Transactions on Machine
Learning and Data Mining
Vol. 12, No. 2 (2019) 55 - 74

P-ISSN: 1865-6781, E-ISSN 2509-9337
ISBN: 978-3-942952-72-9

ibai Publishing
www.ibai-publishing.org

Evaluating Economic Performance with Soft Regression

Moti Schneider[1], Arthur Yosef[2], Eli Shnaider[3]

[1] Netanya Academic College, Netanya, Israel, profmoti@gmail.com
[2]Tel Aviv-Yaffo Academic College, Tel Aviv, Israel, yusupoa@yahoo.com
[3]Peres Academic Center, Rehovot Israel, eli-sh@012.net.il

Abstract. This study demonstrates effective data mining tool under severe limitations of data availability. We present a soft computing method for evaluating economic performance. To avoid computational explosion, we utilize intervals. This will reduce the number attributes in the dataset. Utilizing intervals allows us to overcome difficult modeling problems such as large quantity of missing data, substantial outliers, etc. Finally, case study of evaluating economic performance of the Soviet led East European bloc is presented. In spite of highly unreliable and inaccurate data provided by the officials of the bloc, the method presented here allows to reach solid and reliable conclusions.

Keywords: Data Mining, Soft Computing, Cross-national model, Soft Regression, fuzzy logic.

1 Introduction

In this study we present a method to evaluate economic performance of either individual countries or groups of countries, based on data mining/soft computing tools. The method involves modeling of the general factors facilitating economic performance. The factors facilitating economic performance essentially represent the constraints that either facilitate or limit long-term economic growth.

As a case study, we evaluated economic performance of the Soviet-led East European bloc, which is a particularly challenging task due to a biased data published by the official sources of these countries. The study is based mostly on cross-national data originated from the World Bank databases and hard copy publications and covers the period from 1960 to 1992. More specifically, we constructed a cross-national model for the years 1960, 1970, 1978, 1985 and 1992. We presented results for the year 1992 – to illustrate the situation following the collapse of the Eastern Bloc, because some conclusions can be reached in retrospect. Thus, we followed the performance and illustrated the constraints of the Soviet led alliance till its end.

In order to assure reliability and robustness of the results, the idea was to build a general model based on the data from over 120 countries and then to apply the results specifically to the countries of the East European Bloc. The generality of the model (rather than building a specific model for the East European countries) is one of the points of strength in this study.

The modeling method, “Soft Regression” (SR) [2], is a Soft Computing tool, based on fuzzy logic [7,8,11]. Utilizing SR rather than traditional econometric tools makes it possible to overcome some technical difficulties associated with the traditional modeling tools and allows us to build more reliable and robust model, as will be explained below. There is a large amount of modeling tools, including regression methods based on fuzzy logic [13,14,15]. However, we decided to use Soft Regression, because for our purpose it is most effective and most convenient tool. In particular, the reliability of computing relative importance of explanatory variables (Relimp) and the ability to avoid distortions due to highly correlated explanatory variables (see [8]) point to Soft Regression as an appropriate choice.

The methodology introduced in this study can be applied to any country or group of countries to evaluate where they stand in comparison to the leading performers at any given point of time. Thus, the method presented in this study can generate important information necessary for designing effective long-term policies to "contain" the challengers and it can generate important information for the lagging countries to identify their basic weaknesses. Hence, the method can be a useful tool for economic policy makers as well as for foreign policy strategists.

The method presented here is general enough and can be also be applied for comparative evaluation of corporations, or of various components (sub-divisions) of large organizations, etc..

In Section 2 we describe different soft regression methods. Then, in Section 3 we explain a case study based on economic and socioeconomic data from the Soviet Union. We describe the factors facilitating economic performance, the variables, how to normalize the data and how data preparation is carried out. In Section 4 we present the results. Conclusions are given in Section 5.

2 Soft Regression

SR is a modeling tool based on soft computing concepts (such as Fuzzy Logic – Zadeh (1965 [12])). The technical details of the SR method are described in [8], [11], [10]. Previous works leading to the development of Soft Regression are: [5], [2], [6].

We will briefly describe 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.
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.

2.1 Standard Soft regression

Recall the definition of the fuzzy set: if X is a collection of objects denoted generically by x, then a fuzzy set $\widetilde{\mathrm{A}}$ in X is a set of ordered pairs:

$$\widetilde{\mathrm{A}} = \{(x, \tilde{x}): x \in X\} \text{ where } \tilde{x} = \mu_{\widetilde{\mathrm{A}}}(x) \quad (1)$$

$\mu_{\widetilde{\mathrm{A}}}$ is called the membership function (for computing grade of membership of $\boldsymbol{x}$ in $\widetilde{\mathrm{A}}$)that maps X to [0,1]. Let $Y = (y_1, y_2, \ldots, y_n)$ be the n-dimensional vector of dependent variable to be explained, and let $\{X_j\}_{j=1}^{m}$ be the corresponding n-dimensional vectors of explanatory variables when $X_j = (x_{j,1}, x_{j,2}, \ldots, x_{j,n})$. Based on (1), the fuzzy numerical sets of $\{X_j\}_{j=1}^{m}$ and Y are

$$\tilde{X}_j = \{(x_{j,k}, \tilde{x}_{j,k}\}_{k=1}^{n}, \text{ for all } j = 1,2, \ldots, m \text{ and } \tilde{Y} = \{(y_{\mathrm{k}}, \tilde{y}_{\mathrm{k}}\}, \text{ respectively} \quad (2)$$

where

$$\tilde{x}_{j,k} = \mu_{\tilde{X}_j}(x_{j,k}), \tilde{y}_{\mathrm{k}} = \mu_{\tilde{Y}}(y_{\mathrm{k}}) \text{ and } \mu_{\tilde{X}_j}, \mu_{\tilde{Y}} \text{ are a membership functions of } \tilde{X}_j, \tilde{Y}, \text{ respectively.} \quad (3)$$

We compute the similarity between the dependent variable Y and every explanatory variable $\{X_j\}_{j=1}^{m}$ in the following way:

We define distance for direct relation between variables:

$$d_{Y,X_j}^{direct}(k) = |\tilde{y}_{\mathrm{k}} - \tilde{x}_{j,k}| \text{ for all } j = 1,2, \ldots, m \quad (4)$$

and distance for inverse relation between variables:

$$d_{Y,X_j}^{inverse}(k) = |\tilde{y}_{\mathrm{k}} - (1 - \tilde{x}_{j,k})| \text{ for all } j = 1,2, \ldots, m \quad (5)$$

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

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

$$S_{Y,X_j} = 1 - \frac{1}{n}\sum_{k=1}^{n} d_{Y,X_j}(k) \text{ for all } j = 1,2, \ldots, m \quad (6)$$

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:

$$d_{Y,X_1,\ldots,X_m}^{Min}(k) = \min_{1 \leq j \leq m} d_{Y,X_j}(k) \text{ for all } k = 1,2, \ldots, n. \quad (7)$$

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

$$S^{Comb}_{Y,X_1,\dots,X_m} = 1 - \frac{1}{n}\sum_{k=1}^{n} d^{Min}_{Y,X_1,\dots,X_m}(k) \quad (8)$$

$S^{Comb}_{Y,X_1,\dots,X_n}$ 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 $S^{Comb}_{Y,X_1,\dots,X_n}$ 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 (7) (that was used to compute $S^{Comb}_{Y,X_1,\dots,X_n}$). This is done by finding the difference between the vector of minimum distances $d^{Min}_{Y,X_1,\dots,X_m}(i)$ (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 [4]). Therefore, relative importance 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 a given 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 based on the following principles (for more details [8]):

$$\text{Relimp}_j = \frac{Contrib_j - 0.7}{\sum_{r=1}^{m}(Contrib_r - 0.7)} \text{ for all } j = 1,2,\dots,m, \quad (9)$$

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

$$Contrib_j = 1 - \frac{1}{n}\sum_{k=1}^{n} |d^{Min}_{Y,X_1,\dots,X_m}(k) - d_{Y,X_j}(k)| \text{ for all } j = 1,2,\dots,m. \quad (10)$$

2.2 Soft Regression using Intervals (see [9])

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 supposedly represent the same thing, we can construct a matrix, such that each numerical vector is a column in that matrix. When the matrix of k columns (numerical vectors) is converted into the matrix of intervals, it will become a matrix of two columns: column of minimum values and column of maximum values.

There are numerous studies regarding the application of intervals in modeling and dealing with the issue of missing data [15, 16]. However, in this study we utilize a method presented in [9], where the effectiveness of utilizing data in terms of intervals for modeling purposes is demonstrated for all the aspects important for this article.

There is a very important issue that must be addressed when constructing intervals of values: it is critical to make sure that before we construct the intervals, all variables 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), 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.

Another important issue to consider when constructing intervals is the potential presence of outliers and their implications. The outliers that are expected to appear in various data series can substantially widen the intervals to a degree that is detrimental for successful modeling. The cut-off points applied in membership functions by their nature tend to alleviate, at least to some extend the problem of outliers. In other words, when different measurements are full members of the fuzzy set, they all are assigned the value of 1, no matter how much their original values differ. The same holds for measurements that are definitely not members of the fuzzy set – all of them are assigned the value of 0, no matter how much their original values differ.

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. This way, for every row (in our case – for each country), we construct intervals consisting of grades of membership.

We utilized the Range Reduction Algorithm (RRA), which is explained in detail in [9]. RRA is applied to reduce the range of intervals by deleting outliers. RRA also identifies cases where interval reduction is not working, and the length of the interval is such, as to seriously question the reliability of the data. In such cases the data for that specific country are deleted.

2.3 Range Reduction Algorithm (RRA) (see [9])

The algorithm of range reduction consists of the following main components:

1. Preparation Stage
2. Identifying and eliminating outlying identical (or almost identical) vectors.
3. Reducing range: Deleting outlying elements
4. Additional reduction of the range and deletion of over-extended intervals (optional)

The work has been carried out as follow:

a) **Preparation Stage:**
Let's assume that we have c numerical vectors, each consisting of n elements (In other words, we have a matrix $\mathbf{A} = (x_{k,l})_{n\times c}$ 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.,
A fuzzy matrix of $\mathbf{A}$ is a matrix:
$$\widetilde{\mathbf{A}} = (\tilde{x}_{k,l})_{n\times c} \quad (11)$$
where $\tilde{x}_{k,l} = \mu_l(x_{k,l})$ for all $k = 1,2,\dots,n$ and μ_l is a membership function for all $l = 1,2,\dots,c$. Sort each row of the matrix from the lowest value on the left side to the highest value on the right side.
Note: Following the preparation stage, 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.

b) **Identifying and eliminating outlying identical (or almost identical) vectors:**
The idea behind this part of the algorithm is to correct possible distortion, when due to unique methodology, conversion methods, etc., some vectors become outliers for all or most of their elements. If only one such numerical vector appears in our data, the interval reduction procedure presented in stage 3 will handle it. However, if two or more vectors like that appear, and they are identical or almost identical, then the method presented in stage 3 will not perform effectively. This problem might arise when collecting data series that are having different names, but are essentially the same mathematically. They might differ in scale, which makes it difficult to detect the similarity among them. However, once these data series are normalized, they might become almost identical. In other words, the problem arises not when such situation is encountered in just several rows, but when we are talking about identical vectors for almost all their rows. Thus our objective at this stage is to locate possible outlying pairs or groups of vectors that are identical or almost identical and delete the redundant elements. We should note that having identical or almost identical vectors does not constitute a problem as long as they are confined mostly to the internal portion of the interval. However, if they are located on the edges, they will imperil our ability to reduce the interval.
Another important point to consider: when deleting elements from the matrix, we must keep in mind that some rows (in our case: data for some countries) might consist of very few measurements. No element should be deleted from the matrix, if in that row, there are only four measurements or less. The reason for that is: our objective is to attain better representation of the central tendency, but we want to achieve it without possible loss of information. When amount of elements in a given interval is large, then deleting several outlying elements only brings us closer to the core of the "central tendency". However, when the amount of elements is small (four or less), then deleting a single element can potentially lead to a loss of important information and distort our view of central tendency. In this case it is preferable to keep the whole original interval.

c) **Reducing range; Deleting outlying elements:**
The interval reduction rules are applied for each row (interval) separately, depending on the specific characteristics of that interval. If there are four elements or less in a specific row, leave the raw as is. If there are five elements in that row, one outlying element can be deleted. If there are 6 to 10 elements in the row, two outlying elements can be deleted. For any additional 5 elements in the row, one additional outlying element can be deleted, etc. For example: in the interval of 11-15 elements we delete 3 elements, in the interval 16-20 elements we delete 4, etc. Thus at this stage we determine the

amount of elements to be deleted in a given row. The deleted elements can be located either on the left side of that row, or on the right side or both. Obviously, no element located in the middle of row can be deleted. The idea is to select elements for deletion so as to achieve maximum range reduction of a given interval.

d) **Additional reduction of the interval (optional):**
Following the range reduction process described above, if the new range 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. This portion is optional and involves individual reasoning by a modeling professional, and could differ based on circumstances and constrains. In our case study we decided at this stage to delete rows where the interval exceeded 0.30.
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. In our case study, for each of the variables, we deleted only few cases out of over 125 rows (see Table 1) – which did not affect final results.

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

$$\widetilde{\mathbf{A}}^{\mathbf{RRA}} = \left(\tilde{x}_{k,l}^{RRA}\right)_{n^* \times c^*} \tag{12}$$

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

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

$$\widetilde{\mathbf{A}}_{min}^{\mathbf{RRA}} = \left(\tilde{a}_1^{min}, \tilde{a}_2^{min}, \dots, \tilde{a}_{n^*}^{min}\right) \text{ and } \widetilde{\mathbf{A}}_{max}^{\mathbf{RRA}} = \left(\tilde{a}_1^{max}, \tilde{a}_2^{max}, \dots, \tilde{a}_{n^*}^{max}\right) \tag{13}$$

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).

Let $\mathbf{Y} = \left(y_{k,l}\right)_{n \times c_y}$ be the matrix of dependent variable to be explained, and let $\{\mathbf{X}_j\}_{j=1}^m$ be the corresponding matrices of explanatory variables when $\mathbf{X}_j = (x_{k,l}^j)_{n \times c_j}$ for all $j = 1,2,\dots,m$, where c_y, c_j are a numbers of columns of matrices $\mathbf{Y}$, $\mathbf{X}_j$, respectively. Based on (11), the fuzzy matrices of $\{\mathbf{X}_j\}_{j=1}^m$ and $\mathbf{Y}$ are

$\widetilde{\mathbf{X}}_j = (\tilde{x}_{k,l}^{\,j})_{n \times c_j}$ for all $j = 1,2,\dots,m$ and $\widetilde{\mathbf{Y}} = \left(\tilde{y}_{k,l}\right)_{n \times c_y}$, respectively.

After applying RRA and based on (12) we have: $\widetilde{\mathbf{Y}}^{\mathbf{RRA}} = \left(\tilde{y}_{k,l}^{RRA}\right)_{n^* \times c_y^*}$ is a dependent fuzzy matrix and $\widetilde{\mathbf{X}}_j^{\mathrm{RRA}} = (\tilde{x}_{i,k}^{j,RRA})_{n^* \times c_j^*}$ are explanatory fuzzy matrices for all $j = 1,2,\dots,m$.

Based on (13) we have vectors (fuzzy sets): $\widetilde{\mathbf{Y}}_{min}^{\mathbf{RRA}}, \widetilde{\mathbf{Y}}_{max}^{\mathbf{RRA}}$ and $\widetilde{\mathbf{X}}_{j,min}^{\mathbf{RRA}}, \widetilde{\mathbf{X}}_{j,max}^{\mathbf{RRA}}$, for all $j = 1,2,\dots,m$.

The following example illustrates effectiveness of the method based on intervals vs. conventional regression analysis using traditional regression methods such as MVR: In our case study, just for the year 1985, we ended up with 17 different data series representing our dependent variable. In addition, one of our explanatory variables: exports per capita had 12 data series. Therefore, just to test our model for the year 1985, it would be necessary to perform over 200 regression runs, when trying all possible combinations of those two variables. And what if we decide to test the model for more than one year (in order to have higher degree of confidence in results)? The problem is not only the amount of work, but also the question of how to summarize so many results and to reach meaningful conclusion?

In contrast to the 200 regression runs that would be required by conventional regression methods to cover all possible outcomes, (for the cross national study – year 1985), when using the method presented here, the amount of regression runs 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

In mathematical terms, the four regression runs are as follows:
When we have dependent variable ($\widetilde{\mathbf{Y}}_{min}^{\mathbf{RRA}}, \widetilde{\mathbf{Y}}_{max}^{\mathbf{RRA}}$) and explanatory variables ($\widetilde{\mathbf{X}}_{j,min}^{\mathbf{RRA}}, \widetilde{\mathbf{X}}_{j,max}^{\mathbf{RRA}}$, for all $j = 1,2,\dots,m$) expressed as vectors, the SR process, as explained above, will have to be repeated four times:

1. Vector of min. values of dependent variable ($\widetilde{\mathbf{Y}}_{min}^{\mathbf{RRA}}$) vs. vectors of min. values of explanatory variables ($\widetilde{\mathbf{X}}_{j,min}^{\mathbf{RRA}}$) (i.e., set in (2), $\tilde{Y} = \widetilde{\mathbf{Y}}_{min}^{\mathbf{RRA}}$, $\tilde{X}_j = \widetilde{\mathbf{X}}_{j,min}^{\mathbf{RRA}}$ for all $j = 1,2,\dots,m$).
2. Vector of max. values of dependent variable ($\widetilde{\mathbf{Y}}_{max}^{\mathbf{RRA}}$) vs. vectors of max. values of explanatory variables ($\widetilde{\mathbf{X}}_{j,max}^{\mathbf{RRA}}$) (i.e., set in (2), $\tilde{Y} = \widetilde{\mathbf{Y}}_{max}^{\mathbf{RRA}}$, $\tilde{X}_j = \widetilde{\mathbf{X}}_{j,max}^{\mathbf{RRA}}$ for all $j = 1,2,\dots,m$).
3. Vector of min. values of dependent variable ($\widetilde{\mathbf{Y}}_{min}^{\mathbf{RRA}}$) vs. vectors of max. values of explanatory variables ($\widetilde{\mathbf{X}}_{j,max}^{\mathbf{RRA}}$) (i.e., set in (2), $\tilde{Y} = \widetilde{\mathbf{Y}}_{min}^{\mathbf{RRA}}$, $\tilde{X}_j = \widetilde{\mathbf{X}}_{j,max}^{\mathbf{RRA}}$ for all $j = 1,2,\dots,m$).
4. Vector of max. values of dependent variable ($\widetilde{\mathbf{Y}}_{max}^{\mathbf{RRA}}$) vs. vectors of min. values of explanatory variables ($\widetilde{\mathbf{X}}_{j,min}^{\mathbf{RRA}}$) (i.e., set in (2), $\tilde{Y} = \widetilde{\mathbf{Y}}_{max}^{\mathbf{RRA}}$, $\tilde{X}_j = \widetilde{\mathbf{X}}_{j,min}^{\mathbf{RRA}}$ for all $j = 1,2,\dots,m$).

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 aggregated as ranges between the lowest result and the highest results (see Table 1 and Table 2).
Note: 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.

3 Case Study: Evaluating Economic Prospects of the Soviet led Bloc

3.1 The Model of Factors Facilitating Economic Performance

The model of factors facilitating economic performance was first introduced in [3] and consists of three broad factors that can be considered as facilitating factors (or constraints) for successful long-term economic performance:

1. **International competitiveness**: The term "international competitiveness" reflects the ability of a given country to produce products and services in a competitive manner within international markets. The combination of factors such as product price, quality, reliability, type of warranty, customer support, durability, etc., reflect the various aspects of being competitive. The degree of international competitiveness of an economy at any given time period is a cumulative result of multiple long-term processes.
2. **Human Capital**: Human capital includes factors such as education, knowledge, skills, experience, and tradition. It is reflected by features such as development of new technologies and products, research and development capabilities, advanced technology infrastructure, education and research facilities, organizational and management skills, etc. Human capital is an important factor in determining international competitiveness of the economy, as well as economic efficiency.
3. **Degree of Social Progress**: We characterize socially advanced countries by: Degree of social sophistication and flexibility required for effective functioning of modern and internationally competitive economy, social environment facilitating growth and retention of human capital, higher degree of personal and economic freedom, etc.

There is a definite relation expected between the degree of social progress and the previously defined factor "human capital". In addition, we expect substantial interrelation between human capital (technology, knowhow) and international competitiveness. Hence, the factors included in this model are not independent of each other. This fact constitutes a severe limitation for modeling tools based upon assumption that all explanatory variables are independent (conventional regression methods such as MVR). Therefore, conventional regression methods would not be appropriate modeling tools for this study.

When advancing from the initial stage of theoretical definition of the model to practical implementation, it became apparent that there are no data available in the World Bank databases for the three factors discussed above (international competitiveness, human capital and the degree of social progress). Thus, it was necessary to define proxy variables instead. In order to capture various aspects in the behavior of the original variables, sometimes more than one proxy variable was needed to substitute for the original broad variable, as seen in the section below.

3.2 Proxy Variables

We utilized the following variables as proxies for the three explanatory factors of our model: international competitiveness, human capital and social progress:

1. **Exports per capita (Exports)**- being a proxy for the degree of international competitiveness of a given economy in global markets (adjusted for population size). This variable indicates the bottom line: How much revenue (per capita) was earned by any given country in international markets, no matter what the mix of factors is creating competitive advantages or disadvantages.
2. **Tertiary education enrollment (Tertiary)**- Percentage of the relevant population group that attends tertiary education institutions. Percentage of population attending academic studies can be viewed as a good quantitative proxy for the degree of social progress. It can also be considered as an indicator of investment in human capital – at least from the quantitative viewpoint.
3. **High technology per capita (High-Tech)**- refers to exports (per capita) of products associated with advanced technologies. This variable is an important proxy variable of international competitiveness, representing activities where technologies and human skills are dominant components of competitive advantage. In addition, this variable can supplement "Tertiary Education" variable by illustrating to what extent the skills generated by higher education help to improve competitiveness in the Technology-intensive markets.
4. **Secondary education enrollment (Secondary)**- Percentage of the relevant population group that attends secondary education institutions. This variable represents different aspect of human capital (in comparison to "Tertiary education"). In addition, Secondary Education is also important in influencing social progress based on its unique mix of covered topics, depth of studies and the final outcome of shaping the social characteristics of young generation just entering adulthood.
5. **Birth Rate** - This is a proxy representing a degree of social progress. Large families are in general associated with agrarian economies, where the agricultural sector is usually characterized by traditional (and technologically backward) methods of production. On the other hand, smaller families are usually associated with the aspiration to be part of the middle class (or above), and to acquire education and skills needed for a successful career.

Therefore, as stated above, there is no one-to-one relation between the proxy variables and the variables they supposedly represent:

a. International Competitiveness is represented by: Exports and High Tech.
b. Human Capital is represented by: High Tech, Tertiary and Secondary.
c. Degree of Social Progress is represented by: Tertiary, Secondary and Birth Rate.

It seems that the combinations of proxy variables reflect fairly well the various aspects of variables they supposedly represent. However, it is also clear that the proxy variables are not independent of each other. Therefore, modeling tools assuming independence of explanatory variables cannot be applied successfully in this project.

This is additional argument for using SR, which does not require independence of explanatory variables. This way the integrity and the common sense of the original model have been maintained.

As a dependent variable representing successful long-term economic performance we selected various measures of income/output per capita, such as GDP per capita, GNP per capita and GNI per capita.

3.3 Normalizing Data

We normalize data (see (11) above) by introducing the heuristically determined maximum and minimum thresholds. Data normalizing requires projection of the values from every numerical vector into equivalent normalized numerical vector having values between zero and one, based on predefined function which is expected logically to reflect common sense in projecting such values, while maintaining the integrity of the data. In this study, for every variable we define a group of best economic performers: "High Income Economies". During the normalizing process we assign value of 1 to all the data points which are equal to or greater than the average value for the group of "High Income Economies

The first step in the normalizing process is: we define max_l as the value in a given vector such that all elements equal to or greater than max_l are assigned the value of one. For example, if max_l represents a value of GDP per capita which logically belongs to a category of "High Income Countries", then any country having higher value – will definitely be considered a "High Income Country" as well. We selected "Average of High-Income Economies" as our max_l for the dependent variable as well as for all the explanatory variables. Such average values appear in the data bases and hard copy publications of the World Bank for all variables. By turning all the numbers above max_l into 1, we neutralize the negative effect of the outliers having excessively high values without deleting these data points.

Similarly, we define min_l as the value in that vector such that all elements equal to or smaller than min_l are assigned value of zero, which means they definitely do not belong to the category of "High Income Countries".

We emphasize again: max_l and min_l must be determined based on logic and common sense for each domain (for every variable), so as not to distort the data (for more detailed explanation and example see [3]).

Note: in the cases of several numerical vectors which essentially represent the same variable (see discussion above), the data normalizing procedure explained above brings all these vectors into the same scale, thus helping to express all of them in terms of undistorted intervals (ranges) of values.

In this case, a membership functions in (11) are:

$$\mu_l(x_{k,l}) = \begin{cases} 0 & , x_{k,l} \leq min_l \\ \frac{x_{k,l} - min_l}{max_l - min_l} & , min_l < x_{k,l} < max_l, \\ 1 & , max_l \leq x_{k,l} \end{cases}$$

where $\mathbf{A} = \left(x_{k,l}\right)_{n\times c}$ is a matrix and min_l, max_l are the Maximum cut-off point and Minimum cut-off point as explained above.

3.4 Data Preparation

We utilized cross-national data obtained mostly from the World Bank data bases and hard copy reports. We excluded from the study all the countries having small populations (half a million or less) because small (by population) countries are characterized by different features (such as less diverse and small domestic market, etc.) in comparison to large countries. In particular, when the purpose of the model is to investigate Communist East-European bloc, the exclusion of small countries seems reasonable. Additional countries such as Taiwan and North Korea were excluded due to missing data. The total of over 120 countries were included for the years: 1960, 1970, 1978, 1985 and 1992. We supplemented missing data for individual countries (where it was possible) from adjacent years (this procedure was also used in the world bank hard copy publications). The above-mentioned data supplementing procedure is reasonable in the case of cross section analysis of variables, usually characterized by relatively small annual changes, and in the context of the inherent imprecision of the data in the first place.

There were very few countries in this study, that were deleted by RRA algorithm because of severely unreliable and inconsistent data. This of course had very little influence on the results of a general model where the data for over 120 countries were used. However, one of the problematic countries as far as inconsistency of the data was Bulgaria, which was one of the countries of the Soviet-led bloc, and we excluded it from our study. East Germany was excluded due to excessive amount of missing data.

4 Results

This section consists of the two subsections. The first subsection ("evaluation of the model results") consists of the analysis of the general model, involving its consistency over the years covered under this study, stability, reliability and general conclusions regarding the relative importance of the explanatory variables. The second subsection ("evaluation of the East-European bloc") consists specifically of the analysis of the East-European bloc by its individual countries, based on the results of the model and in comparison to the "High-Income Economies".

4.1 Evaluation of the model results

Similarity results (Table 1) show that the first three proxy variables (Export, High-Tech, Tertiary) are significant every year throughout the period of study (See graphs 1-3). On the other hand, variables Secondary and Birth Rate were significant during 1960 and 1970, but in the following years the lower end of the interval drops into partial significance, and the whole range of the results is gradually declining. In other words, we can see that for both variables higher part of the range is significant (for

1978 and 1985), but the lower part of the range is only partially significant for the same years. It can be interpreted as follows: as more and more countries experienced decrease of their birth rate, as well as managed to enroll increasingly larger percentage of the relevant age group into secondary education, those two variables gradually lost their explanatory power to distinguish between the rich and the poor countries. These two variables are the only proxies used in this study, where the Soviet-led bloc came close to, or actually reached the performance comparable to the "High Income economies". However, the importance of these variables continuously declined towards the end of the period under study, thus undermining these achievements of the communist bloc. (Graphs 4 – 5).

Table. Cross-National model of factors facilitating economic performance

		1960	1970	1978	1985	1992
S_{Y,X_j}	Export	[0.822,0.874]	[0.836,0.894]	[0.791,0.936]	[0.881,0.922]	[0.886,0.924]
	High Tech	[0.856,0.888]	[0.820,0.896]	[0.858,0.897]	[0.886,0.922]	[0.831,0.882]
	Tertiary	[0.853,0.878]	[0.816,0.898]	[0.860,0.863]	[0.845,0.847]	[0.811,0.832]
	Secondary	[0.872,0.891]	[0.867,0.870]	[0.784,0.819]	[0.776,0.819]	[0.701,0.750]
	Birth Rate*	[0.823,0.836]	[0.813,0.845]	[0.784,0.815]	[0.751,0.805]	[0.702,0.739]
Relimp_j	Export	[0.163,0.210]	[0.173,0.240]	[0.154,0.288]	[0.251,0.277]	[0.312,0.379]
	High Tech	[0.196,0.227]	[0.170,0.224]	[0.199,0.281]	[0.230,0.292]	[0.225,0.325]
	Tertiary	[0.193,0.215]	[0.169,0.236]	[0.204,0.238]	[0.189,0.213]	[0.218,0.231]
	Secondary	[0.220,0.229]	[0.200,0.222]	[0.135,0.188]	[0.128,0.163]	[0.058,0.124]
	Birth Rate*	[0.123,0.181]	[0.146,0.199]	[0.135,0.186]	[0.100,0.148]	[0.043,0.109]
$S^{Comb}_{Y,X_1,\ldots,X_n}$		[0.959,0.964]	[0.950,0.960]	[0.957,0.985]	[0.956,0.964]	[0.949,0.965]

*Inverse relation

Note: The dependent variable consisted of the various measurements of income/output per capita

When looking at Relimp_j, we can see that Tertiary Education variable more or less maintains the same relative importance, while Export and High-Tech (which are persistently among the most important variables), having their relative importance gradually increasing due to relative decline of Secondary Education and Birth Rate (Secondary declined continuously since 1970, Birth Rate declined continuously since 1978). By 1985, Export and High-Tech became the two most important variables (both are proxies for "International Competitiveness"). In addition, High-Tech and Tertiary Education, both continuously significant variables are major components of the "Human Capital" factor. Hence, we can summarize Table 1 as follows: the empir-

ical evidence based on cross-national model definitely supports International Competitiveness as well as Human Capital as the major factors facilitating successful economic performance. Since Tertiary education is also a proxy for Social Progress" factor, we can conclude, that based on proxy variables used in this model, Social Progress is also important factor facilitating economic performance, even-though some of its proxy variables became less successful indicators for the later part of the study.

Note: We must keep in mind that the relevant period to evaluate Soviet-led bloc is 1960-1985. The year 1992 represents the situation after the collapse of the bloc. The year is presented because it helps to identify some trends that continued and accelerated after the bloc disintegrated.

Table 1 also displays $\boldsymbol{S}^{\boldsymbol{Comb}}_{Y,X_1,\ldots,X_n}$ measurements: to what extend all the variables combined explain the behavior of the dependent variable. We can see that all the measurements are above 0.949 on the scale between 0 and 1. The high value of $\boldsymbol{S}^{\boldsymbol{Comb}}_{\boldsymbol{Y,X_1,\ldots,X_n}}$ means that the model is highly successful in explaining the behavior of the dependent variable. In addition, the consistency of the results throughout the years under study should be noted.

The consistency and stability of the model between 1960 to 1985 as well as the significance of similarity relations of explanatory variables are important factors determining confidence in the conclusions.

Graphs 1 through 5 display visually the results of the Table 1. In particular, it is important to note that despite the inclusiveness of the study and utilization of all the data series that we could find for every variable, the ranges appear to be fairly narrow except very few cases , and the vast majority of similarity measures are above $\boldsymbol{S}_{Y,X_j} = 0.8$ (which is the lower limit of the significant range). In addition, the graphs show the decline of Secondary Education and Birth Rate in the later years of the study, but still being above 0.7 limit of insignificance – even for the lower end of their range (for the years 1960-1985).

Fig. 1.

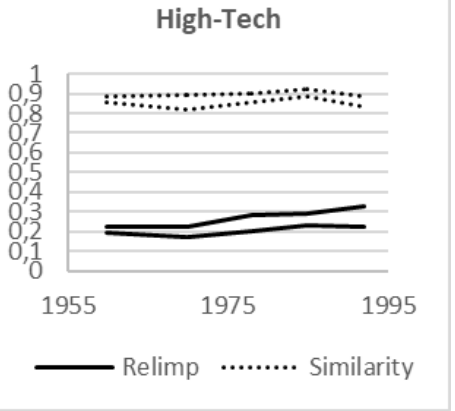

Fig. 2.

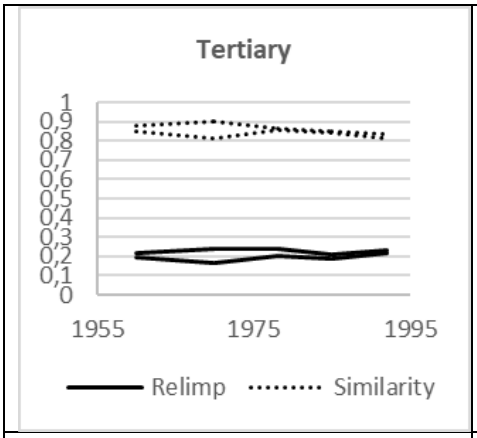

Fig. 3.

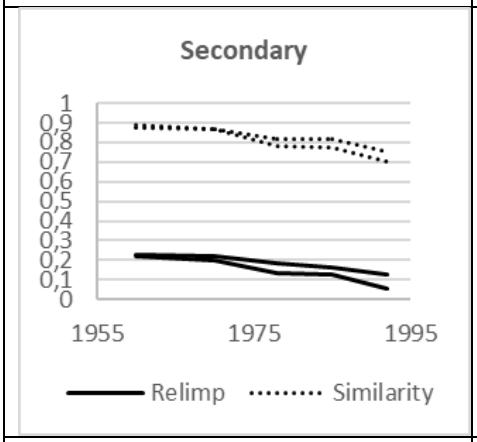

Fig. 4.

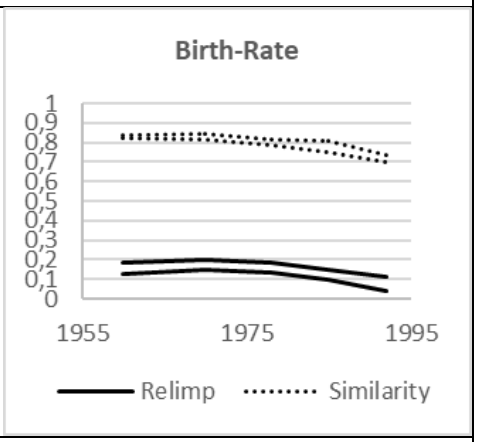

Fig. 5.

4.2 Evaluation of the East-European bloc

Table 2 displays normalized data for individual countries of the Soviet-led bloc. In the cases we had more than one data series for a given variable – we present range of values. If there was only one numerical vector per variable, there is only a single value. The fact that results are normalized makes it easier to compare the status of each Eastern-Bloc country to the average performance of the High-Income developed economies, which have a value of 1.

Table 1. Normalized data for the East-European bloc

		USSR	Poland	Czechoslo-vakia	Hungary	Romania
GDP	1960 1970 1978 1985 1992	[0.391,0.442] 0.475 [0.427,0.444] 0.388 [0.155,0.363]	[0.286,0.330] 0.363 [0.390,0.441] [0.148,0.310] [0.122,0.251]	[0.509,0.620] 0.563 [0.531,0.574] 0.510 0.398	[0.120,0.397] [0.142,0.421] [0.171,0.466] [0.145,0.376] [0.190,0.382]	[0.060,0.120] 0.208 [0.096,0.218] [0.159,0.380] [0.057,0.220]
Export	1960 1970 1978 1985 1992	0.121 0.189 0.144 0.201 [0.063,0.181]	[0.292,0.333] [0.329,0.385] [0.209,0.281] [0.124,0.166] [0.086,0.200]	0.958 0.890 0.613 0.762 0.295	0.768 [0.321,0.520] [0.452,0.499] [0.346,0.520] [0.245,0.373]	[0.246,0.274] [0.270,0.312] [0.217,0.287] [0.178,0.250] [0.031,0.085]
High-Tech	1960 1970 1978 1985 1992	0.154 NA 0.083 NA NA	0.292 0.495 0.346 0.196 0.020	1 1 0.865 NA 0.213	1 0.802 0.435 0.481 0.082	NA 0.291 0.188 NA 0.005
Tertiary	1960 1970 1978 1985 1992	0.600 NA 0.587 0.510 1	0.466 0.437 0.486 0.383 0.428	0.600 0.601 0.432 0.366 0.312	0.33 0.269 0.320 0.367 0.271	0.200 0.264 0.217 0.350 0.154
Second-ary	1960 1970 1978 198	0.648 NA 0.807 1 0.986	0.666 0.773 0.945 0.911 0.901	0.204 0.231 0.344 0.255 0.756	0.592 0.797 0.915 0.939 0.850	0.185 0.467 0.980 1 0.860

	5 199 2					
Birth Rate*	196 0 197 0 197 8 198 5 199 2	0.888 NA 0.931 0.876 1	0.984 1 0.894 0.899 1	1 1 0.931 1 0.995	1 1 1 1 1	1 0.901 0.898 0.982 1

NA-Not Available
*Inverse relation

- **International Competitiveness (represented by proxies Export and High-Tech)**: We can see a general trend in all the countries of the bloc: a major decline in performance by both, Exports and High-Tech variables. Table 2 implies that international competitiveness was one of the major weaknesses of the Soviet-led bloc. We can summarize the performance of the Soviet-led East European bloc regarding international competitiveness as follows: The performance of the bloc was not at its best in 1960, and thereafter continuously deteriorated throughout the period under study. Instead of closing the gap vs. High-Income Developed Economies, the gap continuously widened, thus making the possibility of catching up with the performance level of the leading economies - unattainable.
- **Human Capital (represented by proxies: High-Tech, Tertiary and Secondary)**: Based on Table 2, the performance of Eastern bloc related to Human Capital was far from successful.
  High-Tech: As indicated above (in relation to international competitiveness), the performance regarding the High-Tech variable is indicative of a major failure of the Eastern bloc characterized by widening gap vs High Income economies.
  Tertiary enrollment is another major proxy variable for the "Human Capital" factor. Soviet Union began in 1960 at value of 0.6, and from that point on continuously declined. All other countries of the bloc did not do any better regarding this variable: Czechoslovakia also started at 0.6, but then continuously declined. Poland, Hungary and Romania began in 1960 below 0.5, and remained there throughout the time frame of this study, moving up and down.
  Secondary Enrollment: This is the only component of "Human Capital" factor, where East-European bloc was successful. However, the relative importance of the proxy variable "Secondary Education", has been continuously declining. Thus, the success of the bloc in terms of Secondary Education enrollment had continuously declining impact on the overall performance of the bloc in comparison to the High-Income economies. Hence, the "Human Capital" factor became increasingly determined by the performance in terms

of the two other proxy variables having substantially higher relative importance towards the end of the time frame of this study: High-Tech and Tertiary Education, and in terms of those two parameters, the Soviet-led bloc failed to close the gap vs. High Income economies (in fact the gap widened over time).

- **Degree of Social Progress (represented by proxies: Tertiary, Secondary and Birth Rate)**: Already in 1960, the performance of the bloc was compatible with the High-Income economies and remained more or less at the same level throughout the years under study. However, as in the case of Secondary - the relative importance of the proxy variable "Birth Rate", has been continuously declining since 1970. Combined with the continuous decline in the relative importance of proxy variable "Secondary", this left Tertiary Enrollment as gradually becoming more dominant proxy variable representing the degree of social progress. This is also the variable where Soviet-led bloc failed to improve (see above), while the two other proxies where the bloc was successful, continuously lost their importance by gradually moving from fully significant variables to partially significant variables.

To summarize: based on the method presented in this study for the evaluation of economic performance, the Soviet-led bloc totally failed in the area of International Competitiveness, mostly failed in the area of Human Capital (success in only one proxy variable which continuously declined in its relative importance), and had mixed results in the area of Degree of Social Progress (failure in a major proxy variable, success in two proxy variables which continuously declined in their relative importance). Overall results point overwhelmingly towards the conclusion that the failure of East-European communist bloc to catch-up with the performance of "High-Income Economies" was predictable, based on the data (biased according to the CIA estimates [1]) provided by the government agencies of those countries themselves.
The effectiveness of the method presented in this study is based on the fact that there are certain fundamentals ("Factors Facilitating Economic Performance") which must be satisfied (more or less) for any country to be able to reach and maintain the level of the best performers (High Income Economies). Those fundamentals actually represent constraints that the lagging countries must overcome to reach the level of the best performers, and the Soviet-led countries of Eastern Europe definitely failed to do so.

5 Summary and Conclusions

In this study we presented a Soft Computing/Data Mining method to evaluate economic performance of individual countries or group of countries. As a case study we presented the evaluation of economic performance of the East-European bloc during the period of 1960 – 1985.

We utilized cross-national data to build a general world-wide model of factors facilitating economic performance. We applied the model's results to evaluate the countries of East-European bloc. All the available data series were utilized, including the cases where there were more than one data series for a given variable, which resulted in the application of intervals. Advantages of including all the available data series

and applying intervals in the modeling process were discussed. Soft Regression technique was utilized to build the model. The process, analysis and conclusions are straight-forward and in line with human-logic and common sense. Another important advantage of utilizing Soft Regression in this study was that it allowed successful integration of highly correlated (among themselves) explanatory variables into the same model without being affected by multicollinearity.

The method applied in this study displayed high degree of robustness: the data used for the East-European bloc came mostly from the hard copy publications, published before the disintegration of the bloc. Despite complains (see [1]) regarding the biases and the lack of accuracy of the data provided by the East-European government agencies, the method used in this study managed to identify broadly but accurately, the true standing and prospects of the bloc by its individual countries.

References

1. Harrison M. (2002), Economic Growth and Slowdown, Centre for Russian and East European Studies, University of Birmingham. Appears as a chapter in Bacon E. and Sandle M. – eds. (2002). Brezhnev Reconsidered, London and Basingstoke: Palgrave, pp. 38-67.
2. Kandel A., Last M. and Bunke H. (Eds.) (2001) Data Mining and Computational Intelligence. Physica-Verlag Publishing, pages 251 - 272.
3. Shnaider E. and Haruvy N. (2008), Background Factors Facilitating Economic Growth Using Linear Regression and Soft Regression. Fuzzy Economic Review, vol. XIII, No 1, pages 41-55.
4. Shnaider E., Haruvy N., and Yosef A. (2014). ”The soft regression method – suggested improvements”. Fuzzy Economic Review, Volume XIX, Number 2, pp. 21-33.
5. Shnaider E. and Schneider M. (2000), Heuristic Significance Test for Economic Modeling. Fuzzy Economic Review, Volume V, Number 2, pp. 49-69.
6. Shnaider E., Schneider M. and Kandel A. (1997), Fuzzy Measure for Similarity of Numerical Vectors, Fuzzy Economic Review, Vol. II, No. 1, pp.17 - 38.
7. Shnaider E. and Schneider M. (2005). "Soft Regression”. Published in Maimon O. and Rokach L. Eds. Data Mining and Knowledge Discovery Handbook, Springer Science and Business Media, pages 522-525.
8. Shnaider E. and Yosef A. (2018a), Relative Importance of explanatory variable: Traditional method vs Soft Regression, International Journal of Intelligent System, vol. 33, issue 6, pages 1180-1196.
9. Shnaider E. and 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.
10. Yosef A. and Shnaider E. (2017), On Measuring the Relative Importance of Explanatory Variables in a Soft Regression Method. Advances and Applications in Statistics. Vol. 50, No. 3, p. 201 – 228.
11. Yosef A., Shnaider E. and Haruvy N. (2015), Soft Regression vs Linear Regression. Pioneer Journal of Theoretical and Applied Statistics. Volume 10, Numbers 1-2, 2015, Pages 31-46.
12. Zadeh, L. A. (1965), Fuzzy sets. Information and Control 8 (3): 338.
13. Chang Y.H.O. and Ayyub B.M., (2001) Fuzzy regression methods-a comparative assessment, Fuzzy Sets and Systems, vol. 119, Issue 2, pages 187-203.

14. Savic D.A. and Pedrycz W., (1991) Evaluation of fuzzy linear regression models, Fuzzy Sets and Systems, vol. 39, Issue 1, pages 51-63.
15. Peters G., (1994) Fuzzy linear regression with fuzzy intervals, Fuzzy Sets and Systems, vol. 63, Issue 1, pages 45-55.
16. Schafer J. L. and Graham J. W., (2002) Missing Data: Our view of the state of the art, Psychological Methods, 7 (2), pages 147-177.