Decision Support System based on Fuzzy Logic for Assessment of Expected Corporate Income Performance

Arthur Yosef[1], **Eli Shnaider**[2] , **Rimona Palas**[3] **and Amos Baranes**[4]

[1]Tel Aviv-Yaffo Academic College, Israel, ✉ yusupoa@yahoo.com
[2]Peres Academic Center, Israel, ✉ eli-sh@012.net.il
[3]College of Law and Business, Israel, ✉ rimona@mitzuv.com
[4]Peres Academic Center, Israel, ✉ amos@drbaranes.com

Abstract: This study presents a decision-support method to estimate (in general terms) the performance of corporate Operating Income Margin for next year. It is based on a unique combination of cross section model and the rules-based evaluation mechanism. The model is constructed as follows: its dependent variable (Operating Income Margin) is one year ahead vs. the corresponding explanatory variables. This structure of the model allows us to view its explanatory variables as reflecting a financial potential of the corporations. The evaluation component involves applying a set of rules, that are designed to identify companies, where the status of their "potential" clearly points to a very high possibility for some predefined types of outcome.

For the method presented here to be successful, it is necessary to utilize highly reliable (even if it is "Fuzzy") modeling method. In addition, all the possible proxy variables are included in the model, in order not to miss valuable information.

In the present study we apply Soft Regression (a Soft Computing modeling tool based on Fuzzy Logic), and utilize all the available proxy variables (all the available data representing each factor) by creating intervals of values. Advantages of utilizing Soft Regression, and the intervals' based modeling are extensively discussed. Modeling results for five consecutive years are consistent and stable, thus indicating high degree of reliability. Testing of the decision-support method indicate very high success rate (for the stock market related domain), the lowest being 87.9%

Keywords: Modeling corporate earnings, financial ratios, corporate performance evaluation, decision support, fuzzy logic

Introduction

Predicting approximate direction of corporate future earnings, as a way to determine the future value of the firms, is a major endeavor for investors (Lev & Gu, 2016). In the present study we introduce a decision support method to assess the performance of future corporate income in general terms. For example, if the decision makers find out in very broad terms, that the corporate income of a given firm is expected (for the next year) to be at least as good as at the present year, or better, then it is a very critical information for the potential investors, provided it is reliable. Similarly important information –if the potential investors find out (with high degree of confidence) that a given company's income will not improve next year, and can only go down. Additional, very useful information could be of the type: if a given company has been so far categorized as a "winner", will it stay in that category in the next year as well?, etc. Hence, in contrast to numerous conventional methods that are attempting (mostly unsuccessfully) to forecast numerical value of corporate income for next year, we trade off the attempt to predict a specific numerical value (where the risk of failure is high) for a more general (fuzzy) description or category of outcome. As long as the success rate of such fuzzy output is high (see results below), such a broad, qualitative description is more beneficial to the potential investors in comparison to forecasted numerical values generated by conventional methods, which are usually characterized by a low reliability. The low reliability of results generated by conventional tools and methods, are expected due to some technical deficiencies of such methods (see explanation below), in particular in highly uncertain domains, such as financial markets.

The decision support system presented in this study is not expected to identify potential outcome for all the corporations in our sample. However, the method presented here identifies specific cases (based on existence of some clear-cut predefined conditions), where there is a high confidence regarding the expected outcome.

There are hundreds of different financial ratios, that could be computed from the stock market related data. It is necessary to select the relevant financial ratios, such that: (a) the selected financial ratios must be logically related to the dependent variable, and (b) such relations must be significant.

An important factor which makes it very difficult to build a model of financial ratios, while utilizing conventional modeling tools, is that there is substantial mathematical correlation among various financial ratios. Multicollinearity does not allow to incorporate all the relevant financial ratios into the same model, thus undermining the reliability of the results due to model misspecification. In addition, the exclusion of some explanatory variables due to multicollinearity makes the computation of relative importance of the whole set of variables incorrect, because the excluded variables are implicitly assigned weight of zero even in the cases of important variables, that could have become significant if model specification would be different.

Therefore, in this study we utilize "Soft Regression" (SR), which is a Soft Computing modeling tool based on Fuzzy Logic. SR does not require independence of explanatory variables and thus multicollinearity does not affect the reliability of modeling results. In other words, SR allows to incorporate explanatory variables into the same equation even if they are mathematically correlated. In addition, SR generates reliable computation of relative importance of explanatory variables among themselves (Shnaider & Yosef, 2018a).

Once the modeling process is complete (cross section model based on the data from hundreds of corporations - see details below), the next step is to use the results (generated by the model) for construction of a rule based component.The set of rules is designed to identify corporations where there is a high possibility for fuzzily described, predefined performance (of the next year operating income). The rules are based on the weighted combination of explanatory variables, while the weights are derived from the model. The rules are applied to each firm separately - to assess its next year performance.

Financial reports' ratios are computed utilizing the eXtensible Business Reporting Language (XBRL). The Securities Exchange Commission (SEC) has mandated, since 2011, XBRL format for reporting of financial data for all publicly traded companies. XBRL facilitates information gathering and processing, since it is easily downloaded from the internet and translated into EXCEL format.

Literature Survey

Ou & Penman, (1989) were the first researchers to focus on the usefulness of accounting information to explain directional movement of earnings. The study evaluated whether accounting information can consequently be used as the basis for profitable investment strategy. They began with 58 financial ratios but through feature selection based on Logistic Regression eventually ended up with about 15 ratios to create the final version of their model. The articles that followed, used similar statistical methods (mostly traditional linear statistical modeling techniques) with mixed results (Bernard et al., 1997; Bird et al., 2001; Holthausen & Larcker, 1992; Setiono & Strong, 1998; Stober, 1992).

Due to the fact that not all ratios are informative and can provide high discrimination power, it is necessary to filter out unrepresentative variables from a given data set through feature selection techniques (Tsai & Hsiao, 2010). Many well-known feature selection/extraction techniques have been used to identify relevant financial ratios, such as: correlation matrix (Atiya, 2001), t-test, factor analysis, and logistic regression (LR) (Shin, Lee, & Kim, 2005). LR has also been a key method in feature selection involving accounting ratios (Bernard, Thomas, & Wahlen, 1997; Bird, Gerlach, & Hall, 2001; Holthausen & Larcker, 1992; J. A. Ou & Penman, 1989; Setiono & Strong, 1998; Stober, 1992). Of course, one must always keep in mind that the binary nature of the LR creates difficulties to distinguish among various degrees of successful or failing performance.

Recently the Ou & Penman, (1989) methodology was implemented on XBRL data (Baranes & Palas, 2017b). Machine learnings, specifically Support Vector Machines (Machine Learning method) with PCA as the feature selection method was also applied (Baranes & Palas, 2019).

Chandwani & Saluja, (2014) claimed that financial time series data are characterized by noise, chaos and a high degree of uncertainty.

Soft Regression (SR) is an Artificial Intelligence (Soft Computing) modeling tool based on Fuzzy Logic. It has been evolving since 1990s (Kandel, Last, & Bunke, 2001). Comparison of SR to Multivariate Regression method appears in (Yosef, Haruvy, & Shnaider, 2015). Computing relative importance of explanatory variables (RELIMP) by utilizing SR versus traditional regression methods is presented in (Yosef & Shnaider, 2017). The detailed explanation of RELIMP (based on SR) and evaluation of its reliability are presented in (Shnaider & Yosef, 2018a).

Extensible Business Reporting Language: XBRL (eXtensible Business Reporting Language) is a freely available and global standard designed for exchanging business information. One use of XBRL is to define and exchange financial information, such as financial statements.

The U.S Securities and Exchange Commission (SEC) has created the XBRL U.S. GAAP Financial Reporting Taxonomy. This taxonomy is a collection of accounting data concepts and rules that enables companies to present their financial reports electronically. The SEC's deployment was launched in 2008 in phases, and all public U.S. GAAP companies were required to file their financial reports using the XBRL reporting technology starting from June 15, 2011.

The Model

Dependent Variable: In our study the dependent variable is a leading variable (one period ahead) A20-Operating Income Margin (operating income divided by total revenues). That is if the explanatory variables are for year t, the dependent variable is for year $t + 1$.

Explanatory Variables: Financial ratios have played an important part in evaluating the financial condition of companies (Chen, Kung & Shimerda 1981), different ratios and a variety of different financial ratio classification systems have been suggested (Pinches, Eubank, Mingo, & Caruthers, 1975). This study follows one of the most common classifications as presented in numerous textbooks (Harrison, Horngern, Thomas, & Suwardy, 2011).
The ratios are commonly classified as follows:

1. **Liquidity**-the ability to pay for short term liabilities, current as well as liabilities which mature within the next year. The payment is expected to be in terms of present liquid assets as well as assets which are expected to become liquid within the next year.
2. **Efficiency**- Cash conversion cycle, the ability to sell inventory, collect payment from customers and pay suppliers.
3. **Solvency** -the ability to pay long term debt, liabilities which will mature after more than one year.
4. **Profitability**- the fundamental goal of the business is to earn a profit and therefore there is great importance of profitability measures. Profitability represents the company's ability to create future positive cash flows in excess of liabilities.
5. **Market ratios**- analyzing shares as an investment. Investors buy shares to earn a return on their investment, this return can be achieved in one of two ways: gains (or losses) from selling the shares and/or dividends.

The investment decision is usually based on two important general factors, risk and return. When examining the classifications presented above, the first three classifications represent the company's risk level, its ability to pay its debts and operations and survive, in the short and the long run. The last two classifications represent return to the investor, profitability represents the potential for return, while the market ratios represent the actual return.

It is necessary to find financial ratios, that can be used as proxy variables for the five general financial factors described above. The list of the proxy variables utilized in this study is:

1. **Liquidity**: Sales to Working Capital ratio represents the company's ability to generate revenue in accordance with its ability to pay its short term debt. Working capital is the difference between the company's current assets (which may be realized in the coming year) and the company's current liabilities (which need to be paid off in the coming year) and therefore represents the company's ability to pay its debt in the coming year.
2. **Efficiency**: Sales to Total Cash ratio provides information as to how efficient the company is in using its cash to generate revenues.
3. **Solvency**: The company's solvency is represented by the Interest Coverage Ratio and the Cash Flow from Operations to total debt.
    a. The first ratio measures the proportionate amount of operating income that is used to cover interest payments, since these interest payments are usually made on a long-term basis, they are often treated as an ongoing expense. This ratio is also used to indicate the company's capitalization efficiency, the impact of the company's choices in raising capital.
    b. The second ratio representing the company’s solvency is: Cash Flow from operations to total debt. It indicates how long it will take the company to pay off all of its debt if it devotes all of its cash flow from operations to debt repayment, this ratio provides a snapshot of the overall financial health of the company.

4. **Profitability**: It is reasonable that the classification which will be most prominent and have the most significant variables, are profitability. Profitability ratios represent the relative measures of the earnings (profits) the company created, and therefore have the closest association with the earnings themselves.
5. **Market Ratios**: The market ratios represent the relationship between the company's actual profits and the investor returns (gains from an increase in the price of the shares or from the distribution of dividends). There is a representation of both the gains from shares (P/E ratio) and the gains from dividends (Payment of dividends as a % of operating cash flow).

The list of variables utilized in the model:

Dependent variable:

A20-Next period Operating Income Margin

Proxy explanatory variables:

Liquidity:
A54-Sales to total working capital

Efficiency:
A52-Sales to total cash

Solvency:
A47-Times Interest Earned
A73-Cash From Operations (CFO) to Total Debt

Profitability:
A35-ROA (Return on Assets)
A50-Pre-taxes income over Sales
A51-Net Profit Margin:
A57-Research & Development Expense to Sales
A59-Operating Income to Total assets
A70-EBITDA Margin Ratio: EDITDA (Earnings Before Interest, Taxes, Depreciation and Amortization) to Total Revenue

Market ratios:
A21-P/E Ratio
A75-Payment Of Dividends as % of OCF (Operating Cash Flow)

Data

Using the NASDAQ company list (http://www.nasdaq.com/screening/company-list.aspx) all 6,670 companies (tickers) listed on all of the three major US stock exchanges (AMEX, NASDAQ, and NYSE) were found.

The annual financial data was obtained using XBRL Analyst (created by FinDynamics); an Excel plugin that allows users to access the company's XBRL tagged data from its XBRL SEC filing via the XBRL US database. Using this software not only allows easy access and analysis of the data but also allows the calculation of any missing balances. For example, the balance reported in each XBRL filing for total liabilities is not available on the original XBRL filing but is extracted and calculated using the XBRL Analyst. The obtained data were annual filings from 2012 to 2017 (6 years).

The process of selecting a subset of relevant features to be used in the model construction, was also used to create the financial ratios. 6,670 tickers were originally identified using the NASDAQ company list and 2,561 tickers were removed. The reasons for removal: there wasn't any data reported in XBRL format, tickers for non-common stocks, tickers for companies with IPO's between 2012 and 2017, and tickers for companies with more than one ticker (the same CIK).

The final sample included 4,109 companies (61.6% of all tickers listed) that were publicly traded on 2017. For the purpose of this study it was decided to examine only one industry, the manufacturing industry (SIC code 2000-3999), which represents the largest industry, 1,597 tickers, 38.9% of the total sample out of which 1,585 reported operating income.

Based on the Ou & Penman, (1989) study, 58 variables were extracted from the XBRL filing data base (Appendix 1). It should be noted that some of the variables had to be calculated from the original filing, whereas some other variables were already calculated as part of the XBRL Analyst tool.

Data preparation

1. **Constructing matrix of intervals:** When preparing data for modeling, every variable is treated as a numerical vector. In other words, it is a column of numbers. In the case when several numerical vectors which supposedly represent the same thing, we can construct a matrix, such that each numerical vector is a column in that matrix. Even though all the proxy variables for a given financial variable are supposedly representing the same factor, it is still desirable to include all of them in the model, because each proxy variable represents a unique aspect (unique angle) in measuring of a given financial category. By including all the proxy variables, we utilize all the available information regarding that financial factor. For example, in our case study, we utilize 6 proxy variables representing profitability (A35, A50, A51, A57, A59 and A70). Thus, we create a matrix, where the 6 proxy variables appear as columns in a matrix, while each row represents data for a given company. Therefore, for companies, which appear in all 6 variables (columns), we can view it as a set of values, which consists of 6 numbers. Each such set can be converted into an interval which is defined by its smallest value and its largest value. Obviously, similar intervals can be created for companies that do not appear in all 6 variables – those will be intervals based on fewer measurements. In the extreme cases, where a company appears in only one numerical vector (out of 6), then its range will be represented by the same value as the minimum and the maximum. Thus, the matrix of 6 columns can be transformed into the matrix of two columns: column of minimum values for each row, and column of maximum values of each row. Hence, the matrix of minimum and maximum values essentially contains all the information of the proxy variables, but it reduces modeling complexity, helps to alleviate some problems due to missing observations, etc. (for advantages of utilizing intervals in modeling see Shnaider and Yosef (2018b).

   There is a very important issue that must be addressed when constructing intervals as discussed above: it is critical to make sure that before we construct the intervals, all the proxy variables representing the same factor are converted into the same scale, otherwise the interval is distorted and meaningless. In general, bringing all the different numerical vectors into the same scale is possible by recalculating all of them based on the same reference point. Selected reference point should be reasonable and reliable. When utilizing method based on fuzzy logic (such as Soft Regression, Fuzzy linear regression, Fuzzy Cognitive Maps, etc.), defining all the numerical vectors in terms of membership in the same fuzzy set is an additional (and very effective) way to address the scale problem.

   Once all the values of the matrix are converted into the grades of membership, then we can sort values in each row from the smallest to the largest since now they are all members of the same fuzzy set (and therefore measurable on the same scale). This way, for every row, we construct intervals consisting of grades of membership. The process of converting the data from the regular measurements into the grades of membership in a given fuzzy set involves normalizing the data according to a predefined membership function. It is of crucial importance that the membership function and thus the corresponding data normalizing process will be based on solid human reasoning and common sense – as demonstrated in the next section.

2. **Normalizing the data - the implementation:** All the companies in the data base were divided into three groups, based on the dependent variable Operating Income Margin:
   a. The group of "Winners": contains companies which were continuously profitable, reported a positive net income on annual basis for every year between 2012 to 2016 (including 2012 and 2016).
   b. The group of "Losers": contains companies that reported a negative net income on annual basis for every year between 2012 to 2016 (including 2012 and 2016).
   c. All the remaining companies, the "Middle Group".

   Let's assume that we have c numerical vectors ($\{X_l\}_{l=1}^{c}$), each consisting of n elements ($X_l = (x_{1,l}, x_{2,l}, \ldots, x_{n,l})$). The normalizing of the data is performed based on the equation (1):

$$\mu_l(x_{k,l}) = \begin{cases} 0 & , x_{k,l} < min_l \\ \frac{x_{k,l} - min_l}{max_l - min_l} & , min_l \le x_{k,l} \le max_l \text{ for all } k = 1,2,\dots,n \\ 1 & , x_{k,l} > max_l \end{cases} \quad (1)$$

where μ_l is a membership function for all $l = 1,2,\dots,c$, max_l is a cut-off value defining a full membership in the fuzzy set "winners", and min_l is a cut-off value defining no membership at all in the fuzzy set "winners" (because the relevant companies are full members in the fuzzy set "Losers".

max_l was determined as follows (for every variable): the values of the companies belonging to the group of "Winners" were arranged from the lowest to the highest, and then divided into four quarters. The highest value of the lowest quarter (i.e., the 25th percentile or the first quartile) was selected as max_l.
min_l was determined as follows (for every variable): the values of the companies belonging to the group of "Losers" were arranged from lowest to highest, and then divided into four quarters. The lowest value of the highest quarter (i.e., the 75th percentile or the third quartile) was selected as min_l.

Justification: As was stated above, the process must be in line with human logic and common sense and modelers should be capable of defending their decisions. For example, for max_l, instead of selection made above, we could have selected the minimum measure of the all companies in the category of "Winners". Such selection would include all the companies in the group "Winners" as full members in the Fuzzy Set of "Winners". However, such a selection would include unknown amount of borderline cases, whose corresponding values of explanatory variables (which reflect their performance) often intermix with the more successful performers from the "Middle Group". On the other hand, by defining only the higher 75% of the "Winners" as the full members of the fuzzy set representing the Winner Group, the vast majority of the borderline cases are prevented from being considered as full members of the group, thus making the identification of the group more clear-cut. Moreover, the 25% of the "Winners" which are not assigned the value of 1, which represents the full membership in the fuzzy set, will be assigned grade of membership close to 1, still reflecting accurately the relative strength of their performance, and hence the integrity of the data is maintained. All this in contrast to the Boolean method, where all those who are not assigned the value of 1, get value of 0, thus becoming an important source of distortions in numerous statistical methods.

Similar, but inverse reasoning applies to min_l.

Every variable where $max_l > min_l$, is a valid variable ready for being normalized, utilizing equation (1). If max_l is not greater than min_l, this means that this explanatory variable, if related to the dependent variable, will be inversely related. In other words, this is a variable characterized by large values in the group of "Losers" and small values in the group of "Winners". In this case we define max_l and min_l as follows:

- max_l (for every variable for every year): the values of the companies belonging to the group of "Losers" are arranged from lowest to highest, and then divided into four quarters. The highest value of the lowest quarter was selected as max_l.
- min_l (for every variable for every year): the values of the companies belonging to the group of "Winners" are arranged from lowest to highest, and then divided into four quarters. The lowest value of the highest quarter was selected as min_l.

3. **Outliers vs Central Tendency of Intervals**: By including all the available information (including unavoidable outliers) we will necessarily end up in some cases with intervals that are very extensive, and therefore not very helpful for modeling. Normalizing the data, which is part of the process to convert the numerical vectors into fuzzy sets allows us to reduce and contain to some extent the problem of outliers by redefining each variable in terms of membership of its elements in pre-defined fuzzy set. However, in order to perform successful modeling, it is desirable to identify more limited portion of the interval, containing its core area, reflecting (even in approximate terms) its central tendency. Narrow intervals do not differ much from their core central tendency. However, very extensive intervals require additional work of interval reduction in order (if and when possible) to create a better reflection of their central tendency.

   Therefore, we undertook additional steps designed to reduce the extent of the original range, while attempting at the same time to assure that minimum of valuable information is lost. In other words, the purpose of range reduction process is to eliminate outliers as carefully as possible without distorting the central tendency

of the interval in the process. The algorithm of interval reduction is presented below (For more details see Shnaider and Yosef 2018b).

Range Reduction Algorithm (RRA)

a) Let's assume that we have c numerical vectors, each consisting of n elements (In other words, we have a matrix $\mathbf{A} = \left(x_{k,l}\right)_{n\times c}$ where n is a number of rows and c is a number of columns). First, we normalize all the numerical vectors by applying relevant membership function, such that the resulting elements of the numerical vectors will consist of values [0,1], which represent degree of membership in the same fuzzy set, i.e., Fuzzy matrix $\mathbf{A}$ is $\tilde{\mathbf{A}} = \left(\tilde{x}_{k,l}\right)_{n\times c}$ where

$$\tilde{x}_{k,l} = \mu_l(x_{k,l}) \tag{2}$$

for all $k = 1,2,\dots,n$ and μ_l is a membership function for all $l = 1,2,\dots,c$ (see (1)).

b) Sort each row of the matrix from the lowest value on the left side to the highest value on the right side while arranging all rows to be left-justified. Denote the sorted matrix as:

$$\tilde{\mathbf{A}}^{\text{Left}} = \left(\tilde{x}^{\text{L}}_{k,l}\right)_{n\times c} \tag{3}$$

Note: Following the stage above, the new matrix loses its original structure by its initial vectors. Now we have a matrix, such that in each row, the first element on the left side is the minimum value for that row, the next one is the second smallest value and so on until we reach the last value on the right side, which is the maximum for that row.

c) $col \leftarrow 2; \text{del} \leftarrow 0$

d) Consider the columns 1 and col in matrix $\tilde{\mathbf{A}}^{\text{Left}}$.

If $K = \left\{k: values_k^{\tilde{\mathbf{A}}^{\text{Left}}} > 4\right\} \neq \emptyset$ and $\frac{1}{|K|}\sum_{k\in K}|\tilde{x}^{\text{L}}_{k,1} - \tilde{x}^{\text{L}}_{k,col}| < 0.05$

where $values_k^{\tilde{\mathbf{A}}^{\text{Left}}}$ is a number of values in row k of matrix $\tilde{\mathbf{A}}^{\text{Left}}$, $|K|$ is a cardinal of the set K.

Note: In the expression $\frac{1}{|K|}\sum_{k\in K}|\tilde{x}^{\text{L}}_{k,1} - \tilde{x}^{\text{L}}_{k,col}| < 0.05$, 0.05 can be replaced by 0.01, based on specific characteristics of a given data

Then

1. delete from the column col all the elements from the rows where there are 5 measurements or more (we say that columns 1 and col are almost identical)
2. $col \leftarrow col + 1; \text{del} \leftarrow \text{del} + 1$
3. Go-to step (d)

e) Create similar matrix where all the values of $\tilde{\mathbf{A}}^{\text{Left}}$ are right-justified (Matrix is denoted by $\tilde{\mathbf{A}}^{\text{Right}} = \left(\tilde{x}^{\text{R}}_{k,l}\right)_{n\times c}$).

f) $col \leftarrow c - 1$

g) Consider the columns c and col in matrix $\tilde{\mathbf{A}}^{\text{Right}}$.

If $K = \left\{k: values_k^{\tilde{\mathbf{A}}^{\text{Right}}} > 4\right\} \neq \emptyset$ and $\frac{1}{|K|}\sum_{k\in K}|\tilde{x}^{\text{R}}_{k,c} - \tilde{x}^{\text{R}}_{k,col}| < 0.05$

where $values_k^{\tilde{\mathbf{A}}^{\text{Right}}}$ is a number of values in row k of matrix $\tilde{\mathbf{A}}^{\text{Right}}$

Then

1. delete from the column col all the elements from the rows where there are 5 measurements or more (we say that columns c and col are almost identical)
2. $col \leftarrow col - 1; \text{del} \leftarrow \text{del} + 1$
3. Go-to step (g)

h) Create a new matrix where all the values are left justified (in other words, if there are empty cells in a given row, they appear on the right-hand side of the row). The resulting matrix is denoted by $\tilde{\mathbf{B}} = \left(\tilde{b}_{k,l}\right)_{n\times\tilde{c}}$ when $\tilde{c} = c - \text{del}$

i) Create additional matrix $\tilde{\mathbf{D}} = \left(\tilde{d}_{k,l}\right)_{n\times(\tilde{c}-1)}$ such that $\tilde{d}_{k,l} = \tilde{b}_{k,l+1} - \tilde{b}_{k,l}$. (In other words, we will compute differences in the matrix $\tilde{\mathbf{B}}$ for each row k, between element $\tilde{b}_{k,l+1}$ and element $\tilde{b}_{k,l}$ for $l = 1,2,\dots,\tilde{c}-1$).

j) For any given row k having $values_k^{\tilde{\mathbf{B}}} > 4$ amount of elements, we can delete

$$\beta_k = \left\lceil 0.2 \cdot values_k^{\tilde{\mathbf{B}}} \right\rceil \tag{4}$$

elements, where $values_k^{\tilde{\mathbf{B}}}$ is a number of values in row k of matrix $\tilde{\mathbf{B}}$ and $\lceil\cdot\rceil$ is a ceiling function. We evaluate

$$max\left\{\sum_{l=1}^{\beta_k}\tilde{d}_{k,l}\,,\sum_{l=0}^{\beta_k-1}\tilde{d}_{k,last-l}\,,\sum_{l=1}^{\delta}\tilde{d}_{k,l}+\sum_{l=0}^{\gamma-1}\tilde{d}_{k,last-l}\ \text{where}\ \delta+\gamma=\beta_k\right\} \quad (5)$$

when $\tilde{d}_{k,last}$- is the last element of the interval in row k of matrix $\tilde{\mathbf{D}}$. In (5) the first term represents, for any given row k, sum of β_k elements on the left side of the matrix, the second term represents sum of $\beta_k$elements on the right side of the matrix, and the third term represents all the possible permutations of sums of elements from the left side and the right side such that the total amount of elements remains β_k. Then delete from matrix $\tilde{\mathbf{B}}$, $\beta_k$elements that correspond to the maximum term found in (5).

The Matrix resulting from the reducing range of individual intervals, is denoted as

$$\tilde{\mathbf{R}}=\left(\tilde{r}_{k,l}\right)_{n\times c^*} \quad (6)$$

where $c^*=\max_{k=0,1,\dots,n}\{\tilde{c}-\beta_k\}$.

k) For every row k, find the new interval by subtracting $Range_k=\tilde{r}_{k,last}-\tilde{r}_{k,1}$ ($\tilde{r}_{k,last}$ is the last value of interval in row k of $\tilde{\mathbf{R}}$). We consider interval size of $Range_k>0.25$ as excessively large (See note below).

l) $s\leftarrow 1$

If $Range_k>0.25$ and $values_k^{\tilde{\mathbf{R}}}>4$

where $values_k^{\tilde{\mathbf{R}}}$ is a number of values in row k of matrix $\tilde{\mathbf{R}}$

Then

If $\tilde{r}_{k,s+1}-\tilde{r}_{k,s}>\tilde{r}_{k,last}-\tilde{r}_{k,last-1}$

Then

1. delete $\tilde{r}_{k,1}$
2. $values_k^{\tilde{\mathbf{R}}}\leftarrow values_k^{\tilde{\mathbf{R}}}-1$
3. $s\leftarrow s+1$

Else

1. delete $\tilde{r}_{k,last}$
2. $values_k^{\tilde{\mathbf{R}}}\leftarrow values_k^{\tilde{\mathbf{R}}}-1$
3. $last\leftarrow last-1$

m) If $Range_k=\tilde{r}_{k,last}-\tilde{r}_{k,1}$ (the new range after deletion) is still > 0.25 then if the interval greatly exceeds 0.25, the user might consider deleting that row from the matrix. The user may leave the new interval as is, if it exceeds 0.25 only to a minor degree. The selection of 0.25 is based on individual reasoning by a modeling professional and could differ based on circumstances and constrains.

Note: the very wide range (above 0.25– which is a large portion of the entire numerical domain [0,1]) means that there must be some very serious problem of measurement or error associated with that particular row in the matrix. We must keep in mind, that measurements appearing in a given row represent (from our perspective) different measurements of the same thing and therefore such a wide discrepancy is unreasonable. If such discrepancies appear in just few rows and are relatively small fraction of our data, then we can justify deletion of these intervals (which is a common practice in modeling when there is a reasonable suspicion of problematic data). If, on the other hand, even after applying interval reduction algorithm, - large portion of intervals are still characterized by excessive ranges, then we obviously have a modeling problem, which requires to re-specify the model - redefine the variables included in the model or redefine the set of proxy variables which supposedly represent a given factor.

The matrix created as a result of applying RRA procedure presented above, is denoted as

$$\tilde{\mathbf{A}}^{\mathbf{RRA}}=\left(\tilde{x}_{k,l}^{RRA}\right)_{n^*\times c^*} \quad (7)$$

where c^*, n^* are a number of rows and columns that remain following the RRA process.

Table 1: Data Preparation Using Intervals (Year 2012)

2012	Amount of Columns	Amount of Companies	Amount of Excessive Ranges Before Reduction	Amount of Excessive Ranges After Reduction
Profitability	6	1161	285	32
Solvency	2	978	108	108
Market ratios	2	1023	144	144

Total number of companies is 1161

Table 1 demonstrates data preparation process using as an example the year 2012. In our study, only one variable (profitability) is represented by 6 proxy variables, and thus is applicable to the process of interval reduction. We can see, that when intervals representing all the proxy variables were constructed, out of 1161 intervals (representing corporations), 285 intervals had excessive range. However, following the application of RRA algorithm, the amount of excessive intervals dropped to 32. Those 32 companies were eventually deleted from our sample. This outcome can be compared to the other two variables (Solvency and Market Ratios, each consisting of two proxy variables). Since range reduction is not possible for those two variables, all the companies represented originally by excessive ranges had to be deleted (108 companies in the case of Solvency and 144 companies in the case of Market Ratios. However, since all the variables are part of the same model, it means that if a given interval was deleted in one of the variables, all the measurements for that corporation were deleted (in other words, this corporation was deleted from the sample). In addition, the testing of the rules designed for evaluating expected earnings, required comparison of data for three consecutive years (the procedure is explained below). Therefore, if one of the intervals was either unavailable, or had excessive range in any one variable at least once in the three year sequence, that corporation was deleted from the sample. Thus we ended up with 324 corporations having a complete data for all the years under study, which is more than sufficient for statistical significance and for determining reliability.

Following the range reduction by applying RRA algorithm, we define two vectors on matrix $\widetilde{\mathbf{A}}^{\mathbf{RRA}}$:

$$\widetilde{\mathbf{A}}^{\mathbf{RRA}}_{min} = \left(\tilde{a}_1^{min}, \tilde{a}_2^{min}, \dots, \tilde{a}_{n^*}^{min}\right) \text{ and } \widetilde{\mathbf{A}}^{\mathbf{RRA}}_{max} = \left(\tilde{a}_1^{max}, \tilde{a}_2^{max}, \dots, \tilde{a}_{n^*}^{max}\right) \quad (8)$$

where $\tilde{a}_k^{min} = \min_{l=1,2,\dots,c^*}\left\{\tilde{x}_{k,l}^{RRA}\right\}$ and $\tilde{a}_k^{max} = \max_{l=1,2,\dots,c^*}\left\{\tilde{x}_{k,l}^{RRA}\right\}$ (In other words, $\tilde{a}_k^{min}$ is the minimum value for each row and $\tilde{a}_k^{max}$is the maximum value for each row).

Soft Regression using intervals data

The previous description of the explanatory variables points to a high possibility that there is a mathematical correlation among some of the variables. This means that it becomes impossible to include all of them together in the model when utilizing traditional modeling tools such as Multi-Variate Regression (MVR). Due to multicollinearity, some of the explanatory variables become insignificant not because they are not related enough to the dependent variable, but because of technical limitations of the MVR. The problem is avoided by utilizing SR modeling tool, where explanatory variables are not required to be independent of each other.

SR is a modeling tool based on soft computing concepts (such as Fuzzy Logic - (Zadeh, 1965)). The technical details of the SR method are described in (Shnaider & Yosef, 2018a; Yosef et al., 2015; Yosef & Shnaider, 2017). The following is a brief description of several of the important features of the SR that are preferable in comparison to the traditional Multi-Variate Regression (MVR) when constructing a model characterized by highly interrelated explanatory variables. These features are:

1. Soft regression does not require precise model specification to generate reliable results.
2. The significance of the explanatory variables and the relative importance of those variables among themselves are not affected by adding additional variables to the model or removing some variables from it.
3. Explanatory variables are not required to be independent of each other.
4. There are no technical issues that could cause model distortions. Wrong results are only possible if the model specification contradicts human reasoning and common sense, or if the membership functions used during the data normalization are illogical. As long as logical integrity during the model construction is maintained – the model will be reliable. This means: no unrealistic assumptions (which contradict real world conditions) are allowed.

Time Dependent Soft Regression (standard SR)

Let $Y(t+1) = (y_1(t+1), y_2(t+1), \dots, y_n(t+1))$ be the n-dimensional vector of dependent variable to be explained at time $t+1$, and let $\{X_j(t)\}_{j=1}^m$ be the corresponding n-dimensional vectors of explanatory variables at time t when $X_j(t) = (x_{j,1}(t), x_{j,2}(t), \dots, x_{j,n}(t))$. The fuzzy numerical sets of $\{X_j(t)\}_{j=1}^m$ and $Y(t+1)$ are

$$\tilde{X}_j(t) = \{(x_{j,k}(t), \tilde{x}_{j,k}(t)\}_{k=1}^n \text{ and } \tilde{Y}(t+1) = \{(y_{\mathrm{k}}(t+1), \tilde{y}_{\mathrm{k}}(t+1)\}, \text{ respectively} \quad (9)$$

where

$$\tilde{x}_{j,k}(t) = \mu_{\tilde{X}_j}\left(x_{j,k}(t)\right), \tilde{y}_{\mathrm{k}}(t+1) = \mu_{\tilde{Y}}(y_{\mathrm{k}}(t+1)) \quad (10)$$

and $\mu_{\tilde{X}_j}$, $\mu_{\tilde{Y}}$ are a membership functions of $\tilde{X}_j(t)$, $\tilde{Y}(t+1)$, respectively.

We compute the similarity between the dependent variable $Y(t+1)$ and every explanatory variable $\{X_j(t)\}_{j=1}^m$ in the following way: We define distance for direct relation between variables:

$$d_{Y,X_j}^{direct}(k,t) = |\tilde{y}_k(t+1) - \tilde{x}_{j,k}(t)| \text{ at time } t, \text{ for all } j = 1,2,\dots,m \quad (11)$$

and distance for inverse relation between variables:

$$d_{Y,X_j}^{inverse}(k,t) = |\tilde{y}_k(t+1) - (1-\tilde{x}_{j,k}(t))| \text{ at time } t, \text{ for all } j = 1,2,\dots,m \quad (12)$$

If $\sum_{k=1}^n d_{Y,X_j}^{direct}(k,t) < \sum_{k=1}^n d_{Y,X_j}^{inverse}(k,t)$ then $d_{Y,X_j}(k,t) = d_{Y,X_j}^{direct}(k,t)$ for all $k = 1,\dots,n$ and $\text{sign}_j = +1$, else $d_{Y,X_j}(k,t) = d_{Y,X_j}^{inverse}(k,t)$ for all $k = 1,\dots,n$ and $\text{sign}_j = -1$.

The similarity or closeness (denoted by S_{Y,X_j}) at time t of each explanatory variable X_j to Y is then computed as:

$$S_{Y,X_j}(t) = 1 - \frac{1}{n}\sum_{k=1}^n d_{Y,X_j}(k,t) \text{ for all } j = 1,2,\dots,m \quad (13)$$

The measure of similarity indicates the degree to which explanatory variable behaves in a similar pattern (direct or inverse) in comparison to dependent variable. Therefore, the measure of similarity $S_{Y,X_j}$is an equivalent to the traditional statistical measures of significance (t-tests or sig.). However, in addition to a significant relation (similarity of $S_{Y,X_j} \geq 0.8$), there is an option of partial significance $0.7 < S_{Y,X_j} < 0.8$, so that as $S_{Y,X_j}$is approaching closer to 0.7, it is closer to insignificance. The gradual transition from being fully significant to being fully insignificant adds additional element of stability to the modeling process when utilizing soft regression.

Once similarity measures are computed for all the explanatory variables, the next step is to calculate collective contribution of all the explanatory variables combined in explaining the behavior of dependent variable. For every observation, we select the element from one (or more) of the explanatory variables, that is the most similar (has the shortest distance) to the dependent variable, thus creating the vector of minimum distances:

$$d_{Y,X_1,\dots,X_m}^{Min}(k,t) = \min_{1\leq j\leq m} d_{Y,X_j}(k,t) \text{ at time } t, \text{ for all } k = 1,2,\dots,n. \quad (14)$$

A combined similarity at time t of all the explanatory variables to the dependent variable is

$$S_{Y,X_1,\dots,X_m}^{Comb}(t) = 1 - \frac{1}{n}\sum_{k=1}^n d_{Y,X_1,\dots,X_m}^{Min}(k,t) \quad (15)$$

$S_{Y,X_1,\dots,X_n}^{Comb}$ explains, to what degree all the explanatory variables combined – explain the behavior of the dependent variable, and in this respect, it is parallel to R^2 (in conventional regression methods). One important difference between the two measurements is that in $S_{Y,X_1,\dots,X_n}^{Comb}$ we allow for overlap of explanatory variables in their relations with the dependent variable (which is, of course, more reasonable and more in line with the "real world" behavior), and therefore explanatory variables are not required to be independent of each other.

The way to compute relative importance of the explanatory variables is to find out how much each of them contributes to the vector of minimum distances (14) (that was used to compute $S_{Y,X_1,\dots,X_n}^{Comb}$). This is done by finding the difference between the vector of minimum distances $d_{Y,X_1,\dots,X_m}^{Min}(k,t)$ (overall closeness of all the explanatory variables combined to the dependent variable) and the distance of each explanatory variable from the dependent variable (d_{Y,X_j}) (see in Shnaider & Yosef, 2018a). Therefore, relative importance of a given explanatory variable in the SR (in contrast to traditional regression methods) is not affected by correlation with other explanatory variables, and is determined solely by the contribution of that explanatory variable to explaining the behavior of the dependent variable.

We can calculate relative weight or relative importance (denoted by Relimp) of each explanatory variable in explaining the behavior of the dependent variable as follows (for more details see in Shnaider & Yosef, 2018 a):

$$\text{Relimp}_j(t) = \frac{Contrib_j(t)-0.7}{\sum_{r=1}^m (Contrib_r(t)-0.7)} \text{ at time } t, \text{ for all } j = 1,2,\dots,m, \quad (16)$$

where the contribution of each explanatory variable ($Contrib_j$) is :

$$Contrib_j(t) = 1 - \frac{1}{n}\sum_{k=1}^n |d_{Y,X_1,\dots,X_m}^{Min}(k,t) - d_{Y,X_j}(k,t)| \text{ for all } j = 1,2,\dots,m. \quad (17)$$

Soft Regression using intervals

Let $\mathbf{Y}(t+1) = \left(y_{k,l}(t+1)\right)_{n\times c_y}$ be the matrix of dependent variable to be explained at time $t+1$, and let $\{\mathbf{X}_j(t)\}_{j=1}^m$ be the corresponding matrices of explanatory variables at time t when $\mathbf{X}_j(t) = (x_{k,l}^j(t))_{n\times c_j}$ for all $j = 1,2,\dots,m$, where c_y, c_j are a numbers of columns of matrices $\mathbf{Y}(t+1)$, $\mathbf{X}_j(t)$, respectively. Based on (2), the fuzzy matrices of $\{\mathbf{X}_j(t)\}_{j=1}^m$ and $\mathbf{Y}(t+1)$ are $\tilde{\mathbf{X}}_j(t) = (\tilde{x}_{k,l}^j(t))_{n\times c_j}$ for all $j = 1,2,\dots,m$ and $\tilde{\mathbf{Y}}(t+1) =$

$\left(\tilde{y}_{k,l}(t+1)\right)_{n\times c_y}$, respectively. Following the application of RRA and based on (7) we have: $\widetilde{\mathbf{Y}}^{\mathbf{RRA}}(t+1) = \left(\tilde{y}_{k,l}^{RRA}(t+1)\right)_{n^*\times c_y^*}$ which is a dependent fuzzy matrix and $\widetilde{\mathbf{X}}_j^{\mathrm{RRA}}(t) = (\tilde{x}_{i,k}^{j,RRA}(t))_{n^*\times c_j^*}$, which are explanatory fuzzy matrices for all $j = 1,2,\dots,m$. Hence, based on (8) we represent our model in terms of vectors (fuzzy sets) as follows:

$$\widetilde{\mathbf{Y}}_{min}^{\mathbf{RRA}}(t+1), \widetilde{\mathbf{Y}}_{max}^{\mathbf{RRA}}(t+1) \text{ and } \widetilde{\mathbf{X}}_{j,min}^{\mathbf{RRA}}(t), \widetilde{\mathbf{X}}_{j,max}^{\mathbf{RRA}}(t), \text{ for all } j = 1,2,\dots,m.$$

In contrast to the large amount of regression runs that would be required by conventional regression methods to cover all possible outcomes based on the original proxy variables (24 regression runs for every year, in our present study), when using the method presented here, the amount of regression runs per year drops to 4 (and still covers all the possible outcomes)

1. Regression using only Minimum values
2. Regression using only Maximum values
3. Regression of Minimum for dependent variable vs. Maximum of explanatory variables
4. Regression of Maximum for dependent variable vs. Minimum of explanatory variables

or, using a formal notation, the above 4 regression runs can be expressed as:

1. Column of min. values of dependent variable vs. column of min. values of explanatory variables. (Set in (9) and (10): $\tilde{Y}(t+1) = \widetilde{\mathbf{Y}}_{min}^{\mathbf{RRA}}(t+1)$, $\tilde{X}_j(t) = \widetilde{\mathbf{X}}_{j,min}^{\mathbf{RRA}}(t)$).
2. Column of min. values of dependent variable vs. column of max. values of explanatory variables. (Set in (9) and (10): $\tilde{Y}(t+1) = \widetilde{\mathbf{Y}}_{min}^{\mathbf{RRA}}(t+1)$, $\tilde{X}_j(t) = \widetilde{\mathbf{X}}_{j,max}^{\mathbf{RRA}}(t)$).
3. Column of max. values of dependent variable vs. column of min. values of explanatory variables. (Set in (9) and (10): $\tilde{Y}(t+1) = \widetilde{\mathbf{Y}}_{max}^{\mathbf{RRA}}(t+1)$, $\tilde{X}_j(t) = \widetilde{\mathbf{X}}_{j,min}^{\mathbf{RRA}}(t)$).
4. Column of max. values of dependent variable vs. column of max. values of explanatory variables. (Set in (9) and (10): $\tilde{Y}(t+1) = \widetilde{\mathbf{Y}}_{max}^{\mathbf{RRA}}(t+1)$, $\tilde{X}_j(t) = \widetilde{\mathbf{X}}_{j,max}^{\mathbf{RRA}}(t)$).

The four regression runs generate four results of: similarity (S_{Y,X_j}), combined similarity ($S_{Y,X_1,\dots,X_m}^{Comb}$) and relative importance (Relimp_j) which are converted to ranges between the lowest result and the highest results (see Table 2).

Note:

1. It does not matter how many explanatory variables are expressed in terms of intervals, the method will still require only four regression runs for a specific year.
2. Since in this study the dependent variable is not an interval, but a single numerical vector, the amount of required regression runs per year drops to 2.

The decision support rules based on the model results

By definition, explanatory variables are supposed to explain as much as possible the behavior of the dependent variable: the variable to be explained. The term "dependent variable" means that its behavior is dependent on the behavior of explanatory variables (even if based on correlation and not on causality). Here we take this statement slightly further by stating that the combination of all the explanatory variables represents in fact the potential for the behavior of the dependent variable. The potential determines approximate limits of possible values of the dependent variable Those limits can be exceeded occasionally, but such discrepancies are unsustainable.

The important question is: how to combine the explanatory variables to compute the potential? In this study we utilize weighted linear combination of explanatory variables. Here we apply the previously computed relative importance of explanatory variables (Relimp_j), which enables to differentiate among the different explanatory variables and assign appropriate weight. We denote such weighted linear combination of the explanatory variables WLC_t.

$$WLC_t = \sum_{j=1}^{m} a_j \tilde{x}_{j,k}^{sign}(t) \tag{18}$$

where

$$\tilde{x}_{j,k}^{sign}(t) = \begin{cases} \tilde{x}_{j,k}(t) & ,\mathrm{sign}_j = +1 \\ 1-\tilde{x}_{j,k}(t) & ,\mathrm{sign}_j = -1 \end{cases}$$

In the case of the specific financial study discussed in this study, the linear combination of explanatory variables weighted by their relative importance, can be interpreted as an indicator of weighted financial capabilities (an indicator of financial potential) of various corporations.

The two regression runs (per year) generate two results of WLC_t which are converted to ranges between the low result and the high result (see Table 2).

Since the WLC_t reflects firm's financial potential, it becomes possible to explore the relation between WLC_t and the dependent variable (Operating Income Margin) for next year $t+1$. If such a relation is significant, then it becomes possible to set some rules, that can be very helpful to identify group of next year winners, next year losers, etc. Obviously, being able to identify next year's candidates for improved performance, poor performance, uncertain cases, etc., means highly effective decision support tool. In order to implement such tool, it is necessary to set some rules and perform tests to determine to what degree such rules are reliable.

We postulate the following conditions:

For improved future performance:

1. If $WLC_t^{min} \geq \widetilde{\mathbf{Y}}_{max}^{\mathbf{RRA}}(t)$ it means that during period t, the smallest WLC_t^{min} , which represents the financial capability, or potential of the firm, is greater than its actual performance during the year t. In other words, the actual performance of the firm in year t does not reflect its capabilities and its potential is under-utilized. Hence better performance in the future is likely.
2. If $WLC_{t+1}^{min} \geq WLC_t^{max}$ it means that the firm definitely improved its financial potential from year t to year $t+1$ by all measurements (because the minimum of year $t+1$ is greater than the maximum of year t).

When we combine both rules, 1 and 2 into a single rule, it means that not only the performance of the firm in year t is lower than its full potential (does not reflect its full potential), but that potential is further increasing in the next year $(t+1)$. Under such circumstances we will expect the following outcomes:

a. The firms which were already in the category of "Winners" in the years t and $t+1$ are expected to remain in the category of "Winners" in $t+2$
b. The firms which do not belong to category of "Winners" are expected to improve their performance in the year $t+2$ in comparison to the year $t+1$.

For worse future performance, the reasoning is similar, but in the opposite direction:

3. If $WLC_t^{max} \leq \widetilde{\mathbf{Y}}_{min}^{\mathbf{RRA}}(t)$ it means that during period t, the largest WLC_t^{max}, which represents the financial capability, or potential of the firm, is smaller than its actual performance during the year t. In other words, the actual financial performance of the firm in year t exceeds its financial potential. Hence deteriorating performance in the future is expected.
4. If $WLC_t^{min} \geq WLC_{t+1}^{max}$ it means that the firm's capabilities (financial potential) definitely deteriorated from year t to year $t+1$ by all measurements (because the maximum of year $t+1$ is smaller than the minimum of the year t).

When we combine both rules, 3 and 4 into a single rule, it means that not only the performance of the firm in year t is higher in comparison to what is reflected by its potential capabilities, but those capabilities are further decreasing in the next year $(t+1)$. Under such circumstances we will expect the following outcomes:

a. The firms which were already in the category of "Losers" in the years t and $t+1$ are expected to remain in the category of "Losers" in $t+2$
b. The firms which do not belong to category of "Losers" are expected to experience deteriorating performance in the year $t+2$ in comparison to the year $t+1$.

In other words, for improved future performance:

If $WLC_t^{min} \geq \widetilde{\mathbf{Y}}_{max}^{\mathbf{RRA}}(t)$ and $WLC_{t+1}^{min} \geq WLC_t^{max}$ then expected outcome is: $\widetilde{\mathbf{Y}}_{min}^{\mathbf{RRA}}(t+2) \geq \widetilde{\mathbf{Y}}_{max}^{\mathbf{RRA}}(t+1)$

And for worse future performance:

If $WLC_t^{max} \leq \widetilde{\mathbf{Y}}_{min}^{\mathbf{RRA}}(t)$ and $WLC_t^{min} \geq WLC_{t+1}^{max}$ then expected outcome is: $\widetilde{\mathbf{Y}}_{min}^{\mathbf{RRA}}(t+1) \geq \widetilde{\mathbf{Y}}_{max}^{\mathbf{RRA}}(t+2)$.

In our study, there is only one proxy (A20- Operating Income Margin at time t, denoted by OIM_t) for the dependent variable, thus $\widetilde{\mathbf{Y}}_{min}^{\mathbf{RRA}}(t) = \widetilde{\mathbf{Y}}_{max}^{\mathbf{RRA}}(t) = OIM_t$ and

Improved future performance: If $WLC_t^{min} \geq OIM_t$ and $WLC_{t+1}^{min} \geq WLC_t^{max}$ then $OIM_{t+2} \geq OIM_{t+1}$

Worse future performance: If $WLC_t^{max} \leq OIM_t$ and $WLC_t^{min} \geq WLC_{t+1}^{max}$ then $OIM_{t+1} \geq OIM_{t+2}$

The example of next year performance by several companies appears in Table 2

Table 2: Next year performance of several companies

	Company	OIM_{2013}	WLC_{2013}	WLC_{2014}	OIM_{2014}	OIM_{2015}
Improved future performance	ARAY	0.082	[0.209,0.230]	[0.512,0.560]	0.703	0.713
	IDSY	0.020	[0.329,0.376]	[0.414,0.479]	0.163	0.254
Worse future performance	CEMI	0.899	[0.829,0.862]	[0.446,0.498]	0.686	0.488
	KLIC	1	[0.730,0.746]	[0.710,0.729]	1	0.994

OIM_t: Operating Income Margin at time t (after normalization)

In Table 2, there is an example consisting of four corporations. Columns 3, 4 and 5 display the normalized and computed information regarding the companies (for 2013 and 2014), which is utilized as input into the rules presented above. The last two columns provide actual behavior of Operating Income Margin between 2014 and 2015.

Results

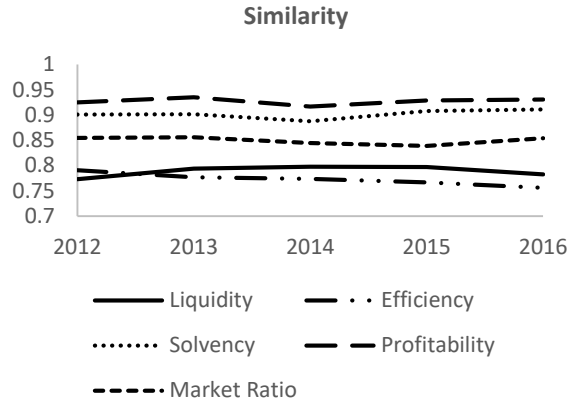
Table 3 displays the results of modeling, based on Soft Regression, and including measures of Similarity (equation 13), S-comb (equation 15) and Relimp (equation 16). Similarity measures to what extent the behavior of any explanatory variable resembles that of the dependent variable. Based on Table 3, explanatory variables Liquidity and Efficiency are both not fully significant, but are fairly close to the border line of being fully significant. The three other variables are highly significant. The differences in the degrees of significance (Similarity) are reflected by a differences in the relative importance of the corresponding variables (Relimp). S^{Comb}, which displays the overall ability of the model to explain the behavior of the dependent variable, is consistently above 0.95. This is a very high score, considering that the model is a cross-section model.

Table 3:The Summary of Modeling Results

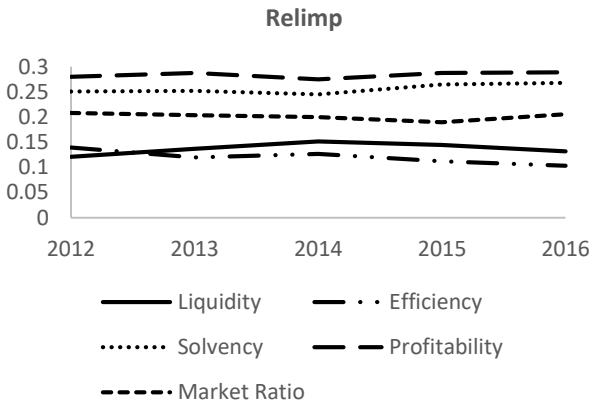
		2012-2013	2013-2014	2014-2015	2015-2016	2016-2017
Similarity (S_{Y,X_j})	Liquidity	0.773	0.794	0.798	0.797	0.783
	Efficiency	0.791	0.777	0.774	0.767	0.756
	Solvency	[0.891,0.902]	[0.896,0.907]	[0.882,0.893]	[0.903,0.913]	[0.905,0.917]
	Profitability	[0.923,0.926]	[0.933,0.936]	[0.915,0.919]	[0.927,0.931]	[0.929,0.933]
	Market ratios	[0.850,0.861]	[0.851,0.861]	[0.838,0.850]	[0.833,0.846]	[0.847,0.859]
Relimp	Liquidity	[0.119,0.122]	[0.136,0.138]	[0.150,0.154]	[0.143,0.147]	[0.130,0.134]
	Efficiency	[0.138,0.141]	[0.119,0.121]	[0.125,0.128]	[0.111,0.113]	[0.102,0.105]
	Solvency	[0.248,0.253]	[0.249,0.254]	[0.242,0.247]	[0.262,0.265]	[0.265,0.270]
	Profitability	[0.278,0.283]	[0.284,0.290]	[0.273,0.279]	[0.285,0.290]	[0.286,0.291]
	Market ratios	[0.205,0.211]	[0.201,0.206]	[0.195,0.203]	[0.186,0.194]	[0.202,0.209]
S^{Comb}		[0.958,0.959]	[0.962,0.964]	[0.951,0.954]	[0.962,0.963]	[0.959,0.960]

Note: In each column, the explanatory variables are for year t and the dependent variable for year $t+1$.

Graph 1



Graph 2



The consistency of the results over the years included in this study can be easily observed visually in Graph1 and Graph 2.

High values of S^{Comb}, as well as consistency of the results over the years are both important parameters in judging model's reliability. Additional factors adding to model's reliability: (a) Utilizing all the available data (as many proxy variables as possible) for each financial factor included in the model and (b) Utilizing robust modeling tool, that does not require restrictive/unrealistic conditions to generate reliable results.

Table 4 presents the results of the tests conducted to evaluate the effectiveness of the rules defined and explained in the previous section. The conditions of the rules are computed for each firm separately based on years t and $t+1$. This expected performance are compared to the actual OIM for year $t+2$. In other words, for each test we need the sequence of three years to either confirm or reject whether the rule successfully anticipated the outcome in year $t+2$.

For each one of the three right-hand columns (in Table 4), there is a ratio of successful outcomes vs total amount of relevant cases. Below the ratio, the success rate is computed as a percentage. The lowest success rate is 87.9%, which is still very high considering the volatility and uncertainty characterizing the stock market data. The ability to anticipate correctly, even in very general and fuzzy terms, above 87.9% of the cases (selected by the predefined rules) on a consistent basis indicates the effectiveness of the decision support tool.

Table 4:The Success Rate of the Rules

	2013-2015	2014-2016	2015-2017
If $WLC_t^{min} \geq OIM_t$ and $WLC_{t+1}^{min} \geq WLC_t^{max}$ then $OIM_{t+2} \geq OIM_{t+1}$ (Improved future performance)	29/33 (87.9%)	21/21 (100%)	23/24 (95.8%)
If $WLC_t^{max} \leq OIM_t$ and $WLC_t^{min} \geq WLC_{t+1}^{max}$ then $OIM_{t+1} \geq OIM_{t+2}$ (Worse future performance)	53/56 (94.6%)	38/43 (88.4%)	47/51 (92.2%)

OIM_t: Operating Income Margin at time t (after normalization)

Summary and Conclusions

This study consists of several unique applications of Soft Computing methods in constructing a financial decision support system. The study can be summarized as follows:

a. A general cross-section model is constructed, based on all the companies for which a complete set of data for all the relevant periods are available.
b. Results of the model display high scores of Similarity (S_{Y,X_j}) and Combined Similarity (S^{Comb}) measures, as well as substantial stability of those results over the years, thus implying high degree of reliability.
c. The use of data is inclusive, which means that if there are several proxy variables that supposedly measure the same factor, then all of the proxy variables are included in the modeling process by applying intervals of values.

d. Soft computing tool utilized for modeling was Soft Regression (SR). SR is based on Fuzzy Logic and has some features that are very critical for the successful modeling of the type performed in this study: (i) Reliability of computing relative weights of the explanatory variables, and (ii) does not require explanatory variables to be independent of each other, and hence allows to incorporate in the same model explanatory variables that are highly correlated mathematically among themselves. Inclusion of all the logically relevant explanatory variables in the model helps to maintain logical integrity of the model.
e. Using linear combination of the explanatory variables weighted by their corresponding relative importance allows to construct rules that can be helpful indicators of future behavior (in broad terms) of the Corporate Operating Income Margin. Those rules are applied for evaluating each one of the corporations for which complete data are available. The rules identify those companies where the logical possibility of a given type of outcome for the next year is very high.
f. The success rate of the rules was tested and for all the years under study exceeded 87.9%. In other words, the output of the system is broad and fuzzy, but mostly correct.

The advantage of Soft Computing in the context of the above study can be stated as follows: It is more important to have imprecise output which is broadly correct, rather than precise output which is incorrect. Such broadly correct, but mostly reliable output is what decision makers need for successful investment decisions in the stock market.

References

1. Atiya, A. F. (2001). Bankruptcy Prediction for Credit Risk Using Neural Networks: A Survey and New Results. IEEE Transactions on Neural Networks, 12(4), 929–935. https://doi.org/10.1109/72.935101
2. Ball, Ray; Brown, P. (1968). An Empirical Evaluation of Accounting Income Numbers. Journal of Accounting Research, 6(2), 159–178. https://doi.org/10.2307/2490232
3. Ball, R. (1992). The Earnings-Price Anomaly. Journal of Accounting and Economics, 15(2–3), 319–345. https://doi.org/10.1016/0165-4101(92)90023-U
4. Baranes, A., & Palas, R. (2017a). Multivariate Imputation of XBRL Data for Financial Statement Analysis. International Journal of Information Research and Review, 04(7).
5. Baranes, A., & Palas, R. (2017b). The Prediction of Earnings Movements Using Accounting Data: Using XBRL. International Journal of Accounting Research, 04(2), 1–7. https://doi.org/10.4172/2472-114X.1000143
6. Baranes, A., & Palas, R. (2019). Earning movement prediction using machine learning-support vector machines (SVM). Journal of Management Information and Decision Sciences, 22(2), 36–53.
7. Beaver, W., Lambert, R., & Morse, D. (1980). The Information Content of Security Prices. Journal of Accounting and Economics, 2(1), 3–28. https://doi.org/10.1016/0165-4101(80)90013-0
8. Bernard, V., Thomas, J., & Wahlen, J. (1997). Accounting-Based Stock Price Anomalies: Separating Market Inefficiencies from Risk. Contemporary Accounting Research, 14(2), 89–136. https://doi.org/10.1111/j.1911-3846.1997.tb00529.x
9. Bird, R., Gerlach, R., & Hall, A. D. (2001). The Prediction of Earnings Movements Using Accounting Data: An Update and Extension of Ou and Penman. Journal of Asset Management, 2(2), 180–195. https://doi.org/10.1057/palgrave.jam.2240044
10. Chandwani, D., & Saluja, M. S. (2014). Stock Direction Forecasting Techniques : An Empirical Study Combining Machine Learning System with Market Indicators in the Indian Context. International Journal of Computer Applications, 92(11), 8–17.
11. Chen, Kung, H., & ShimerdaThomas, A. (1981). An Empirical Analysis of Useful Financial Ratios. Financial Management, 10(1), 51–61. https://doi.org/10.2307/3665113
12. D'Souza, J. M., Ramesh, K., & Shen, M. (2010). The Interdependence Between Institutional Ownership and Information Dissemination by Data Aggregators. Accounting Review. https://doi.org/10.2308/accr.2010.85.1.159
13. Harrison, W. T., Horngern, C. T., Thomas, W. C., & Suwardy, T. (2011). Financial Accounting - International Financial Reporting Standards (Eighth Edi). Singapore: Pearson Education South Asia.
14. Holthausen, R. W., & Larcker, D. F. (1992). The Prediction of Stock Returns Using Financial Statement Information. Journal of Accounting and Economics, 15(2–3), 373–411. https://doi.org/10.1016/0165-4101(92)90025-W

15. Kandel, A., Last, M., & Bunke, H. (2001). Data Mining and Computational Intelligence. Physica-Verlag Publishing.
16. Kinney, M. R., & Swanson, E. P. (1993). The Accuracy and Adequacy of Tax Data in COMPUSTAT. Journal of the American Taxation Association, 15(1), 121.
17. Lev, B., & Gu, F. (2016). The End of Accounting and the Path Forward for Investors and Managers. John Wiley & Sons.
18. Miguel, J. G. S. (1977). The Reliability of R&D Data in COMPUSTAT and 10-K Reports. The Accounting Review, 52(3), 638–641.
19. Ou, J. A., & Penman, S. H. (1989). Financial Statement Analysis and the Prediction of Stock Returns. Journal of Accounting and Economics, 11(4), 295–329. https://doi.org/10.1016/0165-4101(89)90017-7
20. Ou, J. A, & Penman, S. H. (1989). Accounting Measurement, Price-Earnings Ratio, and the Information Content of Security Prices. Journal of Accounting Research, 27(3), 111–144. https://doi.org/10.2307/2491068
21. Pinches, G. E., Eubank, A. A., Mingo, K. A., & Caruthers, J. K. (1975). The Hierarchical Classification of Financiul Ratios. Journal of Business Research, 3(4), 295–310.
22. Rosenberg, B., & Houglet, M. (1974). Error Rates IN CRSP and COMPUSTAT Data Bases and Their Implications. The Journal of Finance, 29(4), 1303–1310. https://doi.org/10.1111/j.1540-6261.1974.tb03107.x
23. Setiono, B., & Strong, N. (1998). Predicting Stock Returns Using Financial Statement Information. Journal of Business Finance and Accounting, 25(5–6), 631–657. https://doi.org/10.1111/1468-5957.t01-1-00205
24. Shin, K. S., Lee, T. S., & Kim, H. J. (2005). An Application of Support Vector Machines in Bankruptcy Prediction Model. Expert Systems with Applications, 28(1), 127–135. https://doi.org/10.1016/j.eswa.2004.08.009
25. Shnaider, E., & Yosef, A. (2018a). Relative Importance of Explanatory Variable: Traditional Method vs Soft Regression. International Journal of Intelligent System, 33(6), 1180–1196.
26. Shnaider, E., & Yosef, A. (2018b). Utilizing Intervals of Values in modeling due to Diversity of Measurements. Fuzzy Economic Review, International Association for Fuzzy-set Management and Economy (SIGEF). vol. 23, num. 2, pages 3-26.
27. Stober, T. L. (1992). Summary Financial Statement Measures and Analysts' Forecasts of Earnings. Journal of Accounting and Economics, 15(2–3), 347–372. https://doi.org/10.1016/0165-4101(92)90024-V
28. Tallapally, P., Luehlfing, M. S., & Motha, M. (2011). The Partnership Of EDGAR Online And XBRL - Should Compustat Care? The Review of Business Information Systems, 15, 39–46. Retrieved from http://search.proquest.com/docview/900720360?accountid=11262
29. Tsai, C. F. (2009). Feature Selection in Bankruptcy Prediction. Knowledge-Based Systems, 22(2), 120–127. https://doi.org/10.1016/j.knosys.2008.08.002
30. Tsai, C. F., & Hsiao, Y. C. (2010). Combining Multiple Feature Selection Methods for Stock Prediction: Union, Intersection, and Multi-Intersection Approaches. Decision Support Systems, 50(1), 258–269. https://doi.org/10.1016/j.dss.2010.08.028
31. Yang, D. C., Vasarhelyi, M. a., & Liu, C. (2003). A Note on the Using of Accounting Databases. Industrial Management & Data Systems, 103(3), 204–210. https://doi.org/10.1108/02635570310465689
32. Yosef, A., Haruvy, N., & Shnaider, E. (2015). Soft Regression vs Linear Regression. Pioneer Journal of Theoretical and Applied Statistics, 10(1–2), 31–46.
33. Yosef, A., & Shnaider, E. (2017). On Measuring the Relative Importance of Explanatory Variables in a Soft Regression Method. Advances and Applications in Statistics, 50(3), 201 – 228.
34. Zadeh, L. . (1965). Fuzzy Sets. Information and Control, 8(3), 338–353.

Appendix 1: All financial ratios

No.	Ratio Classification	Variables [a]
1	Liquidity	Current Ratio
2		Quick Ratio
3		Sales to total working capital
4		Working capital to total assets
5		ΔCurrent Ratio
6		ΔQuick Ratio
7		ΔWorking capital
8		ΔSales to total working capital
9		ΔWorking capital to total assets
10	Efficiency	Account Receivable Turnover
11		Inventory Turnover
12		Days sales in Accounting Recv.

13		Inventory to total assets
14		Sales to total cash
15		Sales to total Inventory
16		Δinventory
17		ΔInventory Turnover
18		ΔDays sales in Accounting Recv.
19		ΔInventory to total assets
20		ΔSales to total Inventory
21	Solvency	Total Debt To Equity
22		Long-Term Debt/Equity
23		Equity/Fixed assets
24		Times Interest Earned
25		Cash From Operations (CFO) to Total Debt
26		ΔTotal Assets
27		ΔTotal Long-Term Debt
28		ΔTotal Debt To Equity
29		ΔCapital Expenditures/total assets
30		ΔLong-Term Debt/Equity
31		ΔEquity/Fixed assets
32		ΔTimes Interest Earned
33		LΔCapital Expenditures/total assets
34	Profitability	ROA
35		ROE
36		Gross Profit Margin
37		Depreciation over Plant
38		Sales/Total Assest
39		Pre taxes income/Sales
40		Net Profit Margin
41		Sales to Fixed assets
42		Operating Income to Total assets
43		EBITDA Margin Ratio
44		Net Income over OCF
45		ΔDepreciation (&Amortization), IS
46		ΔResearch & Development Expense
47		ΔTotal Revenue
48		ΔROE
49		ΔGross Profit Margin
50		ΔDepreciation over Plant
51		ΔSales/Total Assest
52		ΔPre taxes income/Sales
53		ΔNet Profit Margin
54		ΔResearch & Development Expense to Sales
55		ΔOperating Income to Total assets
56		ΔEBITDA Margin Ratio
57	Market Ratios	Payment Of Dividends as % of OCF
58		ΔDividends per share
[a] Δ indicates changes. In calculating % Δ, observations with zero denominators are excluded and absolute values are used in all denominators.		