Data Preparation for Fuzzy Modelling using Intervals

Arthur Yosef[1], Moti Schneider[2], Eli Shnaider[3], Amos Baranes[3] and Rimona Palas[4]
[1]*Tel Aviv-Yaffo Academic College, Israel*
[2]*Netanya Academic College, Israel*
[3]*Peres Academic Center, Israel*
[4]*College of Law and Business, Israel*

Keywords: Data Mining, Fuzzy Logic, Intervals, Central Tendency, Data Preparation.

Abstract: Model-building professionals are often facing a very difficult choice of selecting relevant variable/s from a set of several similar variables. All those variables are supposedly representing the same factor but are measured differently. They are based on different methodologies, baselines, conversion/comparability methods, etc., thus leading to substantial differences in numerical values for essentially the same things. In this study we introduce a method that utilizes intervals to capture all the relevant variables that represent the same factor. First, we discuss the advantages utilizing intervals of values from the stand point of reliability, better and more efficient data utilization, as well as substantial reduction in the complexity, and thus improvement in our ability to interpret the results. In addition, we introduce an interval (range) reduction algorithm, designed to reduce excessive size of intervals, thus bringing them closer to their central tendency cluster. Following the theoretical component, we present a case study. The case study demonstrates the process of converting the data into intervals for two broad economic variables (each consisting of several data series) and two broad financial variables. Furthermore, it demonstrates the practical application of the procedures addressed in this study and their effectiveness.

1 INTRODUCTION

1.1 Description of a Problem

The purpose of modelling is to explain behaviour of a given dependent variable. We must determine on theoretical grounds, what are the explanatory variables that can explain such behaviour. After determining theoretically, what variables should be included in the model, the next step is to locate numerical measurements of those variables (both dependent and explanatory). It is not always simple and straight-forward process. In the case study presented below, we are presenting several examples of variables which can be represented by several (and in some cases large number) of data series. For example, one of the variables presented in our case study is: the measurements of aggregate economic activity per capita, such as GDP per capita, or GNI per capita (or in previous years – GNP per capita). Those are very common and widely used measurements. Which among the three is the most appropriate? There are additional variations of data. For example, if we are utilizing cross-national data, and since all the values are presented in U.S. dollars, there are additional differences among various data series due to currency conversion methods or due to different baselines. There are data in current U.S. dollars (USD), as well as data in constant 1990 USD, in constant 1995 USD, in constant 2000 USD, and in constant 2005 USD. There are data series based on regular currency conversion method vs. PPP (purchasing power parity) conversion method. Also, several data series based on current USD were downloaded in different years and differ substantially from each other due to changes in measurement methodology over time. Thus, despite the fact, that all these measurements are (from our perspective) measuring essentially the same thing, there are very substantial differences among various data series in terms of values, and even in their scale. For example, for the year 1985, we ended up with 17 different data series representing "aggregate economic activity per capita".

In most cases, modellers do not use all the possible data series, but rather select one or several such series. The question is: which of the various data series to select? Most modellers either select

Yosef, A., Schneider, M., Shnaider, E., Baranes, A. and Palas, R.
Data Preparation for Fuzzy Modelling using Intervals.
DOI: 10.5220/0007947102470257
In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019)*, pages 247-257
ISBN: 978-989-758-382-7

the most popular and easiest to obtain variables. In some other cases the decision is based upon the availability of data, amount of missing observations, etc. The less legitimate approach is: to try several different variables, and then select the ones generating results that best facilitate the conclusions these modelers want to reach. Of course, there is always a possibility of criticism: why a given selection among the data series was made, and not another. The method introduced in this study precludes such criticism, since all the data series are utilized.

1.2 Advantages of Utilizing Intervals

In this study we introduce a method of converting numerical vectors into ranges (intervals) of values that are derived from all the available data series. There are several important advantages of transforming available data into intervals of values:

a. The very basic principle in the field of Information Systems is: all available data are valuable (unless suspected of being severely distorted) and should be utilized in the modeling process.

b. Confidence in the modelling results: when the approach is inclusive and involves all the available data series, then obviously the confidence in results is greater vs. modelling process involving selected data series while ignoring others.

c. Efficient handling of missing observations: This issue arises when in many data series there is a large number of missing measurements. For example, in our case study, we utilized economic data from over 125 countries (for variables: aggregate economic activity per capita and exports per capita), but in many data series (numerical vectors), we encountered a problem of missing data for dozens of countries. In addition, the set of missing countries was not the same in different data series. However, the problem of missing data was resolved by constructing intervals for every country, for which there was at least one measurement. Of course, in some intervals there were more data points and in others less, but we included all these countries in the modeling process, and thus increased our confidence in the results.

d. It is much easier to reach meaningful and unambiguous conclusion due to the drastic reduction of the amount of regression runs. For example, if our dependent variable is “aggregate economic activity per capita” (17 data series), and our explanatory variable is “exports per capita” (12 data series), then when trying all possible combinations of these variables, we will have to perform over 200 regression runs. The problem here is not only the amount of work, but also the question of how to summarize so many results and to reach meaningful conclusion? However, when using the method presented here, the amount of regression runs drops to 4:

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

Note: It does not matter how many explanatory variables are expressed in terms of intervals, the method will still require only four regression runs.

The four regression runs generate four results, which again can be reduced to an interval between the minimum and the maximum value of the results, and this interval can be used to draw conclusions as well as for further computations.

1.3 Literature Survey

The idea of utilizing intervals in fuzzy information processing is not new. Schneider and Kandel (1993) introduced the idea of utilizing Fuzzy Expected Intervals (FEI) in order to handle higher degrees of uncertainty in Fuzzy Expert Systems. Wagman et al. (1994) proposed to generate intervals of real numbers to be processed by the fuzzy matching algorithm.

Nguyen and Kreinovich (1996) address the issue of estimating intervals within the domain of physical measurements during the manufacturing process. If here is a variable Y, which cannot be measured directly, (or very difficult to measure directly) then it is estimated indirectly (the procedure is called “indirect measurement”), using a related variable X. Due to imprecision of measurements, numerical values of X are measured in terms of intervals, and the authors address the issue of estimating the corresponding intervals of the computed variable Y.

Hans and Gottwald (1995), define theoretically various implementations of fuzzy intervals. The authors present ways to define fuzzy interval, such as (a) defining a crisp interval to form the kernel, from which the membership function decreases to zero, or (b) by two fuzzy numbers representing the edges of interval. The authors also describe mathematical operations on fuzzy numbers as well as fuzzy intervals. They discuss “interval mathematics” which is a field of numerical mathematics that originated out of the usual calculus of errors and is based on the idea to work directly with intervals (instead of real numbers and their error bounds).

Ip et al. (2003) postulate, that when data are scattered, the obtained regression model generates a possibility range that is too wide. Thus, they apply fuzzy linear regression with fuzzy intervals and used validation experiments to demonstrate effectiveness of the method Bustince and Burillo (1995) introduce the concepts of correlation and correlation coefficients of interval-valued intuitionistic fuzzy sets.

Grzegorzewski (2002) addresses a problem of interval approximation of fuzzy numbers. He proposes a new interval approximation operator designed as a measure of distance between fuzzy values. Cheng and Mon (1993) introduce approach of evaluating fuzzy system reliability by interval arithmetic and α-cuts. They demonstrate through theoretical analysis as well as an example the simplicity and generality of their proposed approach. D’Urso et al. (2015) introduce a fuzzy clustering model for interval-valued data. In order to avoid negative effects of possible outliers on the clustering process, they propose a robust method with a trimming rule.

Fuzzy logic and the theory of Fuzzy sets were introduced by Zadeh (1965) and since then have been widely applied in various branches of information processing. The modelling method based on Soft Regression, where historical data is converted into intervals and where interval related problems are extensively discussed is presented in Shnaider. and Yosef (2018).

2 THE METHOD FOR CONSTRUCTING INTERVALS

2.1 Constructing the 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 supposedly represent the same thing, we can construct a matrix, such that each numerical vector is a column in that matrix. For example, in our case study, we utilize 17 variables representing aggregate economic activity per capita for the year 1985. Thus, we create a matrix, where 17 variables appear as columns in that matrix, while each row represents a data for a given country. Therefore, for countries, which appear in all 17 variables (columns), we can construct an interval of values, which consists of 17 numbers. The interval will be defined by its smallest value and its largest value. Obviously, similar intervals can be created for countries that do not appear in all 17 variables – those will be intervals containing fewer measurements. In the extreme cases, where a country appears in only one numerical vector (out of 17), then its range will be represented by the same value as the minimum and the maximum. Thus, the matrix of 17 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.

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 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, 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. This way, for every row, we construct intervals consisting of grades of membership.

2.2 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.

Moreover, in order to perform successful modeling, it is desirable to identify the core area of each interval which represents, 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 introduce 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 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.

3 RANGE REDUCTION ALGORITHM (RRA)

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

3.1 Preparation Stage

Let's assume that we have c numerical vectors, each consisting of n elements (In other words, we have a matrix $\mathbf{A} = \left(x_{k,l}\right)_{n\times c}$ 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., A fuzzy matrix of $\mathbf{A}$ is a matrix:

$$\widetilde{\mathbf{A}} = \left(\tilde{x}_{k,l}\right)_{n\times c} \tag{1}$$

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 functionfor all $l = 1,2,\dots,c$.

3.2 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. 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, because once we delete a given element, there will remain another one which is almost the same, and then there could be an additional one, etc.

Another important point to consider: when deleting elements from the matrix, we must keep in mind that some rows might consist of very few measurements. No element should be deleted, 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.

a. 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:

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

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.

b. $col \leftarrow 2; \mathrm{del} \leftarrow 0$

c. Consider the columns 1 and col in matrix $\widetilde{\mathbf{A}}^{\mathrm{Left}}$.
if
$K = \left\{k: values_k^{\widetilde{\mathbf{A}}^{\mathrm{Left}}} > 6\right\} \neq \emptyset$
and $\frac{1}{|K|}\sum_{k\in K} |\tilde{x}^{\mathrm{L}}_{k,1} - \tilde{x}^{\mathrm{L}}_{k,col}| < 0.05$

Then

1. delete from the column col all the elements from the rows where there are 7 measurements or more (we say that columns 1 and col are almost identical)

2. $col \leftarrow col + 1; \mathrm{del} \leftarrow \mathrm{del} + 1$

3. Go-to step (c)

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

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

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

e. $col \leftarrow c - 1$

f. Consider the columns c and col in matrix $\widetilde{\mathbf{A}}^{\mathrm{Right}}$.
if $K = \{k:$ 〖$values$〗 _k^($\mathbf{A}$˜^Right) $> 6\} \neq \emptyset$ and $\frac{1}{|K|}\sum_{k\in K} |\tilde{x}^{\mathrm{R}}_{k,c} - \tilde{x}^{\mathrm{R}}_{k,col}| < 0.05$
where $values_k^{\widetilde{\mathbf{A}}^{\mathrm{Right}}}$ is a number of values in row k of matrix $\widetilde{\mathbf{A}}^{\mathrm{Right}}$

Then

1. delete from the column col all the elements from the rows where there are 7 measurements or more (we say that columns c and col are almost identical)

2. $col \leftarrow col - 1; \mathrm{del} \leftarrow \mathrm{del} + 1$

3. Go-to step (f)

g. 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
$\widetilde{\mathbf{B}} = \left(\tilde{b}_{k,l}\right)_{n\times\tilde{c}}$ when $\tilde{c} = c - \mathrm{del}$

3.3 Reducing Range: Deleting Outlying Elements

a. Create additional matrix $\widetilde{\mathbf{D}} = \left(\tilde{d}_{k,l}\right)_{n\times(\tilde{c}-1)}$ 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 $\widetilde{\mathbf{B}}$ for each row k, between element $\tilde{b}_{k,l+1}$ and element $\tilde{b}_{k,l}$ for $l = 1,2,\dots,\tilde{c}-1$).

b. For any given row k having $values_k^{\widetilde{\mathbf{B}}} > 4$ amount of elements, we can delete

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

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

$$max\left\{\sum_{l=1}^{\beta_k} \tilde{d}_{k,l}\, ,\sum_{l=0}^{\beta_k-1} \tilde{d}_{k,last-l}\, ,\sum_{l=1}^{\delta} \tilde{d}_{k,l} + \sum_{l=0}^{\gamma-1} \tilde{d}_{k,last-l} \text{ where } \delta+\gamma=\beta_k\right\} \tag{4}$$

when $\tilde{d}_{k,last}$- is the last element of the interval in row k of matrix $\widetilde{\mathbf{D}}$. In (4) 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 $\widetilde{\mathbf{B}}$, β_k elements that correspond to the maximum term found in (4).

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

$$\widetilde{\mathbf{R}} = \left(\tilde{r}_{k,l}\right)_{n\times c^*} \tag{5}$$

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

3.4 Additional Reduction of the Interval

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 $\widetilde{\mathbf{R}}$). We consider interval size of $Range_k > 0.25$ as excessively large (See note below).

a. $s \leftarrow 1$
If $Range_k > 0.25$ and $values_k^{\widetilde{\mathbf{R}}} > 4$
where $values_k^{\widetilde{\mathbf{R}}}$ is a number of values in row k of matrix $\widetilde{\mathbf{R}}$

```
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. $last \leftarrow last - 1$
```

b. 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 modelling professional and could differ based on circumstances and constrains. In our studies we selected 0.3 as a cut-off point. In other words, if after all the deletions of the RRA there was still an interval wider than 0.3, then the whole row was deleted from the data.

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 modelling 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 modelling problem, which requires to re-specify the model - redefine the variables included in the model. In our case study, for each of the variables, we deleted only a very small percentage of the rows in all the variables (see Tables 3a and 3b) – which is not expected to affect final results of the modelling. It also can be seen in Table 3b, that for a variable, which was inappropriate for RRA reduction process, the percentage of deleted rows is substantially higher.

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

$$\tilde{\mathbf{A}}^{\mathbf{RRA}} = \left(\tilde{x}_{k,l}^{RRA}\right)_{n^* \times c^*} \qquad (6)$$

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 $\tilde{\mathbf{A}}^{\mathbf{RRA}}$:

$$\tilde{\mathbf{A}}^{\mathbf{RRA}}_{min} = \left(\tilde{a}_1^{min}, \tilde{a}_2^{min}, \dots, \tilde{a}_{n^*}^{min}\right) \text{ and}$$
$$\tilde{\mathbf{A}}^{\mathbf{RRA}}_{max} = \left(\tilde{a}_1^{max}, \tilde{a}_2^{max}, \dots, \tilde{a}_{n^*}^{max}\right) \qquad (7)$$
$$\text{where } \tilde{a}_k^{min} = \min_{l=1,2,..,c^*}\left\{\tilde{x}_{k,l}^{RRA}\right\} \text{ and } \tilde{a}_k^{max} = \max_{l=1,2,..,c^*}$$

$\left\{\tilde{x}_{k,l}^{RRA}\right\}$ (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).

4 CASE STUDY

In this study we present examples of using intervals in two domains: economics and finance. In particular, we emphasize heuristics when defining a membership function, such that the data transformation corresponds to the logic of predefined fuzzy set, and the integrity of the data is maintained.

As an example, we present two economic variables and two finance variables. Each of the variables is represented by a number of data series (numerical vectors) as presented in Tables 1 and 2.

Aggregate Economic Activity per Capita (AEA/Cap): The variables that represent aggregate economic activity per capita are: GDP/Cap, GNI/Cap and GNP/Cap. All of them are considered common and legitimate measurements. Some of these data series are in current U.S. dollars (USD), while others are in constant 1990 USD, in constant 1995 USD, in constant 2000 USD, and in constant 2005 USD. There are data series based on regular currency conversion method vs. PPP (purchasing power parity) conversion method. Also, since we are using data bases, downloaded in different years (2004, 2009, 2015), there probably were differences in measurement methodology because the numbers were different. Thus, we ended up with 17 variables representing aggregate economic activity per capita in 1985.

Exports per Capita: The variables that represent “exports per capita”, are: Merchandise Exports, Exports of Goods and Services, Exports of Goods and Services-BoP, Exports of Goods, Services and Income- BoP. We found these variables in current USD, in constant 1995 USD, in constant 2000 USD and in constant 2005 USD. Also, similarly to the case above, since we are using data bases,

Table 1: Economic data series (1985).

Aggregate Economic activity per capita	Exports per capita
1. GNP/Cap. WDR 1987. Current USD 2. GDP/Cap. 1990 intl. Geary-Khamis \$ 3. GDP/Cap (constant 1995 US\$) 4. GNI/Cap – Atlas (current US\$ - 2004) 5. GDP/Cap, PPP (constant 1995 intl. \$) 6. GDP/Cap, PPP (current intl. \$ - 2004) 7. GNI/Cap, PPP (current intl. \$ - 2004) 8. GDP/Cap (current US\$ - 2004) 9. GDP/Cap (constant 2000 US\$) 10. GNI/Cap, Atlas (current US\$ - 2009) 11. GDP/Cap, PPP (constant 2005 intl. \$) 12. GDP/Cap, PPP (current intl. \$ - 2009) 13. GNI/Cap, PPP (current intl. \$ - 2009) 14. GDP/Cap (current US\$ - 2015) 15. GDP/Cap (constant 2005 US\$) 16. GNI/Cap, Atlas (current US\$ - 2015) 17. GNI/Cap (constant 2005 US\$)	1. Merchandise Exports per capita (current USD), WDR 1987. 2. Exports of goods and services per capita (BoP, current US\$ - 2004) 3. Exports of goods and services per capita (constant 1995 US\$) 4. Exports of goods, services and income per capita (BoP, current US\$ - 2004) 5. Merchandise exports per capita (current US\$ - 2004) 6. Exports of goods and services per capita (BoP, current US\$ - 2009) 7. Exports of goods and services per capita (constant 2000 US\$) 8. Exports of goods, services and income per capita (BoP, current US\$ - 2009) 9. Exports of goods and services per capita (current US\$ - 2009) 10. Exports of goods and services per capita (constant 2005 US\$) 11. Exports of goods and services per capita (current US\$ - 2015) 12. Merchandise exports per capita (current US\$- 2015)

GDP/Cap, GNI/Cap and GNP/Cap are GNP per capita, GNI per capita and GNP per capita, respectively.

Table 2: Finance data series (2012).

Solvency	Profitability
1. Interest Coverage Ratio 2. Cash from Operations (CFO) to Total Debt	1. ROA (Return on Assets) 2. Pre-taxes income over Sales 3. Net Profit Margin 4.Research & Development Expense to Sales 5. Operating Income to Total Assets 6. EBITDA Margin Ratio: (Earnings Before Interest, Taxes, Depreciation and Amortization) to Total Revenue

downloaded in different years (2004, 2009, 2015), there probably were differences in measurement methodology because the numbers were different. Thus, we ended up with 12 Export per capita variables in 1985.

Profitability: Profitability ratios represent the relative measures of the earnings the company created, and therefore have the closest association with the profits. Each one of the above proxy variables explains some aspects of profitability.

Since there are six proxy variables, the RRA reduction process is applicable.

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.

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.

Since this variable consists of only two data series, the RRA is not applicable in this case.

4.1 Normalizing Economic Data

As discussed above, in order to create intervals, it is necessary to bring all the different numerical vectors representing a given variable into the same scale. When utilizing computing methods based on fuzzy logic, defining all the data series of a variable as members of the same fuzzy set actually brings all of them into the same scale.

For the economic variables presented in this study, we define a fuzzy set of "Successful Economies". The conversion of data series from numerical vectors consisting of raw data into numerical data of the elements of fuzzy set is done via process of data normalizing. We normalize data by introducing the heuristically determined maximum and minimum thresholds.

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 (they are full members of the fuzzy set "Successful Economies". We selected "Average of High-Income Economies" as our max_l for the both economic variables in this study. 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 numerical 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 "Successful Economies". We selected "Average of Low-Income economies as our min_l . Those average values also appear in the data bases and hard copy publications of the World Bank for all variables. By turning all the numbers below min_l into 0, we neutralize the negative effect of the outliers having excessively low values without deleting these data points.

In this case, a membership functions (in (1) and (2)) for the relevant data series are:

$$\mu_l(x_{k,l}) = \begin{cases} 0 & , x_{k,l} \leq min_l \\ \dfrac{x_{k,l} - min_l}{max_l - min_l} & , min_l < x_{k,l} < max_l \\ 1 & , max_l \leq x_{k,l} \end{cases} \quad (8)$$

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.

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. They also depend on what we are trying to model. If the objective is to build cross-national model involving various world economies, then the procedure presented above is appropriate and logical. All countries can be evaluated in reference to the best performers (High income) and the worst performers (Low income). However, if we are trying to model the best performers within the set of "High-Income Economies" in comparison to other countries within the same group of "High-Income Economies", then the normalizing procedure presented above would be inappropriate and illogical for such task, and different membership function would be required.

4.2 Normalizing Financial Data

The financial variables in this case study were used to construct a model to evaluate earnings of corporations traded in the major U.S. stock markets (AMEX, NASDAQ, and NYSE). The data was extracted from the data base (XBRL) containing financial reports the companies traded in the stock market are required to submit. The data for 1585 manufacturing industry companies were downloaded (years 2012 – 2016).

All the downloaded companies were divided into three groups:

1. The group of "Winners": companies which were continuously profitable, reported a positive net income, on annual basis for every year between 2012 to 2016 (including 2012 and 2016).
2. 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).
3. All the remaining companies, the "Middle Group".

We ended up with 622 companies having positive operating income for consecutive 5 years (2012-2016), 246 companies with negative operating income for consecutive 5 years, and 398 companies that had positive and negative operating income over the 5 years.

max_j for the year 2012 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_j.

min_j for the year 2012 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_j.

Justification: As was stated above, the process must be in line with human logic and common sense and modellers should be capable of defending their decisions. For example, for max_j , 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 a full member 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, we prevent the vast majority of the borderline cases 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_j.

4.3 A Note regarding the Normalizing Process of Inversely Related Variables

When we define fuzzy sets for the purpose of explaining the behaviour of a dependent variable, corresponding fuzzy sets are defined for the explanatory variables, so that the modelling be meaningful. In the case of the four variables presented in this case study, the relations between the relevant variables are direct, and the normalizing procedure presented above is logical and meaningful. However, when there is an inverse relation between a dependent variable and an explanatory variable, applying the same exactly process as in the case of directly related variables will not work. For example, in economic model when there is a direct relation, then max_j (reflecting “High Income Economies”) are associated with high number and min_j (reflecting “Low Income Economies”) are associated with low numbers. However, when the relation between variables is inverse, “High Income Economies” group will be associated with low numbers, and “Low Income Economies” will be associated with high numbers. The membership function as defined above still remains the same, but the relevant fuzzy set is defined in inverse: Full members of the fuzzy set are the economies that are definitely not members in “High Income Economies” group. Therefore, the “Low Income Economies” will be defined as full members of this fuzzy set and be assigned the value of 1.

Thus, the Average of Low Income Economies will become max_j , and conversely, the Average of High Income Economies” will become min_j . Hence, the equation (8) applies in both cases – direct and inverse.

In the case of Financial model, if a given variable is characterized by large values in the group of “Losers” and small values in the group of “Winners”, then we have an inverse relation.

In this case we define max_j and min_j *as* follows:

max_j: 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 is selected as max_j.

min_j: the values of the companies belonging to the group of “Winners” were arranged from lowest to highest, and then divided into four quarters. The lowest value of the highest quarter was selected as min_j.

5 SOME RESULTS

Tables 3a and 3b demonstrate the effectiveness of range reduction process. In particular, the two tables focus on the amount of intervals characterized by an excessive range. In the Table 3a, the variable “Economic Activity per Capita” consists of 17 data series. Following the normalizing process (which reduced the problem of outliers to some extent), there were still 28 intervals (out of 131) having excessive ranges. However, the RRA process reduced the amount of excessive intervals to 7. Eventually, 4 intervals out of 7 were retained because their range did not exceed the bench-mark of 0.3, while 3 rows had still excessive intervals and were deleted.

Variable “Exports per Capita” consists of 12 data series. Following the data normalizing process, 22 intervals still had excessive ranges. However, the RRA procedure reduced the amount of the excessive intervals to just 2, and eventually only one row was deleted.

Table 3a: Interval Reduction – Economic Variables.

1985	Amount of Columns	Amount of Countries	Amount of Excessive Ranges before reduction	Amount of Excessive Ranges after reduction	Amount of retained excessive intervals (range<0.3)	Amount of deleted rows (range>=0.3)
AEA/Cap	17	131	28	7	4	3
Exports/Cap	12	131	22	2	1	1

Table 3b: Interval Reduction – Finance Variable.

2012	Amount of Columns	Amount of Companies	Amount of Excessive Ranges before reduction	Amount of Excessive Ranges after reduction	Amount of retained excessive intervals (range<0.3)	Amount of deleted rows (range>=0.3)
Profitability	6	6331	1641	253	232	21
Solvency	2	6331	652	652	0	652

Table 4a: Singapore.

1985	income/output per capita	Num. of Elements	Interval	Range
Raw Data	6466.3, 6781.8, …,12192.9, 12333	17	[6466.3, 12333]	5867.2
Normalizing Data	0.520, 0.521, …, 0.819, 0.848	17	[0.520, 0.848]	0.328
After RRA	0.520, 0.521,…, 0.683, 0.735	13	[0.520, 0.735]	0.215

Table 4b: India.

1985	income/output per capita	Num. of Elements	Interval	Range
Raw Data	264.8, 266.3, …, 1003.5, 1078	17	[264.8, 1078.6]	813.8
Normalizing Data	0, 0, …, 0.006, 0.008	17	[0, 0.008]	0.008
After RRA	0, 0,…, 0.001, 0..005	13	[0, 0.005]	0.005

The variable “Profitability” consists of 6 data series, and therefore still allows to apply RRA procedure. Following the normalizing procedure, there were still 1641 intervals having excessive ranges. RRA procedure reduced this amount to merely 253 intervals. Since most of those intervals did not exceed our bench-mark of 0.3, only 21 rows (out of 6331) were deleted.

The results for Solvency are much worse. The reason is: the variable “Solvency” consists of only two data series, and therefore, the RRA procedure is not applicable. Therefore 652 rows were eventually deleted. The contrast vs. other variables presented above demonstrates the effectiveness of the RRA procedure.

Tables 4a-4e demonstrate the effectiveness of the RRA algorithm in terms of individual cases. We decided to choose as an example the variable “Economic Activity per Capita”, because this variable consists of 17 data series. First thing we can observe is that the original raw data are characterized by a very wide discrepancy, even by orders of magnitude. Therefore, without normalizing the data, the intervals will be meaningless group of numbers. In some of the examples below, just the process of normalizing creates a narrow interval instead of wide range visible in raw data (see India and Switzerland below). In our example (for year 1985), the second stage of RRA was inactive because the conditions were not applicable (there were no numerical vectors on the edges that were almost identical). Therefore, we present an example of Hungary for the year 2000, where before applying RRA there were 20 elements in the interval, 9 elements were deleted during the stage 2 of RRA (due to almost identical numerical vectors on the edges of data matrix) and 3 elements were deleted during the stage 3of RRA. This example demonstrates the capability of the RRA to prevent a bias in determining the central tendency of interval due to the prevalence of large number of very similar data series that can potentially misrepresent the interval due to their quantity in one of the edges of the data matrix.

Table 4c: Switzerland.

1985	income/output per capita	Num. of Elements	Interval	Range
Raw Data	14921, 16340, …, 44897, 46496	17	[14921,46496]	31575
Normalizing Data	1, 1,…, 1, 1	17	[1, 1]	0
After RRA	1, 1,…, 1, 1	13	[1, 1]	0

Table 4d: Hungary.

1985	income/output per capita	Num. of Elements	Interval	Range
Raw Data	1880, 1880, …, 9759.7, 11845.9	11	[1880, 11845.9]	9965.9
Normalizing Data	0.146, 0.148, …, 0.508, 0.526	11	[0.146, 0.526]	0.38
After RRA	0.146, 0.148, …, 0.209, 0.377	7	[0.146, 0.377]	0.231

Table 4e: Hungary.

2000	income/output per capita	Num. of Elements	Interval	Range
Raw Data	4620, 4650.2, …,16838.1, 17706	20	[4620, 17706.9]	13086.9
Normalizing Data	0.161, 0.162, …, 0.496, 0.538	20	[0.161, 0.538]	0.377
After RRA	0.163, 0.169, …, 0.257, 0.359	8	[0.163, 0.359]	0.196

6 SUMMARY AND CONCLUSIONS

In this paper we presented a method for data preparation when the variables are represented as intervals. First we have discussed the advantages of utilizing intervals for variables. Then we presented the algorithm for converting singletons into intervals. This included a description of the methods to handle outliers and different scales for representing information. Following the explanation of the algorithm for creating the intervals, we showed an algorithm to reduce the intervals. Finally, a case study was presented to demonstrate practical use of the process presented in this article. The case study demonstrated effectiveness and efficiency of the techniques presented in this study within the domain of economic and financial modeling.

REFERENCES

Bustince H. and Burillo P. (1995). “Correlation of interval-valued intuitionistic fuzzy sets”, Fuzzy Sets and Systems, vol. 74, Issue 2, pages 237-244.

Cheng C.H. and Mon D.L.(1993). “Fuzzy system reliability analysis by interval of confidence”, Fuzzy Sets and Systems, vol. 56, Issue 1, pages 29-35.

D’Urso P., Giovanni L. D. and Massari R. (2015). “Trimmed fuzzy clustering for interval-valued data”, Advances in Data Analysis and Classification, vol. 9, Issue 1, pages 21–40.

Grzegorzewski P. (2002). “Nearest interval approximation of a fuzzy number”, Fuzzy Sets and Systems, vol. 130, Issue 3, pages 321-330.

Hans B. and Gottwald S. (1995), “Fuzzy sets, fuzzy logic, fuzzy methods”. Chichester: Wiley.

Ip C.K.W., Kwong C.K., Bai H. and Tsim Y. C. (2003). “The process modelling of epoxy dispensing for microchip encapsulation using fuzzy linear regression with fuzzy intervals”, The International Journal of Advanced Manufacturing Technology, vol. 22, Issue 5–6, pages 417–423.

Nguyen H. T. and Kreinovich V. (1996). "Nested intervals and sets: concepts, relations to fuzzy sets, and applications." Applications of interval computations. Springer, Boston, MA, pages 245-290.

Schneider M. and Kandel A. (1993). “Expectations in Fuzzy Environments”, Journal of Fuzzy Logic and Intelligent Systems, vol. 3, num. 1, pages. 76-89.

Shnaider E. and Yosef A. (2018). “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.

Wagman D., Schneider M. and Shnaider E. (1994). “On the use of interval mathematics in fuzzy expert systems”, International Journal of Intelligent Systems, vol. 9, Issue 2, pages 241–259.

Zadeh L. A. (1965). “Fuzzy sets”, Information and Control 8 (3), pages 338-353.