ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS

Volume 62 159 Number 6, 2014

http://dx.doi.org/10.11118/actaun201462061527

MISSING CATEGORICAL DATA IMPUTATION AND INDIVIDUAL OBSERVATION LEVEL IMPUTATION

Pavel Zimmermann[1], Petr Mazouch[1], Klára Hulíková Tesárková[2]

[1] Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic

[2] Department of Demography and Geodemography, Faculty of Science, Charles University in Prague, Albertov 6, 128 00 Prague 2, Czech Republic

Abstract

ZIMMERMANN PAVEL, MAZOUCH PETR, HULÍKOVÁ TESÁRKOVÁ KLÁRA. 2014. Missing Categorical Data Imputation and Individual Observation Level Imputation. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis,* 62(6): 1527–1534.

Traditional missing data techniques of imputation schemes focus on prediction of the missing value based on other observed values. In the case of continuous missing data the imputation of missing values often focuses on regression models. In the case of categorical data, usual techniques are then focused on classification techniques which sets the missing value to the 'most likely' category. This however leads to overrepresentation of the categories which are in general observed more often and hence can lead to biased results in many tasks especially in the case of presence of dominant categories. We present original methodology of imputation of missing values which results in the most likely structure (distribution) of the missing data conditional on the observed values. The methodology is based on the assumption that the categorical variable containing the missing values has multinomial distribution. Values of the parameters of this distribution are than estimated using the multinomial logistic regression. Illustrative example of missing value and its reconstruction of the highest education level of persons in some population is described.

Keywords: missing data, categorical data, multinomial regression

1 INTRODUCTION

Popular methods for a completion of (individual) observation as for example mean imputation, regression imputation or maximal likelihood imputation are usually focused on imputation of a continuous variable. Those methods mostly classify the missing values as "most likely" or "expected" values. Overview of those methods can be found for example in Schafer, Graham, 2002. List of methods for imputation of categorical variable is less extensive. In the case of categorical data, usual techniques are then focused on classification techniques which sets the missing value to the 'most likely' category (see Sentas *et al.*, 2004). This however leads to overrepresentation of the categories which are in general observed more often and hence can lead to biased results in many tasks especially in the case of presence of dominant categories.

The aim of the paper is to introduce multinomial logistic regression as very effective tool for missing data imputation. Motives for using this technique could be described by the following three requirements:

- to impute data set in form which can be re-used for variety of different analysis and applications; this means single imputation is required,
- to impute data in the most detailed level; optimally on individual observation level,
- to impute data in a way that will respect "expected" ratios of categories in general.

In the following text the methodology and its specific features will be described.

2 MISSING DATA TYPOLOGY

In this article the widely renowned typology of missing data structures developed in Rubin, 1976 will be adopted. Rubin considered the missingness as a probabilistic phenomenon, i.e. a set of random indicator variables R indicating non-missingness of a particular observation was considered. Also the partition of the complete dataset Y_{com} into set of observed values Y_{obs} and set of missing values Y_{mis}, i.e.

$$Y_{com} = (Y_{obs}, Y_{mis}),$$

was considerd. Missing data are called missing at random (MAR) in the case where the distribution of the missingness does not depend on Y_{mis}, i.e. when

$$P(R/Y_{mis}) = P(R/Y_{obs}).$$

This is the case where 'MAR allows the probabilities of missingness to depend on observed data but not on missing data'. A special case of the MAR is then MCAR (missing completely at random), where the probabilities of missingness do not depend on the observed data either:

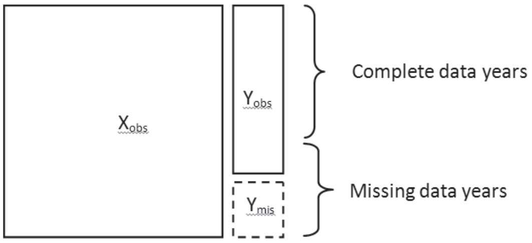
$$P(R/Y_{com}) = P(R).$$

If MAR is violated, data are missing not at random (MNAR).

The Task Solved Within the Paper

In this article a methodology for a specific task is developed which can be however reused in many similar tasks.

We assume a univariate pattern of categorical data, i.e. data where several variables are completely observed (X_{obs}) and one variable contains missing values. This was schematically expressed as in Schafer, Graham, 2002.



1: *Data set structure*

More precisely, we assume n_t complete observations of p categorical variables observed over T time periods denoted as $X_{i,t,j}$, $i = 1, \ldots, n_t$, $t = 1, \ldots, T$, $j = 1, \ldots, p$. Observations of $X_{i,t,j}$ are complete for all years. Furthermore we observe another categorical variable $Y_{i,t}$, $i = 1, \ldots, n_t$, $t = 1, \ldots, T$. For the years $t = 1, \ldots, c < T$ observations of $Y_{i,t}$ are complete (or with just negligible amount of missing data). These years will be referred as 'complete data years'. For the outstanding years $t = c + 1, \ldots, T$ the amount of missing data for $Y_{i,t}$ is rather large and MAR is not guaranteed. These years will be referred as 'missing data years'.

We assume that c is 'sufficiently large' to reasonably extrapolate trends for the missing data years and we assume that the trends observed during the complete data years are relevant for the predictions for the missing data years. Observations of $Y_{i,t}$ are assumed conditionally independent conditioning on $X_{i,t,j.}$

The Time Structure of Data Set According to Missing Data

From the time point of view three types of missing data position could be distinguished. The first is situation where we have complete information from some moment (year) but before this time missing data occur. In such a situation the aim is to reconstruct data before some point in time.

The second example is situation where data are complete, however, from some moment in time some (or all) data are missing. The aim in such a situation is to estimate the missing information for that period after any concrete moment. The third type is a situation where we have missing data "in the middle" of the time period, i.e. for some (limited) period of time the information is partially or completely missing. The aim is to bridge this part, estimate the missing data respecting trends before and after this missing period.

3 THE IMPUTATION ALGORITHM

In the following text, we will mean by **determinants** the original or rediscretized variables that have a significant impact on the distribution of the variable containing missing data. (The significance is measured over the years with complete data.) By **profile** we then mean a group of data with the same combination of values of the determinants. The set of determinants will be denoted as X. The matrix containing observations of determinants up to the time t will be denoted as X_t.

The basic steps of our imputation algorithm:

1. Based on the data observed in complete data years $X_{i,t,j}$, $i = 1, \ldots, n_t$, $t = 1, \ldots, c$, $j = 1, \ldots, p$ and Y_i, $i = 1, \ldots, n_t$, $t = 1, \ldots, c$, find determinants X of the missing data structure.
2. Define profiles of observations with missing data based on the values of the observed determinants X_t. The conditional distribution of Y is different conditioning on different profiles.
3. Estimate probabilities of each category q, $q = 1, \ldots, k$ of the missing variable Y for each profile, i.e. estimate $P(Y = q/X)$.
4. Based on the probabilities, find "appropriate" count of missing observations of each category in each profile and distribute these counts to each individual in the profile.

Multinomial Logistic Regression Application

Based on the assumption of conditional independence (independence of the observations the within profile) the categorical variable containing the missing values (Y) follows for a given profile the multinomial distribution. This fact immediately suggests using the multinomial logistic regression on the complete data years ($t \leq c$) as the methodology for finding the determinants (X) of the distribution (structure) of Y (as the response variable) and predicting the conditionally expected probabilities of each category of the response variable for each profile of data at each time point (for both $t \leq c$ and $t > c$). This requires assessing the time variable as covariate and assuming some (possibly polynomial) trend. That is the probability distribution of the categories of Y for each profile in each year $P(Y_t = q|X, t)$ is fitted as the outcome of the regression analysis (steps 1–3 of the above outlined imputation algorithm).

3.2 Multinomial Logistic Model

The multinomial regression method is a generalization of the logistic regression to multiclass problems. It is assumed that the response variable is a categorical variable with k possible outcomes. One of the k categories are selected as the 'reference' category. For every other category q, a regression equation is assumed in the model to describe the logarithmic odds of the category q to the reference category, i.e. equations

$$\log\left(\frac{p_q}{p_0}\right) = \beta_{q,0} + \beta_{q,1}x_1 + \beta_{q,2}x_2 + \ldots$$

are assumed, where p_q is the probability of the outcome q of the response variable, p_0 is the probability of the reference category, $\beta_{q,j}$ are the parameters and x_j are the regressors. Parameters $\beta_{q,j}$ are fitted using the maximum likelihood method. (See Hosmer, Lemeshow, 2004 for details.) Probabilities p_q may then be calculated using the parameter estimates using the formulas

$$p_q = \frac{\exp(\beta_{q,1}x_1 + \beta_{q,2}x_2 + \ldots)}{1 + \exp(\beta_{q,1}x_1 + \beta_{q,2}x_2 + \ldots)}$$

and

$$p_0 = \frac{1}{1 + \exp(\beta_{q,1}x_1 + \beta_{q,2}x_2 + \ldots)}.$$

3.3 Partially Missing Data

Based on the above described analysis we obtain the predicted distribution of the variable containing the missing data (Y) also for the years containing missing data ($t > c$) for each profile and each time point (conditioning on X and t will be leaved out in the notation of this section for simplicity). However, for these years we may have some amount of observed data (supposing partially missing data in the data set). Therefore we can estimate two distributions of missing values, first based on complete data years and second based on missing data years:

1. First prediction of the distribution $P(Y = q)$ for each category $q = 1, \ldots, k$ and a given profile and each time point as the prediction based on the complete data years, i.e. $X_t, Y_t\, t < c$.
2. Second distribution $P(Y = q|R = 0)$ fitted based on the observed data in the missing data years $X_t, Y_t, t > c$, i.e. distribution conditional on the fact that an observation is not missing.

Besides these distributions, we can also estimate the probability of missing values ($P(R = 0)$). The (marginal) distribution $P(Y = q)$ equals

$$P(Y = q) = P(Y = q, R = 0) + P(Y = q, R = 1),$$

where $P(Y = q, R = 0)$ (or $P(Y = q, R = 1)$) is the (joint) probability that the observation is certain category and is missing (or is not missing respectively) which equals

$$P(Y = q, R = 0) = P(Y = q|R = 0)\, P(R = 0)$$

and

$$P(Y = q, R = 1) = P(Y = q|R = 1)\, P(R=1).$$

We can write for the distribution of the observations that are missing (i.e. for which we already know that $R = 0$) as:

$$P(Y = q|R = 0) =$$

$$= [P(Y = q) - P(Y = q|R = 1)\, P(R = 1)]/\, P(R = 0).$$

Furthermore the estimates of the differences between the distributions $P(Y = i)$ and $P(Y = i|R = 1)$ may suggest the (non)randomness in missingness 'mechanism'.

3.4 Finding the Appropriate Count of Missing Observations of Each Category

Let us assume one particular profile of the data in a given year. Based on the above described regression analysis we can get the estimated distribution of the categories of Y for the missing observations, denoted as $P(Y = q|R = 0) = p_q, q = 1, \ldots, k$ for this given profile and year. Furthermore we know that in this profile and year, there is certain amount of missing data n. The distribution of the missing observations is (under the assumption of conditional independence) multinomial with the given parameter vector $p = (p_1, p_2, \ldots, p_k)$ and n. Given the probability distribution of the categories of the response variable and the number of missing observations we still need to determine how many of the missing observations correspond with each category (step 4 of the above outlined imputation algorithm). This variable will be denoted as U_q, $q = 1, \ldots, k$. Note that it is required that the missing values are imputed on the individual level and hence we need to determine counts (integers) of missing observations for each category.

Normally the expected value would be the first choice for the predictions as it yields predictions with the lowest least square error. The expected value of the multinomial distribution in a particular category *q* is simply the count of missing observations (in the particular profile in the particular year) times its probability, i.e.

$$E(U_q) = n\, p_q,\ q = 1, \ldots, k.$$

However, the expected values are generally real numbers (not necessarily integers) and hence do not allow for imputation on individual observation level. Therefore we suggest using the maximum likelihood criterion where the maximization is performed only on the discrete (integer) numbers. This means finding such u_q, $q = 1, \ldots, k$ that the joint distribution $P(u_1,u_2,\ldots,u_k\ |p, n)$ is maximized, i.e. we are looking for a vector $u = (u_1, u_2, \ldots, u_k)$ for which

$$\arg\max_u P(u; p, n)$$

where P denotes the probability function of the multinomial distribution. This in fact means we are looking for the mode of the multinomial distribution.

Mode of the Multinomial Distribution

There is no closed form formula for the mode of the multinomial distribution. There are however several iterative algorithms developed for this task. See for example Lloyd *et al.*, 1997, Finucan, 1964 or Le Gall, 2003. In our computations we selected the **Finucan's algorithm** published in Finucan, 1964.

Distribution of Estimated Data on the Individual Level

Having found the mode of the multinomial distribution for a particular profile we have a vector of counts (integers) of missing values of each category of the variable of the concern which has the highest probability. Within the profile, these counts may be 'assigned' randomly to the individuals as all individuals of the given profile have the same probability vector p_i, $i = 1, \ldots, k$ of being in *i*-th category.

4 APPLICABILITY OF THE ALGORITHM

The proposed method of estimation of missing data could be used in many spheres of application. In this paper we demonstrated the algorithm on (completely or partially unknown) education structure of a population. Education attainment could be taken as a typical example of categorical data. Moreover, when studying the population, this type of data is relatively often incomplete. Other example could be e.g. the marital status, age profile, etc.

The described algorithm is based on the assumption of continuous trend in the data within the missing data years. It corresponds with situation where data are missing because of some administrative changes etc. which does not affect the trend in the data. Application of the described method in situations where this condition is not fulfilled (e.g. where the missingness of the data is at least partially related to some changes affecting also the long-term trend – wars, etc.) would mean some sort of extrapolation of "unaffected" development – how the structure (partially or completely missing) would have developed if there had not been any interruption of the trend.

5 PRACTICAL APPLICATION

5.1 Dataset

The following variables are available within our dataset: Education, Sex, Marital status, Diagnoses of death, Age and Year of death. The dataset contains individual deaths in the Czech Republic 1995–2011. As a practical application we assumed educational attainment of a studied population as the variable containing missing data (*Y*).

The education is perceived as an important proxy for social status and behavioural habits and therefore it is a factor driving mortality. The education was collected obligatory until 2009 only and almost 40% of cases are missing in the year 2010 and 2011 and the other 60% are unreliable. Due to this fact it is necessary to impute the education conditioning on the other observed variables (i.e. using the information contained in the other variables *X*) and some regression model is necessary to forecast the probability of a death being in a given educational category given the other observed values in the year 2010 and 2011.

5.2 The Multinomial Model

Fitted Model

If the imputation algorithm described above is applied, it is first necessary to fit the conditional probabilities of each educational level using the multinomial logistic regression. The second educational category (low education) was selected as the reference category. Results are interpreted in relation to the reference category and log odds of the given category to the reference category are modeled. In the case of 4 education levels 3 equations are estimated: Basic vs Low, Middle vs Low, High vs Low. The determinants identified are displayed in the following table.Besides the main effects, interaction of the age and sex, marital status and sex were used and also some other interactions were identified and consequently reduced into indicators of sex and cancer, and sex and year of death < 2003. Based on the likelihood ratio test, all these effects were statistically significant (p-value < 0.0001). The profiles then consists

I:

Variable	Nr. levels	Levels
Sex	2	Male/Female
Marital status	4	Single/Divorced/Widower/Married
Age	4	0–16/16–39/40–59/80–110
Cause of death	3	Cancer/Circulatory/Other

of combinations of the levels of the above listed determinants.

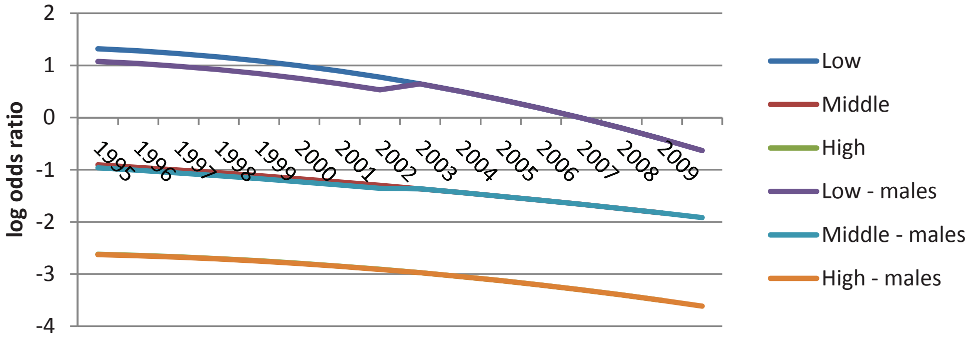
Further Interesting Relations Observed

We present some of the results observed that are in particular interesting. In order to be able to extrapolate the trend into unobserved years 2010 and 2011, we need to use a parametric function for the effect of the year of death. In this case second order polynomial was particularly suitable especially with an extra effect of the male gender until 2002. The trend curves are displayed in the Fig. 2.

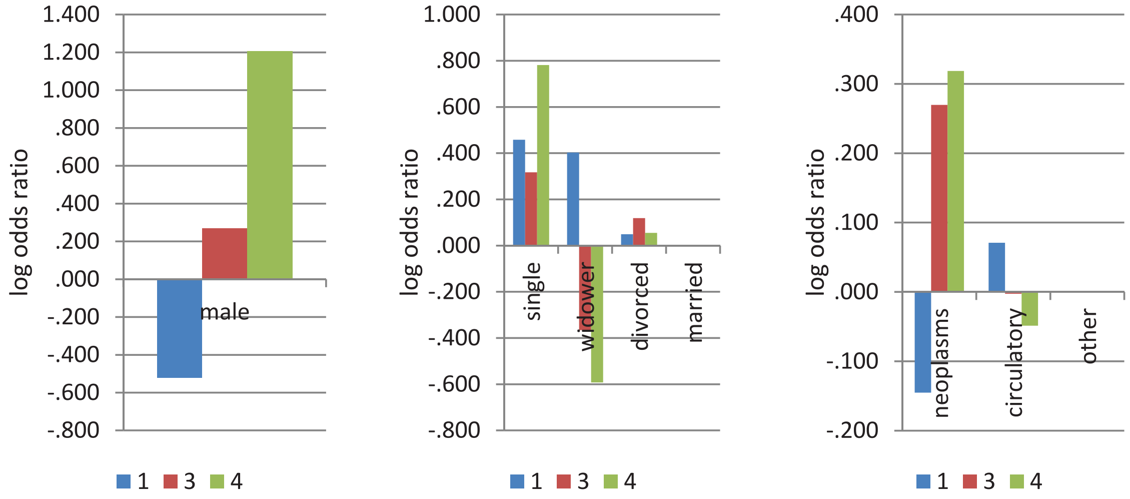
Main effect of the male gender reduces the log odds ratio of having basic education (relatively to having low education) and increases the log odds ratio of having middle or high education. Results are in Fig. 3a.

Main effect of being single increases relatively the chances of having basic, middle or high education. Main effect of being a widower increases the chances of having basic education and decreases the chances of having middle or high education (Fig. 3b).

The interaction of the male gender with the single status was in particular significant. We can interpret the result that single status increases the log odds ratio of having basic education, especially for males. Single status increases the log odds ratio (to low education) of females of having middle or high



2: *Trend curves of the effect of the year of death*



3a 3b 3c

3: *Main effect of gender (3a), marital status (3b) and cause of death (3c) reducing/increasing the log odds ratios of having particular education level (1 – Low education, 3 – Middle education, 4 – High education).*

education and the interaction mitigates the single effect for males. Interactions are described in appendix (App. 1).

Neoplasms diagnose generally decreases the odds of having basic education and it strongly increases the odds of having middle or high education (Fig. 3c).

Circulation diseases strongly increase the odds of having basic education and decrease the odds of having high education. Difference between genders was not observed.

There are significant differences between the impact of the cancer on education for different genders (interaction gender and cause): The impact of cancer on the decrease of the odds of having basic education and the increase of the odds of having middle or high education is much stronger for females than for males. So the neoplasm cause of death is determining the education much more for females than for males (see App. 2).

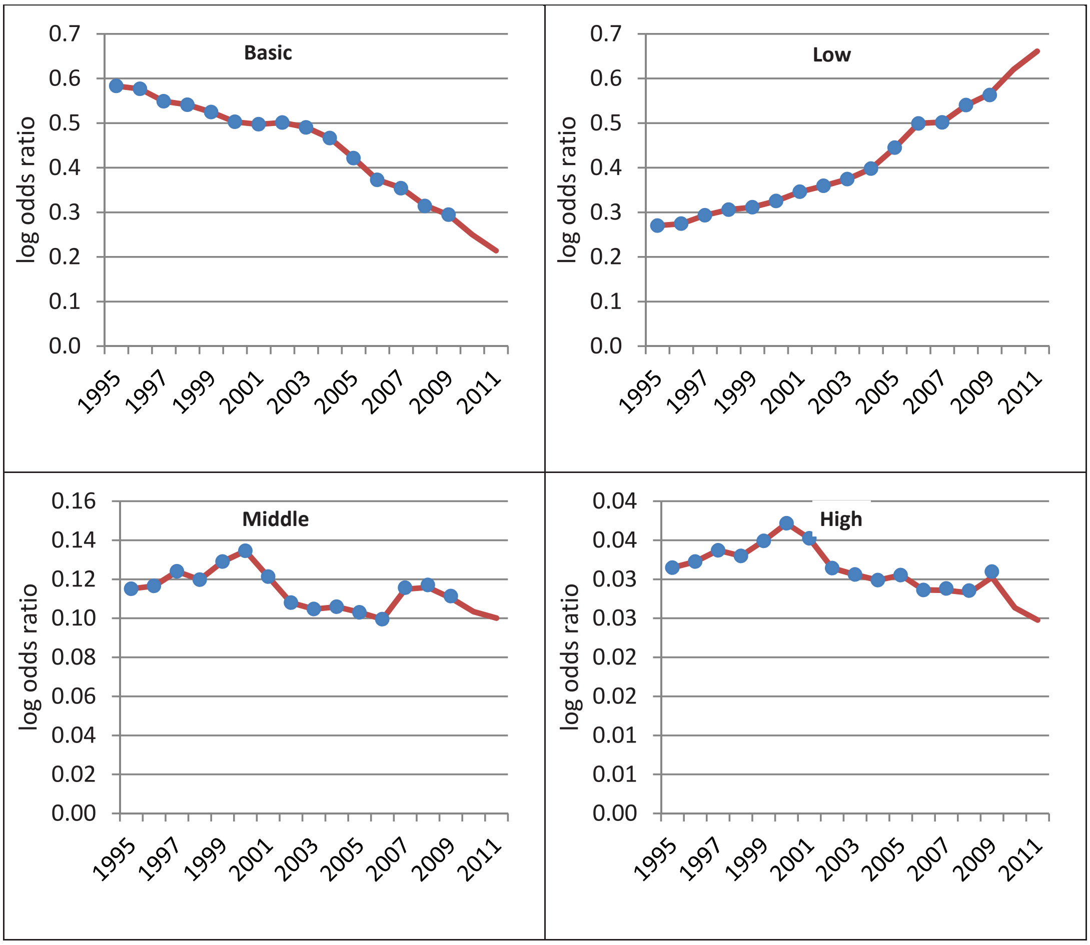
5.3. Predicted Probabilities and Imputation

Based on this model, it is possible to determine the conditional probability distribution of the educational levels for each combination of the values of the regressors (for each profile). These probabilities together with the number of missing observations for each profile specify the multinomial distribution. The mode is searched for each profile using the Finucan's algorithm. These modes are then the number of imputed observations for each educational level in each profile. The resulting imputation is for each education level displayed in comparison with the observed sample in Fig. 4.

6 CONCLUSION

Aim of this paper was to introduce multinomial logistic regression as very effective tool to missing data imputation. To the authors' knowledge the combination of the multinomial regression and mode searching algorithm was used for the first time for the missing data imputation task. The outcome of the proposed algorithm follows expected structure of the variable containing the missing values.

As a by-product the outcomes of the intermediate steps of the algorithm may be used for further analyses such as analyses of the dependencies (determinants) of the variable of our concern, or analysis of the missingnes mechanism.



4: *Distribution of deaths by education level, points are empirical values, lines are modeled values with prediction 2010 and 2011*

Future steps in the research will be to proof this method in some other practical situation. Demographic data (with incomplete information about the education attainment occurring in the latest years of the involved time period – as in the Fig. 3b) were used for the very first verification of the model and first results seem to be acceptable. Next part of the research will be to find more datasets with missing data, both MAR and MNAR and with different structure of missing data from the time point of view (length of missing, time of missing) and to prepare more detailed analysis of complemented data files.

SUMMARY

Paper presents original methodology of imputation of missing values which results in the most likely structure (distribution) of the missing data conditional on the observed values. The aim of the paper is to introduce multinomial logistic regression as very effective tool for missing data imputation. Motives for using this technique could be described by the following three requirements:

1. to impute data set in form which can be re-used for variety of different analysis and applications; this means single imputation is required,
2. to impute data in the most detailed level; optimally on individual observation level,
3. to impute data in a way that will respect "expected" ratios of categories in general.

To the authors' knowledge the combination of the multinomial regression and mode searching algorithm was used for the first time for the missing data imputation task. The outcome of the proposed algorithm follows expected structure of the variable containing the missing values.

The methodology is based on the assumption that the categorical variable containing the missing values has multinomial distribution. Values of the parameters of this distribution are than estimated using the multinomial logistic regression. The multinomial regression method is a generalization of the logistic regression to multiclass problems.

As a by-product the outcomes of the intermediate steps of the algorithm may be used for further analyses such as analyses of the dependencies (determinants) of the variable of our concern, or analysis of the missingnes mechanism.

Demographic data (with incomplete information about the education attainment occurring in the latest years of the involved time period) were used for the very first verification of the model and first results seem to be acceptable.

Acknowledgement

The article was written with the support provided by the Grant Agency of the Czech Republic to the project No. P404/12/0883 "*Generační úmrtnostní tabulky České republiky: data, biometrické funkce a trendy*". Earlier version of this paper with some preliminary results was presented at the Mathematical Methods in Economics 2013 conference held in Jihlava. Authors would like to thank anonymous referees for their extremely valuable and extended suggestions on how to improve the paper. All errors and omissions remain the authors' own.

REFERENCES

FINUCAN, H. M. 1964. The Mode of a Multinomial Distribution. *Biometrika*, 51(3–4): 513–517.

HOSMER JR, D. W., LEMESHOW, S. 2004. *Applied logistic regression*. New York: John Wiley & Sons.

JOHNSON, L. J., KOTZ, S. and BALAKRISHNAN, N. 1997. *Discrete multivariate distributions*. Vol. 165. New York: Wiley.

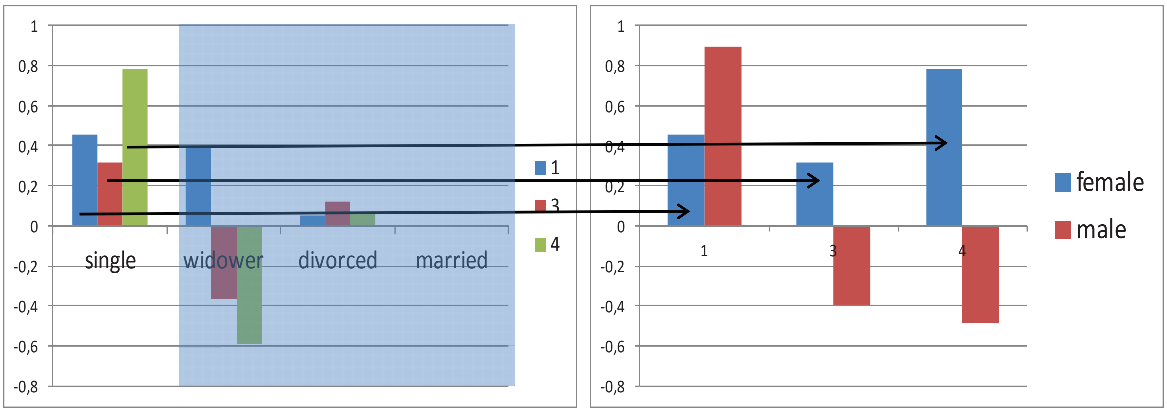
LE GALL, F. 2003. Determination of the modes of a Multinomial distribution. *Statistics & Probability Letters*, 62(4): 325–333.

RUBIN, D. B., 2002. Inference and missing data. *Biometrika*, 63(3): 581–592.

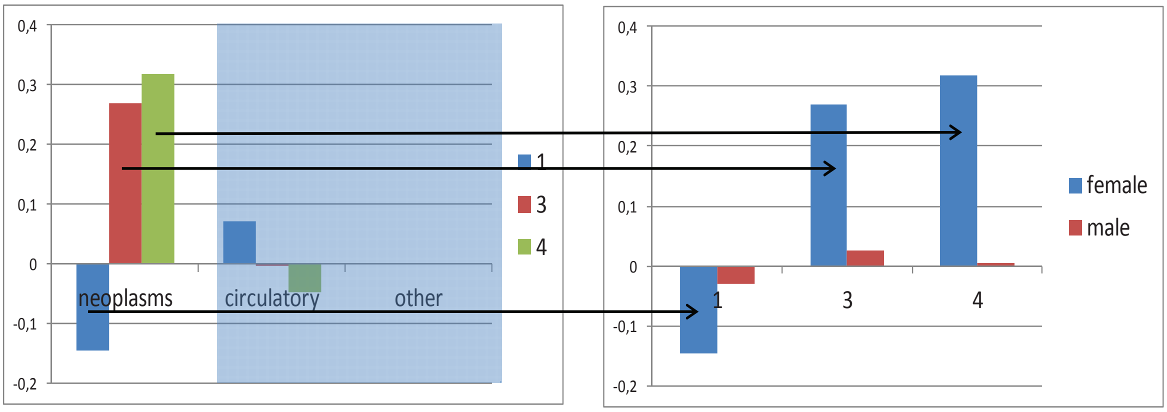
SCHAFER, J. L., GRAHAM, J. W. 2002. Missing data: our view of the state of the art. *Psychological methods*, 7(2): 147–177.

PANAGIOTIS, S., LEFTERIS, A., STAMELOS, I. 2004. Multiple logistic regression as imputation method applied on software effort prediction. In: *Proceedings of the 10th International Symposium on Software Metrics, 2004*. Chicago: IEEE Computer Society.

ZIMMERMANN, P., MAZOUCH, P., HULÍKOVÁ TESÁRKOVÁ, K. 2013. Categorical data imputation under MAR missing scheme. In: *Proceedings of the 31st International Conference Mathematical Methods in Economics, 2013*. Jihlava: College of Polytechnics Jihlava.



Appendix 1: The interaction of the male gender with the single status



Appendix 2: The interaction of the male gender with neoplasm cause of death

Contact information

Pavel Zimmermann: zimmerp@vse.cz
Petr Mazouch: mazouch@vse.cz
Klára Hulíková Tesárková: klara.tesarkova@gmail.com