}(\omega)$) of the fuzzy regression is applied in the total score calculation. For example, if X_i and X_j are selected in the model, and X_k is the independent variable for scoring, $S_{\hat{y}_{X_i}}(\omega)$ is given as:

$$S_y(\omega) = p_1 S_{X_i}(\omega) + p_2 S_{X_j}(\omega) + p_3 S_{X_k}(\omega) + S_\varepsilon(\omega)$$
$$S_{\hat{y}_{X_k}}(\omega) = p_1 S_{X_i}(\omega) + p_2 S_{X_j}(\omega) + p_3 S_{X_k}(\omega) \quad ,$$

where $i, j,$ and k are some real numbers. There are two criteria in factor ranking:

- The percentage of the total number of peaks matched between the dependent variable and the subset. For example, Fig. 2 shows three peaks exist in the

dependent variable (peaks of the black line above the red line), but only two peaks found in the independent variable (peaks of the blue dot above the red line). Hence, the percentage is 66.66% (i.e. 2/3).

- Correlation between the dependent and the subset based on the spectrum with peaks in dependent variable only, where ω_y is assumed as the peak position for the dependent variable. $\rho_{\hat{y}_{X_i},y} = \frac{\overline{S_{\hat{y}_{X_i}}(\omega_y)}^{\mathrm{T}} S_y(\omega_y)}{\sigma_{S_{\hat{y}_{X_i}}(\omega_y)} \sigma_{S_y(\omega_y)}}$ is the correlation between the dependent and the explained variable when the peak spectrum in dependent variable is only used. Fig. 2 shows that spectral values for cycle period only exist in the 200th, 25th, and 4th months.

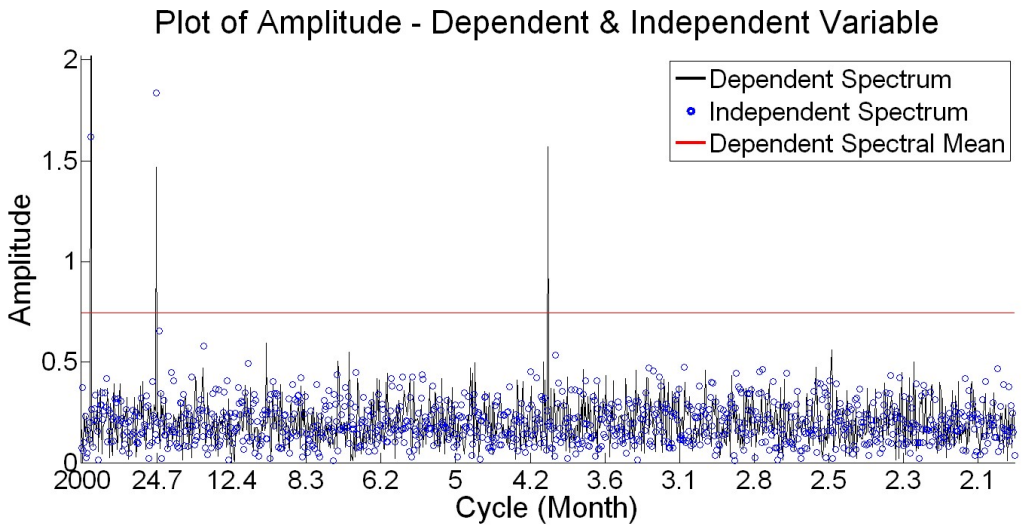

Figure 2. Score calculation to prioritize independent variables

Step 3. Factor Replacement

- Independent variable is added one by one until the pre-defined number of factor (r) is reached. Then, one additional independent variable is added to the model, and is kept in order to contribute the smallest total fuzziness when one of the independent variables is eliminated in the model. For example, if the independent variables X_i and X_j are selected in the model, then X_k is used as the independent variable in the replacement process. Three combinations are considered to generate the lowest total fuzziness:

  $y = (c_{1,1}^L, p_{1,1}, c_{1,1}^U)X_i + (c_{1,2}^L, p_{1,2}, c_{1,2}^U)X_j$

  $y = (c_{2,1}^L, p_{2,1}, c_{2,1}^U)X_i + (c_{2,2}^L, p_{2,2}, c_{2,2}^U)X_k$

  $y = (c_{3,1}^L, p_{3,1}, c_{3,1}^U)X_k + (c_{3,2}^L, p_{3,2}, c_{3,2}^U)X_j$

  $R_{1...3}$ are the total fuzziness of the three combinations respective, and total fuzziness can be computed by $\sum_{i=1}^{r}\sum_{t=1}^{T}(c_i^U |X_{it}| + c_i^L |X_{it}|)$. In the example, the model with $\min(R_{1...3})$ is kept for further processing.

Step 4. Termination

The process can be terminated when all independent variables are considered in the replacement process (step (iii)) but the total fuzziness can be reduced further (i.e. the best model still consists of X_i and X_j after considering all other independent variables in the example). If any independent variable is replaced in the model (i.e. X_i and X_j are excluded in the selected model the example), the remaining independent variables are prioritized again for the replacement process (step (iii)).

3. Results and Discussion

In the empirical analysis, the performance of our proposed factor search algorithm was evaluated by selecting technical indicators from the dataset as independent variables. There are over 1900 technical indicators (independent variables) available for search and selection; they can be classified into three main groups namely spread, trend, and oscillation indicators. In fact, some technical indicators are highly correlated as some of them contribute to the same indicator, but only different in the input parameters. Hence, the conventional method is ineffective to perform the factor selection. In the training dataset, we used raw data with 1,000 daily sample points for both transacted volume and technical indicators in the fuzzy regression, while testing dataset is the daily trading figure of the next two months. The data was collected from Jun, 2013 to 31 May, 2017 via the Bloomberg Terminal system. When performing the factor search process, we only considered spectrum with cycles not greater than 333.33 days (i.e. starting from the fourth position of the spectrum) in the search algorithm. Since the majority of economic data have similar spectral shape, considerable amplitudes are concentrated at low frequencies with some business cycles in common [23]. By disregarding low frequencies in the regression analysis, the common peaks can be avoided among the independent variables in low frequency ranges, as large amplitudes at low frequencies may mask the difference among variables in high frequencies due to relatively small amplitudes. In addition, coefficients should be in the same sign for each independent variable in the fuzzy regression, and its values in lower spread should be equal to or above zero, since transaction volume should not be a negative number; otherwise, the model should be disqualified during the search process. Finally, the number of factors chosen as the final model is determined by their AIC values in models with 5 to 10 factors, and the details of the selection process are discussed in section 2.2 – 2.3.

In this section, we will discuss the results of five equity examples which are the international giants, they are HSBC (HK: 5), Apple Inc. (US: AAPL), 3M Company (US: MMM), Home Depot Inc. (US: HD), and The Goldman Sachs Groups (US: GS). Moreover, we also include crude oil future traded in New York Mercantile Exchange in this study. As the commodities may fluctuate in different magnitude, the result may be more persuasive to show that our proposed model is applicable in both international giants and also commodities. We have also compared the results of the proposed algorithm with the ones obtained by the

commonly used heuristic method namely genetic algorithm (GA). GA is considered as it has been applied to determine parameters in fuzzy systems which have been used to perform decision making in the stock markets [6-8]. The algorithmic performance is evaluated based on the AIC for the fuzzy model which is included with the maximum of 10 indicators in the fuzzy model, and the solution selected should fulfill all the required conditions stated above. Table 1 shows the results generated by the GA method and our proposed method for the six examples mention, and Table 2 summarizes the results of t-score by comparing the performance between GA method and our proposed method. The comparison procedures and the setting of GA for indicators' selection are listed as below:

1. Each independent variable is assigned a number (i) for the selection process, where $i \geq 1$; for example, X_i is *i*-th variable in the dataset.

2. In the GA process, $\mathbf{\Omega}$ is a vector to store the variables to be selected in term of the i, when zero is stored, that mean no independent variable is selected in that particular item of the vector $\mathbf{\Omega}$. In the study, we fix the vector size as 1×10, since our proposed search method generates better results with less than 10 factors in a model in general.

3. In the GA setting, a population with 50 models in one generation is evaluated by AIC in Eq. 5, and 15,000 generations were generated for the final result. Matlab program 'ga' was run to select the best fuzzy model using GA method.

4. For each model, the GA method was run by 30 times independently, and the mean value (μ_{GA}) and standard deviation (σ_{GA}) of the AIC values are calculated respectively from the results generated.

5. In the comparison with our proposed search method to the result distribution by using GA method, the value of the t-score was calculated to show that the result of our proposed method is significantly better when compared to the GA method. Where β is the AIC value of the fuzzy model using our proposed method.

$$t = \frac{\beta - \mu_{GA}}{\sigma_{GA}}$$

Table 1. Results of the six examples for the GA method and our proposed method

Equity example	Mean Value by GA method			Our proposed method		
	Number of factors	Total Fuzziness (In million)	AIC for the fuzzy model	Number of factors	Total Fuzziness (In million)	AIC for the fuzzy model
HSBC (HK: 5)	10	38,741	17,532.4	8	36,116	17,450
Apple Inc. (US: AAPL)	10	130,515	18,747.0	6	122,209	18,657
3M Company (US: MMM)	10	3,988	15,258.8	10	3,694	15,182
Home Depot Inc. (US: HD)	10	9,496	16,126.4	10	8,951	16,067
The Goldman Sachs Groups (US: GS)	10	6,282	15,713.2	10	5,423	15,566
Crude Oil Future (CL1)	10	775	13,620.8	9	733	13,558

Table 2. Performance comparison between genetic algorithm and our proposed method

Equity example	GA		AIC for the fuzzy model using our proposed method	T-score (Significant)
	Mean AIC for the fuzzy models	Sample Standard deviation of the AIC for the fuzzy models		
HSBC (HK: 5)	17,532.4	30.86	17,450	-2.67 (99.4%)
Apple Inc. (US: AAPL)	18,747.0	44.32	18,657	-2.03 (97.4%)
3M (US: MMM)	15,258.8	22.80	15,182	-3.37 (99.9%)
Home Depot Inc. (US: HD)	16,126.4	24.19	16,067	-2.46 (99.0%)
The Goldman Sachs Groups (US: GS)	15,713.2	51.24	15,566	-2.87 (99.6%)
Crude Oil Future (CL1)	13,620.8	12.78	13,558	-4.91 (99.9%)

Based on AIC as the evaluation criteria, the results in Table 1 show that our propoosed method outperforms the GA in all the six examples, and the three examples use less number of technical indicators as the final chosen model compared to GA method. Table 2 shows that our proposed method can achieve better models compared to the GA method with over 95% significant with reference to t-score results. Then, Table 3 - 8 shows the evaluation results of the selected models when different numbers of variables are used in the six examples. Those results show that our search algorithm is able to effectively select a model with suitable number of variables. As an illustration, we will futher discuss the technical indicators chosen in the final model in subsection 3.1 (HSBC) and subsection 3.2 (3M Company) respectively, and some plots are provided to indicate fuzziness, and its spread of the final model, and also evaluate the fuzziness of the model using the testing data.

Table 3. Models Summary with Total Fuzziness and AIC Index for HSBC (HK: 5)

Number of factor in model	Total Fuzziness (In million)	AIC for the fuzzy model
5	42,298	17,590
6	41,089	17,567
7	37,440	17,480
8	36,116	17,450
9	38,735	17,526
10	37,328	17,495

Table 4. Models Summary with Total Fuzziness and AIC Index for Apple Inc. (US: AAPL)

Number of factor in model	Total Fuzziness (In million)	AIC for the fuzzy model
5	128,090	18,698
6	122,209	18,657
7	128,475	18,713
8	132,520	18,750
9	120,872	18,664
10	123,191	18,689

Table 5. Models Summary with Total Fuzziness and AIC Index for 3M Company (US: MMM)

Number of factor in model	Total Fuzziness (In million)	AIC for the fuzzy model
5	5,022	15,459
6	5,226	15,505
7	4,245	15,303
8	3,891	15,222
9	3,938	15,240
10	3,694	15,182

Table 6. Models Summary with Total Fuzziness and AIC Index for Home Depot Inc. (US: HD)

Number of factor in model	Total Fuzziness (In million)	AIC for the fuzzy model
5	10,214	16,169
6	9,667	16,120
7	10,001	16,160
8	9,132	16,075
9	9,485	16,119
10	8,951	16,067

Table 7. Models Summary with Total Fuzziness and AIC Index for The Goldman Sachs Groups (US: GS)

Number of factor in model	Total Fuzziness (In million)	AIC for the fuzzy model
5	6,950	15,784
6	6,751	15,761
7	5,759	15,608
8	5,753	15,613
9	5,617	15,595
10	5,423	15,566

Table 8. Models Summary with Total Fuzziness and AIC Index for Crude Oil Future (CL1)

Number of factor in model	Total Fuzziness (In million)	AIC for the fuzzy model
5	762	13,573
6	765	13,583
7	752	13,572
8	749	13,574
9	733	13,558
10	740	13,574

3.1 HSBC (5 HK)

The results in Table 3 shows that the 8-factor model is chosen with minimum AIC value compared to other models in the table. Table 9 summaries the details of the fuzzy regression model.

Table 9. Summary of Technical Indicators for the Selected Model

#	Technical Indicator	Coefficient Value – Central Line	Coefficient Value – Upper Line	Coefficient Value – Lower Line
		Value in Million		
1	Fear / Greed Index - 21 days period (Daily)	6.788	11.395	1.501
2	Moving Average Envelopes (Lower) - 85 days period (Daily)	769.260	1297.720	206.030
3	Hurst Exponent - 50 days period (Daily)	575.010	1062.590	93.260
4	Rate of Change - 15 days period (Daily)	0.376	0.644	0.183
5	Fear / Greed Index - 2 days period (Daily)	12.143	20.600	2.479
6	Bollinger Bands (Percentage - Standard deviation) - 65 days period (Daily)	2.318	26.213	2.138
7	Williams' %R - 70 days period (Daily)	-0.099	-0.025	-0.173
8	Exponential Moving Average - 15 days period (Daily)	16.046	31.570	14.043

Table 9 shows that only Factor 7 has negative coefficients, and the rest of technical indicators have positive coefficients in the three lines. Fig. 3 and 4 show the estimates which are generated by the fuzzy regression model. The spread of the upper line is much larger than the lower spread, since a larger spread is used to capture sudden ascents which exist in the transacted volume for some periods in the samples.

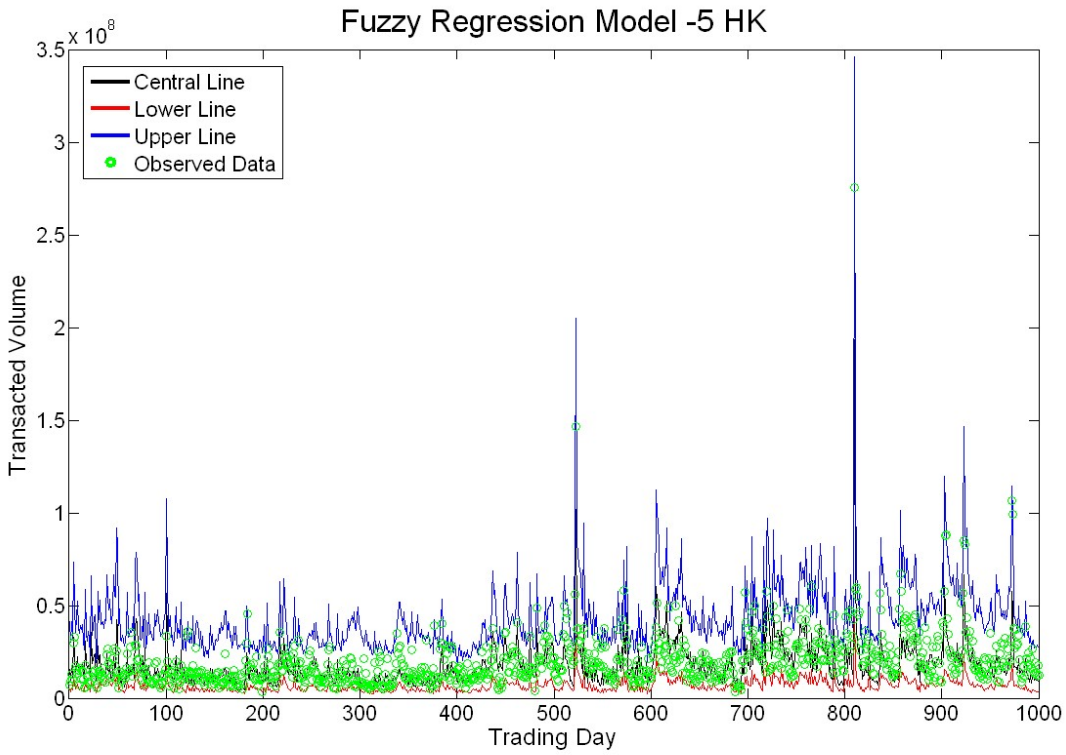

Figure 3. Selected Fuzzy Regression Model for HSBC

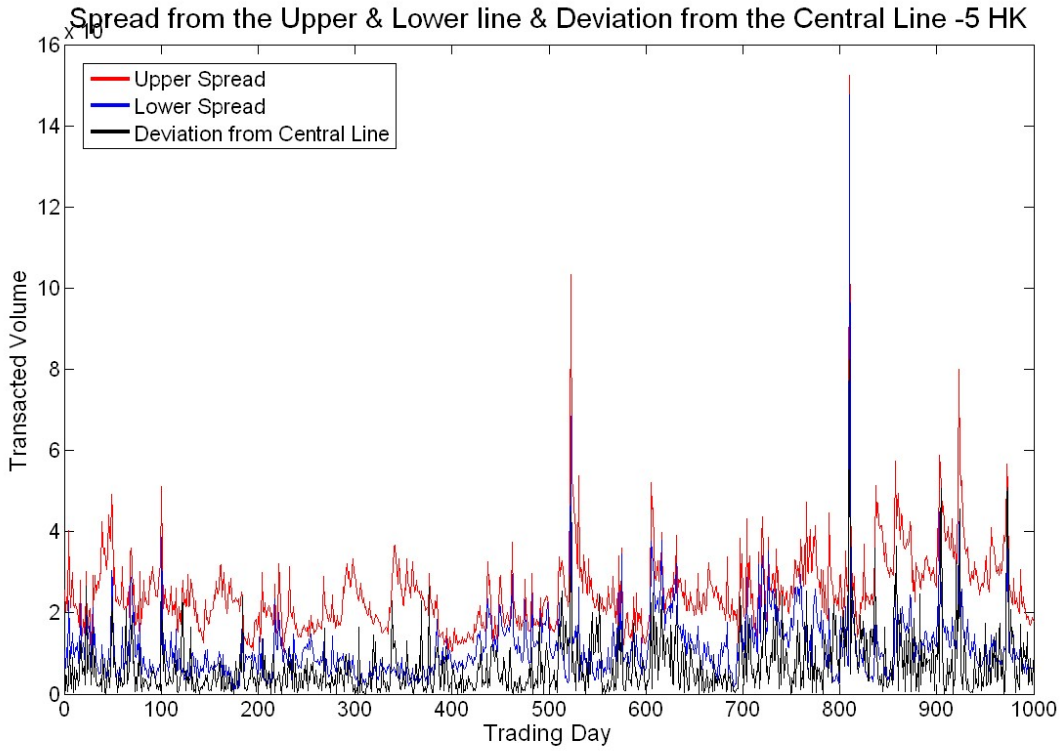

Figure 4. Spread and Deviation for Fuzzy Regression Model - HSBC

In the testing sample, Fig. 5 shows that all testing data are bounded within the lower and upper line, and also it shows that the central line can mostly follow the variation of the transacted volume.

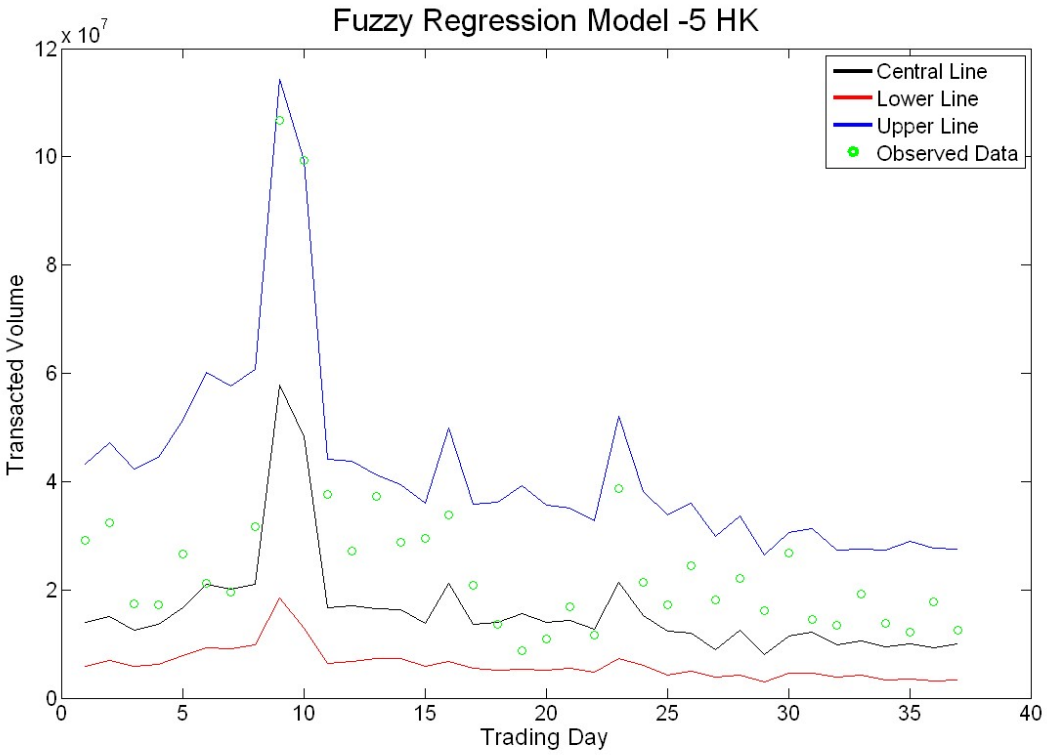

Figure 5. The Plot of testing data by Selected Fuzzy Regression Model for HSBC

When analyzing the fuzziness of the coefficients, Factors 6 and 8 have coefficients in central line which is quite close to coefficients in the lower line; however, coefficients in these factors increase significantly when they represent the upper line. Thus, these factors should have more impact to the financial product when the transacted volume goes above the central line. While the rest of the factors (i.e. factor 1, 2, 3, 4, 5 and 7) have coefficients around the middle of the upper and lower line ranges, this result shows that the impact of the three factors is related to the magnitude of the volume transacted. Based on the information, investors can utilize our model under different scenarios of the volume transacted.

3.2 3M Co. (MMM US)

Table 5 shows the results of the 10-factor model which is chosen based on the AIC value. Table 10 summaries the details of the fuzzy regression model.

Table 10. Summary of Technical Indicators for the 10-factor Model Selected

#	Technical Indicator	Coefficient Value – Central Line	Coefficient Value – Upper Line	Coefficient Value – Lower Line
		Value in Million		
1	Directional Movement Indicator (Moving Average) - 95 days period (Daily)	0.087	0.139	0.035
2	Williams' %R - 95 days period (Daily)	0.029	0.041	0.018
3	Hurst Exponent - 20 days period (Daily)	1.328	2.213	0.443
4	Channel (Retracement) - 100 days period (Daily)	1.072	1.868	0.199
5	Directional Movement Indicator (Downward Trend Strength) - 100 days period (Daily)	0.270	0.290	0.249
6	Commodity Channel Indicator - 30 days period (Daily)	-0.008	-0.008	-0.008
7	Directional Movement Indicator (Moving Average) - 15 days period (Daily)	0.176	0.207	0.138
8	Commodity Channel Indicator - 25 days period (Daily)	0.008	0.010	0.008
9	Directional Movement Indicator (Moving Average) - 25 days period (Daily)	0.615	0.876	0.351
10	Rate of Change - 50 days period (Daily)	0.499	0.831	0.160

Table 10 shows that only Factor 6 has negative coefficients, and the rest of technical indicators have positive coefficients in the three lines. Fig. 6 and 7 show the results of the fuzzy regression model. The figures show that the spread of the magnitude of the upper line is quite similar to the lower spread in diagrams, and transacted volume data are quite closed to the central line.

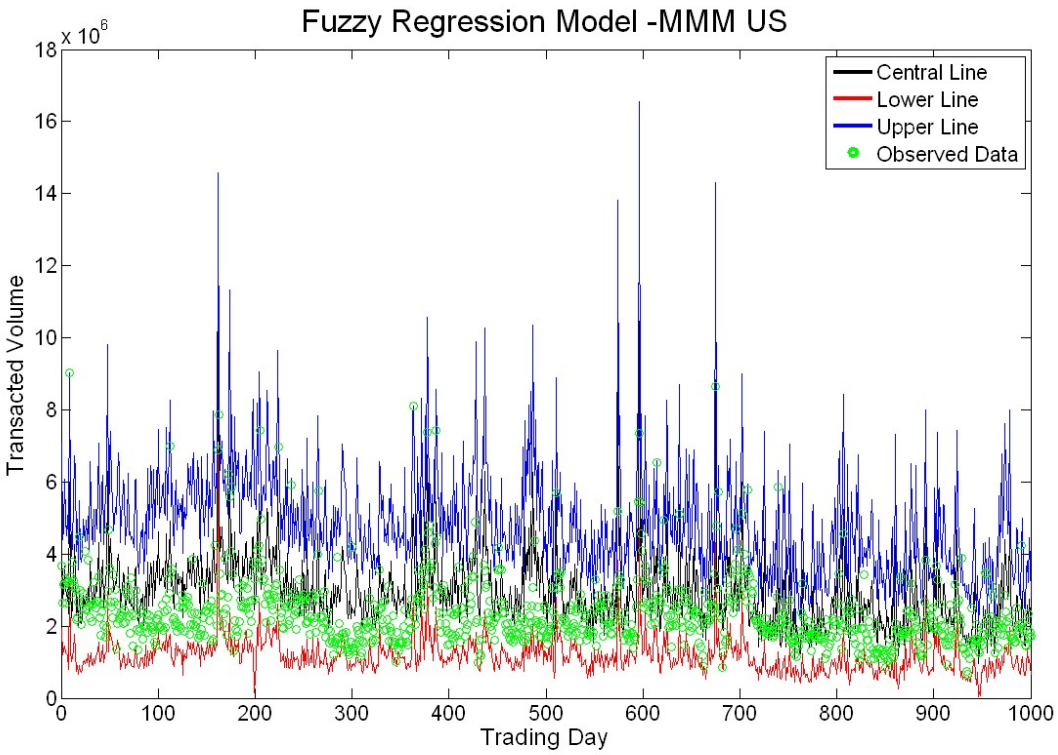

Figure 6. Selected Fuzzy Regression Model for 3M Co.

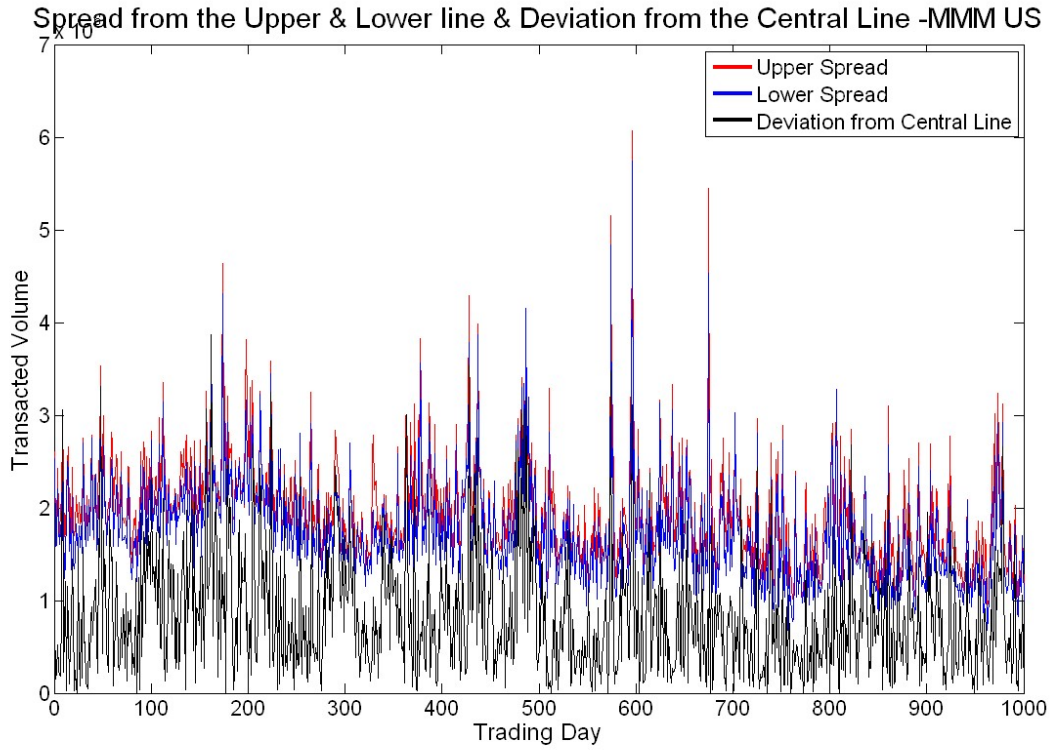

Figure 7. Spread and Deviation for Fuzzy Regression Model - 3M Co.

Fig. 8 shows the forecasting result when the selected model is used. The figure shows that the testing data are bounded within the lower and upper line; while the central line can mostly follow the variation of the transacted volume.

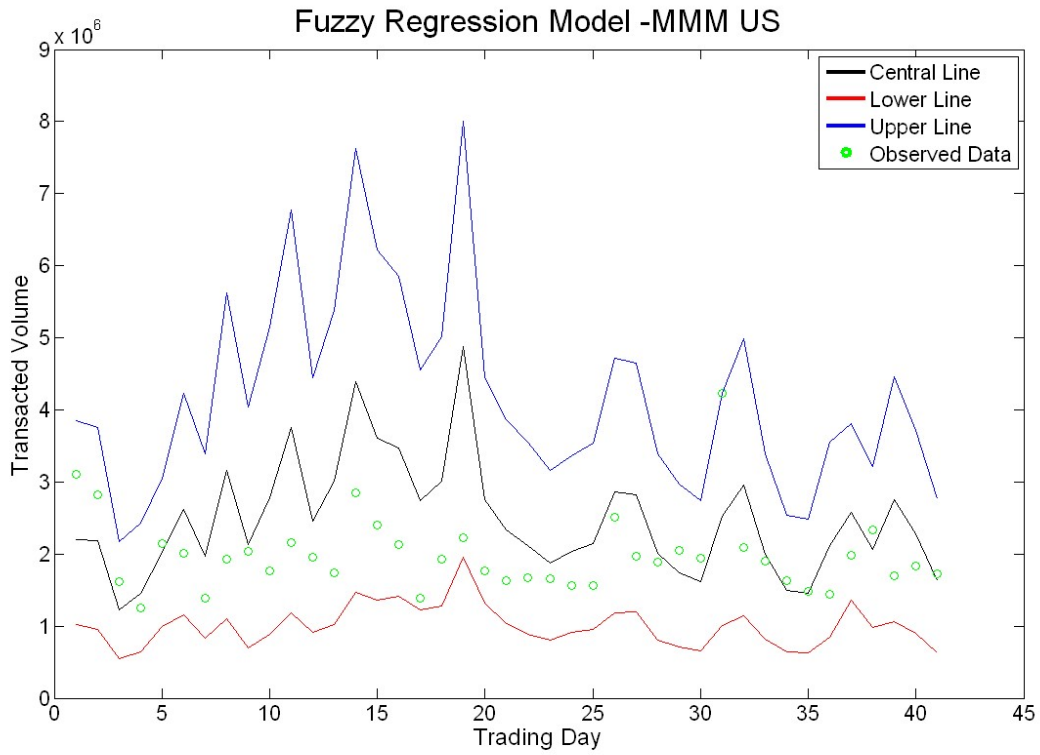

Figure 8. The Plot of Forecast data by Selected Fuzzy Regression Model for 3M Co.

When analyzing the fuzziness of the coefficients, Factor 6 and 8 have relatively small variations in their coefficients of the three lines. This result shows that the impact of these indicators is relatively insensitive to the changes in the transacted volume. While the rest of the coefficients are around the middle of the upper and lower line ranges; thus, the impact of this factor tends to follow the change in volume level. This information may provide investors more insights to which technical factors are related with different scenarios of the volume transacted.

4. Conclusion

In this article, we propose a novel recommendation system of technical indicators. A method based on fuzzy subset selection is proposed to select relevant indicators which are significant to the variation in the transaction history. The proposed approach attempts to enable automatic customization of indicators for different financial products in different markets. The approach is incorporated with the least absolute distance fuzzy regression with non-symmetric lower and upper bounds, in order to avoid extreme values in the data. Furthermore, in order to lower computational complexity in the subset selection, the factor search algorithm is developed in the frequency domain. Also the frequency domain based algorithm attempts to identify and match key patterns and peaks in transacted volumes with the technical indicators. Equity and commodity future examples, which are traded in Hong Kong and the US, have been considered and discussed to illustrate the subset selection process. In these examples, a large number of technical indicators with different parameters have been considered. The results demonstrated the effectiveness of the resulting models using our proposed search method, which were generated by the fuzzy regression technique. The resulting models are more capable of indicating more specific characteristics for the individual stock, and are more capable of relating certain technical indicators with the trading volume. For the future research, we will further extend the approach to analyze high frequency data and formulate trading strategies with reference to the selected subset under fuzzy environment.

5. Reference

1. Lo, A.W., Hasanhodzic, J.: The heretics of finance: Conversations with leading practitioners of technical analysis, vol. 16. John Wiley and Sons, (2010)
2. Blume, L., Easley, D., O'hara, M.: Market statistics and technical analysis: The role of volume. The Journal of Finance **49**(1), 153-181 (1994).
3. Lo, A.W., Mamaysky, H., Wang, J.: Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. The journal of finance **55**(4), 1705-1765 (2000).
4. Smirlock, M., Starks, L.: An empirical analysis of the stock price-volume relationship. Journal of Banking & Finance **12**(1), 31-41 (1988).
5. Lo, A.W., Wang, J.: Stock market trading volume. Handbook of financial econometrics **2**, 241-342 (2009).
6. Brillinger, D.R.: The digital rainbow: some history and applications of numerical spectrum analysis. Canadian Journal of Statistics **21**(1), 1-19 (1993).
7. Novy-Marx, R.: Testing strategies based on multiple signals. Working Paper (2016).
8. Brill, F.Z., Brown, D.E., Martin, W.N.: Fast generic selection of features for neural network classifiers. IEEE Transactions on Neural Networks **3**(2), 324-328 (1992).
9. Raymer, M.L., Punch, W.F., Goodman, E.D., Kuhn, L.A., Jain, A.K.: Dimensionality reduction using genetic algorithms. IEEE transactions on evolutionary computation **4**(2), 164-171 (2000).
10. Huang, C.-F., Chang, B.R., Cheng, D.-W., Chang, C.-H.: Feature Selection and Parameter Optimization of a Fuzzy-based Stock Selection Model Using Genetic Algorithms. International Journal of Fuzzy Systems **14**(1) (2012).
11. Tsang, E.C., Yeung, D.S., Wang, X.: OFFSS: optimal fuzzy-valued feature subset selection. IEEE transactions on fuzzy systems **11**(2), 202-213 (2003).
12. Papadopoulos, B., Tsagarakis, K.P., Yannopoulos, A.: Cost and land functions for wastewater treatment projects: Typical simple linear regression versus fuzzy linear regression. Journal of environmental engineering **133**(6), 581-586 (2007).
13. He, T., Lu, Q.: Fuzzy Varying Coefficient Bilinear Regression of Yield Series. Journal of Data Analysis and Information Processing **3**(03), 43 (2015).
14. Chen, S.-P., Dang, J.-F.: A variable spread fuzzy linear regression model with higher explanatory power and forecasting accuracy. Information Sciences **178**(20), 3973-3988 (2008).
15. Asai, H.T.-S.U.-K.: Linear regression analysis with fuzzy model. IEEE Transaction Systems Man and Cybermatics **12**(6), 903-907 (1982).
16. Savic, D.A., Pedrycz, W.: Evaluation of fuzzy linear regression models. Fuzzy sets and systems **39**(1), 51-63 (1991).
17. Ishibuchi, H., Nii, M.: Fuzzy regression using asymmetric fuzzy coefficients and fuzzified neural networks. Fuzzy Sets and Systems **119**(2), 273-290 (2001).
18. Ishibuchi, H., Tanaka, H.: Fuzzy regression analysis using neural networks. Fuzzy sets and systems **50**(3), 257-265 (1992).
19. Hao, P.-Y., Chiang, J.-H.: Fuzzy regression analysis by support vector learning approach. IEEE Transactions on Fuzzy Systems **16**(2), 428-441 (2008).
20. Hu, Y.-C.: Functional-link nets with genetic-algorithm-based learning for robust nonlinear interval regression analysis. Neurocomputing **72**(7-9), 1808-1816 (2009).
21. D'Urso, P., Gastaldi, T.: An “orderwise” polynomial regression procedure for fuzzy data. Fuzzy Sets and Systems **130**(1), 1-19 (2002).
22. Bingham, N.H., Fry, J.M.: Regression: Linear models in statistics. Springer Science & Business Media, (2010)
23. Granger, C.W.: The typical spectral shape of an economic variable. Econometrica: Journal of the Econometric Society, 150-161 (1966).