European Research Studies Journal
Volume XXVII, Special Issue 3, 2024

pp. 320-332

Imputing Data Gaps in Economic Surveys Using Fuzzy Sets and Artificial Intelligence Technique

Submitted 05/09/24, 1st revision 15/09/24, 2nd revision 21/10/24, accepted 30/10/24

Adam Kiersztyn[1], Krystyna Kiersztyn[2], Korneliusz Pylak[3], Jakub Bis[4], Michał Dolecki[5], Anna Żelazna

Abstract:

Purpose: *This paper develops a novel approach to impute data gaps in economic surveys. In contrast to classical methods relying on statistical analysis of survey data, more advanced prediction techniques combined with fuzzy sets are applied to effectively address missing data.*
Design/Methodology/Approach: *The paper proposes an unconventional approach that integrates advanced prediction methods with fuzzy sets for imputing missing data. The effectiveness of the method is tested on the extensive dataset from the Polish Panel Survey (POLPAN), which was conducted every five years from 1988 to 2018. The survey contains a wide range of questions asked over successive waves, enabling a comprehensive analysis of the method for imputing data gaps.*
Findings: *The results of numerical experiments show that the proposed method performs highly effectively, regardless of the proportion of observations assigned to the training set. Some methods, such as Support Vector Machine (SVM), did not prove suitable for imputing this dataset. The choice and number of explanatory variables play a crucial role in the method's effectiveness, with cases where a single variable was sufficient for accurate imputation.*
Practical Implications: *The proposed method offers practical applications for improving data quality in economic surveys, especially in large-scale longitudinal surveys like POLPAN. It provides new insights into handling missing data and optimizing the selection of explanatory variables, which can enhance the robustness of imputation techniques in complex surveys.*
Originality/Value: *This paper contributes an original and valuable approach by combining advanced prediction techniques with fuzzy sets, providing a highly effective tool for imputing missing data. This unconventional method offers new avenues for further research in economic surveys and beyond.*

Keywords: *Missing value imputation, fuzzy sets, field surveys, POLPAN.*
JEL codes: *C6, C8, C83, D7.*

Paper type: *Research article.*

[1]*Lublin University of Technology, Poland, e-mail: a.kiersztyn@pollub.pl;*
[2]*Lublin University of Technology, Poland, e-mail: k.kiersztyn@pollub.pl;*
[3]*Lublin University of Technology, Poland, e-mail: korneliusz.pylak@pollub.pl;*
[4]*Lublin University of Technology, Poland, e-mail: j.bis@pollub.pl;*
[5]*Lublin University of Technology, Poland, e-mail: m.dolecki@pollub.pl;*

1. Introduction

The increasing volume of research in economics and management demands gathering more and more data, digging for new data sources, and adapting existing ones to new purposes or analytical techniques (Fayolle, Landstrom, Gartner, and Berghund, 2016). Research is also becoming more complex, multistage, and multidimensiona (Griffith, Cavusgil, and Xu, 2008; Xi, Kraus, Filser, and Kellermanns, 2015).

Therefore, the challenge is to collect data sets on the one hand and to ensure their completeness, consistency and thus reliability on the other. Consistency of data sets is challenging not only because of the increasing complexity and the need to combine data sets. It is also due to other reasons, such as ignorance, measurement error, equipment failure (Raja, Sasirekha, and Thangavel, 2020). And missing values in data sets may lead to improper results and misleading conclusion.

Therefore, data consistency testing is an important step in pre-processing. A key element for empirical data is the proper handling of missing data. In the economic sciences, a significant amount of research relies on surveys. Survey research extremely often experiences data gaps, the sources of which are manifold.

Most classical approaches to impute data gaps employ methods derived from statistical analysis (Falge, 2001; Łopucki, Kiersztyn, Pitucha, and Kitowski, 2022; Kiersztyn and Kiersztyn, 2022). Fuzzy techniques are increasingly used in data gap imputation research (Kiersztyn, Kaczmarek, Łopuski, Pedrycz, Al, Kitowski, and Zbyryt, 2020). However, the combination of prediction methods with fuzzy techniques is basically unknown.

Specifically, such an approach is not known and utilized in the economic sciences for the analysis of surveys. Therefore, this paper attempts to apply the latter approach to one of the longest-running surveys for Poland, i.e., the Polish Panel Survey POLPAN carried out since 1988 in 5-year waves.

The paper is structured as follows: the next section is devoted to the description of the analyzed set, followed by the theoretical assumptions of the proposed method in the third section. Section four reports the results of detailed numerical experiments. The last section provides conclusions and further research avenues.

2. Description of the Analyzed Set

The data set used for the numerical experiments concerns, among other things, the first job, which has a decisive impact on the subsequent career of working individuals. Studies have shown that the choice of a first job, especially when there is a mismatch between qualifications and job requirements, can have a negative impact on further career opportunitie (Scherer, 2004).

Therefore, the characteristics of the first job are extremely important for predicting individuals' careers. The details of the first job that were used in the numerical experiments were taken from the POLPAN data set (POLPAN).

POLPAN is a unique panel research program carried out since 1988 at 5-year intervals, focused on describing the social structure and its changes over the last 30 years in Poland. POLPAN is carried out by the Comparative Analysis of Social Inequalities (CASIN) Team at the Institute of Philosophy and Sociology of the Polish Academy of Sciences (IFiS PAN) and in cooperation with researchers from other Polish and international scientific institutions.

Initially, in 1988, the survey was conducted among a nationwide sample representing the adult population of Poland (aged 21–65 years), with N = 5,817. In 1993, this sample was randomly reduced to 2,500 individuals, who were attempted to be reached in each of the subsequent five-year waves of the survey.

To ensure appropriate age balance, additional sub-samples including younger cohorts were subsequently completed. In 2013, CASIN and IFiS PAN attempted to contact all respondents (7 261) who had ever participated in the survey. The most recent wave of the survey was conducted in 2018 (2,161 respondents were interviewed).

This study firstly utilizes the classification of the respondents' first occupation. The most detailed one is the standard classification of occupations (SCO) based on the International Standard Classification of Occupations (ISCO) developed by the International Labor Organization (ILO). In the Polish Classification of 2010, similarly to ISCO-08, modified principles of dividing occupations into groups were adopted, and the classification was extended by, among others, groups of managerial occupations.

Over the years, research has required modifications of SCO in line with changes in the economy. At present, there are 9 main socio-professional groups of SCO in Poland (divided into over two hundred occupations): senior officials and managers, specialists, technicians and specialized office workers, other non-manual middle personnel, sales and service workers, skilled manual workers, semi-skilled and unskilled manual workers, farmers, owners of production and service companies (Domański, Sawiński, and Słomczynski, 2009).

For the purposes of analysis, however, more aggregated divisions are used which help researchers to understand patterns of socio-economic status and the distribution of wealth and opportunities in the population.

One of the most popular is the division into 7 socio-occupational groups and 14 socio-occupational groups giving a more detailed classification.

The second type of variables addressed in this paper are the characteristics of the occupation, the key ones being the Scale of Prestige 1979 and the Scale of Prestige 2009. Referred to as 'the general desirability of occupations' (Goldthorpe and Hope, 1974), the prestige scales involve assessments made either by a full selection (the entire population) or by a group of experts or informed members of the public (Ganzeboom, De Graaf, and Treiman, 1992).

In general, the following scales are being used:

(1) the skill requirements scale, which is the result of assigning to each SCO category a value based on the most detailed level of three variables: general educational development, 'special' vocational skills, and desired level of formal education;

(2) the level of job complexity (elementary task) describing elementary activities involving (a) interpersonal contacts; (b) information processing; and (c) physical effort. Thus, there are three distinct dimensions of occupational description referred to as 'data', 'people' and 'things' (Domański, Sawiński, and Słomczynski, 2009);

(3) material remuneration;

(4) professional prestige are variables identified as rewards received for performing occupational roles.

Remuneration is the result of the education received, which can be identified with the effort an individual puts into preparing for occupational roles, hence remuneration is one form of reward for occupational performance.

Duncan (Duncan, 1961) considered occupation to be a mediating factor between education and income, and proposed the socioeconomic index (SEI), which set the standard in terms of the theoretical rationale and logic for measuring occupations. Indeed, the SEI is the product of education, the prestige of occupations and the level of earnings.

3. Theoretical Description of the Developed Method

It is relatively common to employ statistical attributes of the analyzed dataset to impute missing data. The approach developed in the paper utilizes known prediction methods, but additionally introduces fuzziness to the results obtained. For each of the methods examined, the explained variable was predicted from a different set of explanatory variables. The research used algorithms implemented on the KNIME Analytic Platform (Berthold, 2007).

Furthermore, different prediction methods were used, i.e., Fuzzy Rule (FR) (Berthold, Cebron, Dill, Gabriel, Kotter, Meinl, Ohl, Sieb, Thiel, and Wiswedel,

2003), Decision Trees (DT) (Shafer, Agrawal, and Mehta, 1996), Gradient Boosted Trees (GBT) (Friedman, 2002), Random Forest (RF) (Pal, 2005), Tree Ensemble (TE) (Coppersmith, 1999) and Support Vector Machine (SVM) (Keerthi, Shevade, Bhattacharyya, and Murthy, 2001; Platt, 1999). Prediction results were fuzzified with a strongly intuitive (trapezoidal) membership function. The selection of an example membership function was driven by its intuitiveness and versatility.

Let X[i,j], where 1≤i≤N and 1≤j≤K, denote the data set under consideration. This set is classically identified with a matrix in which the columns correspond to successive variables (questionnaire responses). Given the above denotations, the set consists of K columns and N rows corresponding to successive observations (responses of a single respondent).

In this matrix, x_{i_0,j_0} denotes a single (j_0-th) component of the observation with index i_0. If this value is unknown, a set of the above-mentioned prediction methods was suggested for its reconstruction. In each case, the set of indexes of the explanatory variables is denoted by J.

Thus,

$$\tilde{x}_{i_0,j_0} = f(x_{i_0,j}) : j \in J$$

is the value imputed by one of the prediction methods.

Empirical studies are naturally subject to a certain uncertainty. The optimal way to model uncertainty is to apply fuzzy sets and numbers. Therefore, each predicted value must be converted into a fuzzy number. As mentioned earlier, the approach developed employs a trapezoidal membership function of the form:

$$\mu(x, x_0, \alpha, \beta) = \begin{cases} 1, for\ x \in (x_0 - \alpha, x_0 + \alpha) \\ \dfrac{x - x_0 + \beta}{\beta - \alpha}, x \in [x_0 - \beta, x_0 - \alpha] \\ \dfrac{x_0 + \beta - x}{\beta - \alpha}, x \in [x_0 + \alpha, x_0 + \beta] \\ 0, x \notin (x_0 - \beta; x_0 + \beta) \end{cases}$$

A key aspect of data gap imputation is the high efficiency of the developed method. Clearly, the effectiveness of the method may be most readily tested by artificially introducing gaps into the data and then imputing the missing values. The measure of effectiveness in the proposed solution may be the level of compliance of the reproduced value calculated as calculated as the value of the membership function at a given point.

This measure for a single observation takes values within the range [0,1] and the

higher the value, the better the degree of reconstruction of the missing value. When analyzing a larger number of missing data, it is possible to calculate basic statistics describing efficiency on a larger dataset.

4. Results of Numerical Experiments

The effectiveness of the proposed method was tested on two selected variables describing the respondents' first occupation: (1) its SCO code and (2) its level of prestige. The former is a categorical ordered variable and takes 273 levels from '0000' ('Top government administrators and political officials') to '8600' ('Owners of shops and other commercial establishments'); while the latter takes continuous values in the range [0; 97.3]. The input data set contains 7,796 rows and 5,423 columns. For clarity of consideration, 122 columns describing the respondent's employment and 614 observations containing full information about the individual's employment were selected.

The available dataset was then randomly divided into a training and a testing set to perform the numerical experiments. A different percentage of the data was allocated to the training set, while the remaining observations were included in the test set. To avoid the influence of the random selection of elements into the sets, the whole procedure was repeated 10 times. In each case, a formula was derived from the given training set to make a prediction using the indicated set of explanatory variables.

Table 1 demonstrates the effect of different proportions of observations included in the training set on prediction performance of the first variable. It is worth noting that increasing the number of observations in the training set is not a performance enhancer for some methods. Furthermore, it is worth noting that some of the classical methods are not a suitable tool for this particular task, specifically SVM, which has the lowest level of performance.

It is also worth noting that with 90% of the elements assigned to the training set for the DT method, the average value of the proposed efficiency index is 0.944. In addition, for this method (as well as for the GBT, RF and TE methods), the median value for 50% of the elements in the training set is 1. Interestingly, even the 1st quartile (Q1) is equal to 1. This result should be interpreted as follows: for more than 75% of cases the model value (with accuracy to the proposed fuzzing) coincides with the actual value.

At this point, it should also be noted that the proposed fuzziness with a kernel of 10 and a support of 20 appears to be small. The support value is just over 2 per mille of the range of analyzed value, which may be considered negligible.

In the following discussion, it was assumed that by nature missing data should be rare. It was therefore assumed that 90% of the observations are always assigned to the training set. This assumption appears to be fully justified. From the other hand,

the problem of selecting explanatory variables for the model appears to be an interesting topic. Table 2 shows how the selection of explanatory variables affects the effectiveness of the model describing the variable 'Social Classification of Occupations (SCO) code for the responder's first job'. The same level of fuzziness was assumed as before, and the assumption described above regarding the number of observations in the training set.

Table 1. *Effect of the proportion of observations in the training set on the predictive performance of the first variable describing the social classification of occupations (sco) code for the responder's first job, with all other variables as explanatory variables. kernel = 10, support =20*

Proportion of training set	**Values**	**Prediction method**					
		FR	**DT**	**GBT**	**RF**	**TE**	**SVM**
50%	Mean	0.299	0.791	0.779	0.890	0.874	0.174
	Median	0.000	1.000	1.000	1.000	1.000	0.000
	Q1	0.000	1.000	1.000	1.000	1.000	0.000
60%	Mean	0.336	0.848	0.797	0.832	0.841	0.436
	Median	0.000	1.000	1.000	1.000	1.000	0.200
	Q1	0.000	1.000	1.000	1.000	1.000	0.000
70%	Mean	0.321	0.900	0.913	0.802	0.838	0.442
	Median	0.000	1.000	1.000	1.000	1.000	0.200
	Q1	0.000	1.000	1.000	1.000	1.000	0.000
80%	Mean	0.289	0.886	0.743	0.957	0.903	0.503
	Median	0.000	1.000	1.000	1.000	1.000	0.300
	Q1	0.000	1.000	0.500	1.000	1.000	0.000
90%	Mean	0.311	0.944	0.906	0.889	0.833	0.383
	Median	0.000	1.000	1.000	1.000	1.000	0.000
	Q1	0.000	1.000	1.000	1.000	1.000	0.000

Source: *Authors' calculations.*

Table 2. *Effect of explanatory variable choice on the model effectiveness with the variable describing the social classification of occupations (sco) code for the responder's first job, with 90% of elements in the training set. kernel =10, support =20*

Explanatory variable	**Value**	**Prediction method**					
		FR	**DT**	**GBT**	**RF**	**TE**	**SVM**
Variable describing the scale of skill requirements for the first job	Mean	0.000	0.778	0.778	0.778	0.778	0.667
	Median	0.000	1.000	1.000	1.000	1.000	1.000
	Q1	0.000	1.000	1.000	1.000	1.000	0.000
Variable describing the scale of work complexity for the first job	Mean	0.778	0.889	0.889	0.889	0.889	0.667
	Median	1.000	1.000	1.000	1.000	1.000	1.000
	Q1	1.000	1.000	1.000	1.000	1.000	0.000
Variable describing Polish socio- economic index (PSEI) for the	Mean	1.000	0.889	1.000	1.000	1.000	1.000
	Median	1.000	1.000	1.000	1.000	1.000	1.000
	Q1	1.000	1.000	1.000	1.000	1.000	1.000

first job							
Variable describing the material remuneration for the first job	Mean	0.889	0.889	0.889	0.889	0.889	0.889
	Median	1.000	1.000	1.000	1.000	1.000	1.000
	Q1	1.000	1.000	1.000	1.000	1.000	1.000
Variable describing the scale of prestige (1979) of the first job	Mean	0.778	0.778	0.778	0.778	0.778	0.778
	Median	1.000	1.000	1.000	1.000	1.000	1.000
	Q1	1.000	1.000	1.000	1.000	1.000	1.000
Variable describing the scale of prestige of the first job	Mean	0.889	0.833	0.889	0.889	0.889	0.889
	Median	1.000	1.000	1.000	1.000	1.000	1.000
	Q1	1.000	1.000	1.000	1.000	1.000	1.000
Variable 1 (SCO for all years)	Mean	0.478	0.478	0.544	0.656	0.572	0.428
	Median	0.300	0.400	0.750	1.000	1.000	0.000
	Q1	0.000	0.000	0.000	0.000	0.000	0.000
All explanatory variables describing the first variable	Mean	0.778	0.944	1.000	1.000	1.000	0.622
	Median	1.000	1.000	1.000	1.000	1.000	0.950
	Q1	1.000	1.000	1.000	1.000	1.000	0.050
Variable 1 (SCO for only 1988)	Mean	0.561	0.572	0.489	0.506	0.506	0.517
	Median	0.900	0.900	0.400	0.550	0.550	0.550
	Q1	0.000	0.000	0.000	0.000	0.000	0.000
Variable 1 (SCO for only 1993)	Mean	0.389	0.556	0.389	0.444	0.444	0.378
	Median	0.000	1.000	0.000	0.000	0.000	0.000
	Q1	0.000	0.000	0.000	0.000	0.000	0.000
Variable 1 (SCO for only 1998)	Mean	0.350	0.406	0.606	0.517	0.517	0.183
	Median	0.000	0.100	0.900	0.550	0.550	0.000
	Q1	0.000	0.000	0.050	0.000	0.000	0.000
Variable 1 (SCO for only 2003)	Mean	0.500	0.522	0.522	0.500	0.500	0.467
	Median	0.500	0.700	0.700	0.500	0.500	0.200
	Q1	0.000	0.000	0.000	0.000	0.000	0.000
Variable 1 (SCO for only 2008)	Mean	0.350	0.589	0.283	0.350	0.350	0.278
	Median	0.000	1.000	0.000	0.000	0.000	0.000
	Q1	0.000	0.000	0.000	0.000	0.000	0.000
Variable 1 (SCO for only 2013)	Mean	0.256	0.367	0.311	0.311	0.311	0.311
	Median	0.000	0.000	0.000	0.000	0.000	0.000
	Q1	0.000	0.000	0.000	0.000	0.000	0.000
Variable 1 (SCO for only 2018)	Mean	0.306	0.417	0.400	0.400	0.400	0.350
	Median	0.000	0.150	0.000	0.000	0.000	0.000
	Q1	0.000	0.000	0.000	0.000	0.000	0.000

Source: *Authors' calculations.*

Examining the results presented in Table 2, some very interesting insights may be drawn. Firstly, the construction of a model based on single explanatory variables tends to generate deteriorating levels of performance. On the other hand, the right choice of explanatory variables and forecasting method may yield surprisingly good results. An excellent example of this is the explanatory variable describing the Polish socio- economic index (PSEI) for the first job, which, for most of the methods considered, gives a perfect fit to the actual facts.

An interesting observation is that the values of the SCO variables for successive editions of the survey are not appropriate explanatory variables. None of the methods considered have a satisfactory level of performance.

Interestingly, a model built on data describing SCO in all years of the survey tends to be worse than a model built on one of these variables. A very interesting observation emerges here: too much information leads to wrong conclusions.

An analogous consideration of the effect of the choice of variables on the efficiency of the model was carried out for the second variable describing the prestige of the respondent's first job. Table 3 shows the efficiency statistics for the model describing this variable.

Table 3. *Effect of variable selection on the efficiency of the model with the second variable describing the prestige of the respondent's first job. 90% elements in the training set, kernel =5, support =10*

Explanatory variable	**Value**	**Prediction method**					
		FR	**DT**	**GBT**	**RF**	**TE**	**SVM**
Variable describing occupation with 7 SCO classes	Mean	0.388	0.967	0.967	0.967	0.967	0.388
	Median	0.070	1.000	1.000	1.000	1.000	0.070
	Q1	0.000	1.000	1.000	1.000	1.000	0.000
Variable describing occupation with 14 SCO classes	Mean	0.426	0.944	0.944	0.944	0.944	0.222
	Median	0.370	1.000	1.000	1.000	1.000	0.000
	Q1	0.000	1.000	1.000	1.000	1.000	0.000
All explanatory variables describing the first variable	Mean	1.000	0.890	0.903	0.944	0.944	0.944
	Median	1.000	1.000	1.000	1.000	1.000	1.000
	Q1	1.000	1.000	1.000	1.000	1.000	1.000
Variable describing the scale of skill requirements for the first job	Mean	0.839	0.986	0.986	0.986	0.986	0.546
	Median	1.000	1.000	1.000	1.000	1.000	0.920
	Q1	0.940	1.000	1.000	1.000	1.000	0.000
Variable describing the scale of work complexity for the first job	Mean	0.776	0.831	0.831	0.831	0.831	0.747
	Median	1.000	1.000	1.000	1.000	1.000	1.000
	Q1	0.640	0.710	0.710	0.710	0.710	0.420
Variable describing Polish socio- economic index (PSEI) for the first job	Mean	0.769	0.778	0.778	0.769	0.778	0.778
	Median	1.000	1.000	1.000	1.000	1.000	1.000
	Q1	0.880	1.000	1.000	0.880	1.000	1.000
Variable describing the material remuneration for the first job	Mean	1.000	1.000	1.000	1.000	1.000	0.944
	Median	1.000	1.000	1.000	1.000	1.000	1.000
	Q1	1.000	1.000	1.000	1.000	1.000	1.000
Variable describing the scale of prestige (1979) of the first job	Mean	1.000	0.946	1.000	1.000	1.000	1.000
	Median	1.000	1.000	1.000	1.000	1.000	1.000
	Q1	1.000	1.000	1.000	1.000	1.000	1.000
Variable describing the Social Classification of Occupations (SCO) code for the first job'	Mean	1.000	0.889	1.000	1.000	1.000	1.000
	Median	1.000	1.000	1.000	1.000	1.000	1.000
	Q1	1.000	1.000	1.000	1.000	1.000	1.000
All explanatory variables	Mean	0.553	0.866	0.936	0.936	0.927	0.833
	Median	0.980	1.000	1.000	1.000	1.000	1.000
	Q1	0.000	1.000	1.000	1.000	1.000	1.000
Variable 2 (occupational prestige for all other years)	Mean	0.532	0.594	0.833	0.722	0.722	0.722
	Median	0.790	1.000	1.000	1.000	1.000	1.000
	Q1	0.000	0.000	1.000	0.250	0.250	0.250

Variable 2 (occupational prestige for 1988)	Mean	0.522	0.717	0.633	0.549	0.522	0.633
	Median	0.690	1.000	1.000	0.690	0.690	1.000
	Q1	0.000	0.225	0.005	0.000	0.000	0.005
Variable 2 (occupational prestige for 1993)	Mean	0.769	0.880	0.889	0.880	0.880	0.878
	Median	1.000	1.000	1.000	1.000	1.000	1.000
	Q1	0.920	1.000	1.000	1.000	1.000	0.970
Variable 2 (occupational prestige for 1998)	Mean	0.489	0.547	0.570	0.570	0.570	0.594
	Median	0.400	0.920	0.920	0.920	0.920	1.000
	Q1	0.000	0.000	0.000	0.000	0.000	0.000
Variable 2 (occupational prestige for 2003)	Mean	0.809	0.813	0.813	0.869	0.869	0.758
	Median	1.000	1.000	1.000	1.000	1.000	1.000
	Q1	0.940	1.000	1.000	1.000	1.000	0.730
Variable 2 (occupational prestige for 2008)	Mean	0.620	0.717	0.684	0.684	0.684	0.573
	Median	0.920	1.000	1.000	1.000	1.000	0.750
	Q1	0.100	0.425	0.405	0.405	0.405	0.000
Variable 2 (occupational prestige for 2013)	Mean	0.713	0.598	0.444	0.602	0.602	0.574
	Median	1.000	0.960	0.000	1.000	1.000	0.970
	Q1	0.210	0.000	0.000	0.000	0.000	0.000
Variable 2 (occupational prestige for 2018)	Mean	0.707	0.560	0.498	0.592	0.592	0.451
	Median	1.000	0.790	0.540	0.960	0.960	0.100
	Q1	0.465	0.000	0.000	0.000	0.000	0.000

Source: *Authors' calculations.*

Tables 4 and 5 indicate how the level of efficiency is affected by the level of fuzziness adopted. It is readily apparent that even a small amount of fuzziness provides very high efficiency. The use of fuzzy sets to impute data gaps is an approach that is completely unknown in economic and management sciences. The proposed method opens up completely new avenues for further research.

Table 4. *Effect of kernel and support selection on the effectiveness of the model with the first variable describing the Social Classification of Occupations (SCO) code for the responder's first job, with all other variables as explanatory variable*

Kernel	**Support**	**Value**	**Prediction method**					
			FR	**DT**	**GBT**	**RF**	**TE**	**SVM**
1	2	Mean	0.667	0.778	0.833	0.944	0.944	0.222
		Median	1.000	1.000	1.000	1.000	1.000	0.000
		Q1	0.000	1.000	1.000	1.000	1.000	0.000
5	10	Mean	0.711	0.911	0.944	0.944	0.944	0.444
		Median	1.000	1.000	1.000	1.000	1.000	0.000
		Q1	0.200	1.000	1.000	1.000	1.000	0.000
10	20	Mean	0.778	0.944	1.000	1.000	1.000	0.622
		Median	1.000	1.000	1.000	1.000	1.000	0.950
		Q1	1.000	1.000	1.000	1.000	1.000	0.050
20	40	Mean	0.814	0.944	1.000	1.000	1.000	0.853
		Median	1.000	1.000	1.000	1.000	1.000	1.000
		Q1	1.000	1.000	1.000	1.000	1.000	0.963
30	60	Mean	0.833	0.944	1.000	1.000	1.000	0.889
		Median	1.000	1.000	1.000	1.000	1.000	1.000
		Q1	1.000	1.000	1.000	1.000	1.000	1.000

40	80	Mean	0.833	0.963	1.000	1.000	1.000	0.894
		Median	1.000	1.000	1.000	1.000	1.000	1.000
		Q1	1.000	1.000	1.000	1.000	1.000	1.000
50	100	Mean	0.833	0.981	1.000	1.000	1.000	0.916
		Median	1.000	1.000	1.000	1.000	1.000	1.000
		Q1	1.000	1.000	1.000	1.000	1.000	1.000
100	200	Mean	0.833	1.000	1.000	1.000	1.000	0.944
		Median	1.000	1.000	1.000	1.000	1.000	1.000
		Q1	1.000	1.000	1.000	1.000	1.000	1.000
500	1000	Mean	0.833	1.000	1.000	1.000	1.000	0.955
		Median	1.000	1.000	1.000	1.000	1.000	1.000
		Q1	1.000	1.000	1.000	1.000	1.000	1.000

Source: *Authors' calculations.*

Table 5. *Effect of kernel and support selection on the effectiveness of the model with the variable describing the prestige of the respondent's first job, with all other variables as explanatory variables*

Kernel	**Support**	**Value**	**Prediction method**					
			FR	**DT**	**GBT**	**RF**	**TE**	**SVM**
1	2	Mean	0.889	0.833	0.889	0.889	0.889	0.889
		Median	1.000	1.000	1.000	1.000	1.000	1.000
		Q1	1.000	1.000	1.000	1.000	1.000	1.000
2	4	Mean	0.981	0.947	0.981	0.981	0.981	0.925
		Median	1.000	1.000	1.000	1.000	1.000	1.000
		Q1	1.000	1.000	1.000	1.000	1.000	1.000
2	6	Mean	0.990	0.974	0.990	0.990	0.990	0.935
		Median	1.000	1.000	1.000	1.000	1.000	1.000
		Q1	1.000	1.000	1.000	1.000	1.000	1.000
3	4	Mean	1.000	0.989	1.000	1.000	1.000	0.944
		Median	1.000	1.000	1.000	1.000	1.000	1.000
		Q1	1.000	1.000	1.000	1.000	1.000	1.000
3	5	Mean	1.000	0.994	1.000	1.000	1.000	0.944
		Median	1.000	1.000	1.000	1.000	1.000	1.000
		Q1	1.000	1.000	1.000	1.000	1.000	1.000
3	6	Mean	1.000	0.996	1.000	1.000	1.000	0.944
		Median	1.000	1.000	1.000	1.000	1.000	1.000
		Q1	1.000	1.000	1.000	1.000	1.000	1.000
0.5	1	Mean	0.889	0.833	0.889	0.889	0.889	0.889
		Median	1.000	1.000	1.000	1.000	1.000	1.000
		Q1	1.000	1.000	1.000	1.000	1.000	1.000
0.5	2	Mean	0.889	0.833	0.889	0.889	0.889	0.889
		Median	1.000	1.000	1.000	1.000	1.000	1.000
		Q1	1.000	1.000	1.000	1.000	1.000	1.000
0.25	0.5	Mean	0.889	0.833	0.889	0.889	0.889	0.889
		Median	1.000	1.000	1.000	1.000	1.000	1.000
		Q1	1.000	1.000	1.000	1.000	1.000	1.000

Source: *Authors' calculations.*

5. Conclusions and Further Research

The results of the numerical experiments confirm the high effectiveness of the proposed

method for imputing missing data in surveys. A skillful, novel combination of predictive methods and fuzzy techniques has produced a very effective method. It seems that such a tool will fundamentally improve the quality of research in economic and management sciences based largely on surveys.

It appears that further research should provide a rationale for developing effective methods for selecting appropriate explanatory variables. In addition, it seems reasonable to develop methods for aggregating the results returned by different prediction methods. For the time being, the choice of prediction method and degree of fuzziness is up to the author of the study, based on his or her knowledge.

References:

Berthold, M.R. 2003. Mixed fuzzy rule formation. Int. J. Approx. Reason., 32(2-3), 67-84.

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kotter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B. 2007. KNIME: The Konstanz Information Miner. In: Studies in Classification, Data Analysis, and Knowledge Organization (GfKL2007), Springer.

Coppersmith, D., Hong, S.J., Hosking, J.R. 1999. Partitioning nominal attributes in decision trees. Data Min. Knowl. Discov., 3(2), pp. 197-217.

Domański, H, Sawiński, Z., Slomczynski, K. 2009. Sociological tools measuring occupations: New classification and scales. Warsaw: IFiS PAN.

Duncan, O.D. 1961. A Socio-economic Index for all Occupations. In: Occupations and Social Status, A.J. Reiss, Ed. New York: Free Press, pp. 116-117.

Falge, E., et al. 2001. Gap filling strategies for defensible annual sums of net ecosystem exchange. Agricultural and forest meteorology, 107.1, 43-69.

Fayolle, A., Landstrom, H., Gartner, W., Berglund, K. 2016. The institutionalization of entrepreneurship: Questioning the status quo and re-gaining hope for entrepreneurship research. Entrepreneurship and Regional Development, vol. 28, no. 7–8, pp. 477-486. doi: 10.1080/08985626.2016.1221227.

Friedman, J.H. 2002. Stochastic gradient boosting. Comput. Stat. Data Anal., 38(4), 367-378.

Ganzeboom, H.B.G., De Graaf, P.M., Treiman, D.J. 1992. A standard international socio-economic index of occupational status. Social Science Research, vol. 21, no. 1, pp. 1-56. doi: 10.1016/0049-089X(92)90017-B.

Goldthorpe, J.H., Hope, K. 1974. The Social Grading of Occupations: A New Approach and Scale. Clarendon Press.

Griffith, D., Cavusgil, S., Xu, S. 2008. Emerging themes in international business research. Journal of International Business Studies, vol. 39, no. 7, pp. 1220-1235. doi: 10.1057/palgrave.jibs.8400412.

Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K. 2001. Improvements to Platt's SMO algorithm for SVM classifier design. Neural computation, 13(3), pp. 637-649.

Kiersztyn, A., Kiersztyn, K. 2022. The Impact of Data Preprocessing on Prediction Effectiveness. In: Artificial Intelligence and Soft Computing: 21st International Conference, ICAISC 2022, Zakopane, Poland, June 19-23, Proceedings, Part I, Cham: Springer International Publishing, pp. 353-362.

Kiersztyn, A., Karczmarek, P., Łopucki R., Pedrycz W., Al. E., Kitowski I., Zbyryt A. 2020. Data imputation in related time series using fuzzy set-based techniques. In: 2020 IEEE international conference on fuzzy systems (FUZZ-IEEE), IEEE 2020, pp. 1-8.

Łopucki, R., Kiersztyn, A., Pitucha, G., Kitowski I. 2022. Handling missing data in ecological studies: Ignoring gaps in the dataset can distort the inference. Ecological Modelling, 468, 109964.

Pal, M. 2005. Random forest classifier for remote sensing classification. Int. J. Remote Sens., 26(1), pp. 217-222.

Platt, J.C. 1999. Fast training of support vector machines using sequential minimal optimization, advances in kernel methods. Support vector learning, pp. 185-208.

POLPAN Documentation, Polskie Badanie Panelowe POLPAN, n.a. http://polpan.org/en/data-and-documentation/.

Raja, P.S., Sasirekha K., Thangavel K. 2020. A Novel Fuzzy Rough Clustering Parameter-based missing value imputation. Neural Computing and Applications, vol. 32, no. 14, pp. 10033-10050. doi: 10.1007/s00521-019-04535-9.

Scherer, S. 2004. Stepping-stones or traps? The consequences of labour market entry positions on future careers in West Germany, Great Britain and Italy. Work Employment and Society, vol. 18, no. 2, pp. 369-394. doi: 10.1177/09500172004042774.

Shafer, J.C., Agrawal, R., Mehta, M. 1996. A scalable parallel classifier for data mining. In: Proceedings of 22th International Conference on Very Large Data Bases, vol. 96, pp. 544-555.

Xi, J., Kraus, S., Filser, M., Kellermanns F. 2015. Mapping the field of family business research: past trends and future directions. International Entrepreneurship and Management Journal, vol. 11, no. 1, pp. 113-132. doi: 10.1007/s11365-013-0286-z.