MULTIPLE IMPUTATION AS A SOLUTION TO THE MISSING DATA PROBLEM IN SOCIAL SCIENCES

CLAUDIU D. TUFIŞ

In this paper I analyze a series of techniques designed for replacing missing data. From the extensive literature on political values in post-communist countries, I selected one of the most discussed models – the one proposed by Reisinger et al. (1994). In analyzing political values in Russia at the beginning of the transition, their model represents a significant contribution. The main disadvantage of the analyses of this model, however, is given by the substandard treatment of the missing data: listwise deletion. Since statistical theory suggests alternative techniques that offer unbiased estimators, in this paper I replicate the model using three different methods (mean imputation, regression-based imputation, and multiple imputation) to test the robustness of its findings. The results of this replication show that the initial findings are not robust and indicate the multiple imputation method as a solution for obtaining unbiased estimators in the presence of missing data.

***Key words**: missing data, multiple imputation, methodology, statistical software.*

MISSING DATA – THEORETICAL ASPECTS

Missing or incomplete data cause significant problems in the analysis of survey data. Despite the negative effects of missing data on the results of statistical analyses (e.g. biased estimators) social scientists rarely use newly developed techniques for dealing with missing data. Based on content analysis of three leading journals in political science (*American Political Science Review*, *American Journal of Political Science*, and *British Journal of Political Science*), King *et al*. (2001) estimated that approximately 94% of the articles published between 1993 and 1997 that used some form of survey analysis used listwise deletion, reducing their sample by one third on the average.

In any survey it is very likely that some of the respondents will refuse to participate in the survey. Although this may pose significant problems in terms of response rates and the representativeness of the sample, these respondents are not of interest within the scope of this paper (for an analysis of unit non-response in the Romanian context, see Comşa, 2002). Moreover, Brehm (1993) proved that unit non-response is usually not a significant source of bias in analysis in the social sciences.

Adresa de contact a autorului: **Claudiu Tufiş**, Institutul de Cercetare a Calităţii Vieţii, Calea 13 Septembrie, nr. 13, sector 5, 050711, Bucureşti, România; e-mail: ctufis@iccv.ro.

CALITATEA VIEŢII, XIX, nr. 1–2, 2008, p. 199–212

There are, however, other mechanisms that lead to non-responses in surveys (see Voicu, 1999, and Comşa, 2003, for interesting analyses of the factors that affect item non-response in Romanian surveys). Some of the respondents may fail to answer a number of items. Some may not like a question and refuse to answer it. Others may be anxious to finish the interview and thus refuse to answer the last questions. Once the survey is done, the reasons for not answering are irrelevant however, because they all have the same effect – they generate missing data in the dataset and the analyst has to deal with this problem. Item non-response is thus the main cause of missing data. There are other possible reasons why missing data may appear (including interviewer and coder error), but carefully designing the instrument and controlling its application in the field may eliminate these alternative causes.

The problem of missing data is rather simple: since some of the respondents did not answer all items in the questionnaire, there are no records for particular respondent – question combinations. For any statistical analysis that contains a variable for which there are missing data, the cases with missing data have to be excluded from analysis, if the data are not imputed. This decision is associated with a series of negative outcomes: the sample size is reduced, the representativeness of the sample decreases, and the information offered by the respondents by answering other items is lost. Different solutions have been offered for the problem of missing data, and I discuss the most important in this paper. Since these solutions depend on different conceptualizations of the relationship between respondents with complete data and respondents with missing data, I focus next on the main assumptions encountered in the treatment of missing data (see Rubin 1976).

Data are considered to be missing completely at random (MCAR) if the probability of missing data on a variable is independent of both the values of that variable and of the values of the other variables in the dataset. This is a strong assumption that is usually not met in survey data. In the rare cases where the missing data are MCAR, ”the set of individuals with complete data can be regarded as a simple random subsample from the original set of observations” (Allison, 2002, 3). If the assumption is true, then one could perform the analyses on the subset of cases with complete information without having to worry about obtaining biased estimators.

The missing at random assumption (MAR) considers that the probability of missing data on a certain variable is independent on the values of that variable, once the effects of the remaining variables in the dataset are taken into consideration. The main consequence of this assumption is that the information available in other variables in the dataset could be used for imputing the missing data.

Finally, missing data are considered to be non-ignorable (NI) if the probability of missing data on a certain variable is dependent on the values the variable is taking. Possible examples of non-ignorable missing data include income

(it is possible that the higher the income of a person the higher the probability of refusing to report the income) or certain values or beliefs (in which case, the more extremist the respondent's belief, the higher the probability of missing data).

As these assumptions indicate, NI missing data require special models for the estimation of missing values, MCAR missing data can be excluded from analysis, while for MAR missing data, different imputation models could be used. I discuss next the most important solutions for dealing with missing data, under the MAR assumption.

Listwise deletion. The easiest and simplest solution to the problem of missing data is to assume that, by excluding the cases with missing data from analysis, the problem is solved. As previous studies indicate, this assumption is almost always incorrect. Mackelprang shows that "distortion can occur with as little as two percent missing data […] Five percent missing data produced distortion in the simulated data set which clearly exceeds the acceptable limits for most social science research" (Mackelprang, 1970, 501). The assumption holds true only when the cases with complete information represent a random subsample of the original sample (Little and Rubin, 1987). This, however, is a rare occurrence in social sciences. If the missing data are not MCAR, the most likely outcome of using listwise deletion as a solution to the missing data problem is that the parameter estimates will be biased. King *et al*., indicate that "the point estimate in the average political science article is about one standard error farther away from the truth because of listwise deletion" (King *et al*., 2001, 52). A similar approach is to create a new variable indicating for each respondent whether data are missing or not, and to use this variable in analysis. It has been argued, however, that "the missing-indicator methods show unacceptably large biases in practical situations and are not advisable in general" (Jones, 1996, 222).

Mean imputation. An alternative easy solution is to replace the missing values with the means of the corresponding variables. While it may seem an appealing way to solve the problem, mean imputation dramatically reduces the variance of the variables with missing data, because it uses the same value for all cases with missing data. Moreover, it is problematic to use mean imputation with variables measured at the nominal or even at the ordinal level (e.g., it is not very helpful to replace the missing data for a dichotomous variable like gender with the average for the variable, since any values other than 0 or 1 do not have any meaning). When using mean imputation, "inferences (tests and confidence intervals) are seriously distorted by bias and overstated precision […] Unconditional mean imputation cannot be generally recommended" (Little, 1992, 1231).

Similar response pattern imputation. This is also known as a hot-deck imputation. The missing data are imputed from a respondent with complete data (donor) that has similar answers on a set of variables with the respondent with missing data. While this method is somewhat better than listwise deletion or mean imputation, it still results in a single completed data set, which "may lead to inferences that are grossly in error" (Wang, Sedransk, and Jinn, 1992, 961).

Regression imputation. In this case, the missing data are imputed using a regression model. There are several variations of this method, including the use of a set of regression equations within sample strata, defined by variables not included in the regression equation (in this case, the procedure is called best-subset regression and it is actually a combination of regression and hot-deck imputation).

The four methods presented above represent the ”traditional” approaches to the missing data problem. While some perform better than others, all four are, in fact, nothing else but educated guesses about what would have been the respondent’s answer, if recorded. There is always an uncertainty in missing data imputation and by imputing only one value this uncertainty is artificially reduced to zero. As a result of eliminating the uncertainty from the model, the standard errors of the estimated coefficients are biased towards zero, making it easier to find significant relationships in the data. The next two solutions address the uncertainty issue directly. The distinction between the methods presented above and the two methods I discuss next could be understood as deterministic versus probabilistic missing data imputation.

Full information, maximum likelihood. The advantage of this method consists of the fact that the algorithm makes use of all the information in the observed data, in the presence of an unlimited number of missing-data patterns. ”FIML assumes multivariate normality, and maximizes the likelihood of the model, given the observed data” (Wothke, 2000), and the FIML estimate ”includes information about the mean and variance of missing portions of a variable, given the observed portion(s) of other variables” (Wothke, 2000). The procedure computes unbiased parameter estimates when the data is missing at random, and it is known to produce good results when the missing pattern is ”somewhat nonignorable” (Arbuckle and Wothke, 1999, 333). There are no conventional limits establishing the acceptable amount of missing data with nonignorable patterns, but with randomly missing data, research has demonstrated that the FIML estimation procedure yields comparable regression estimates and standard errors in a sample with complete data and in a sample with 75% missing data on one variable (Arbuckle and Wothke, 1999, 349–358). With data missing at random or completely at random, FIML yields consistent and efficient estimates (Arbuckle and Wothke, 1999, 333).

The main disadvantage of this method is given by its inability to handle different models. While FIML can be used to estimate the most common models used in the social sciences (the linear and log-linear models), it cannot accommodate other models of interest (e.g., duration models, event history models, etc.). This problem is solved by using multiple imputation.

Multiple imputation. This solution is advocated, among others, by Rubin (1987, 1996), Schafer and Olsen (1998), and Allison (2002). The multiple imputation technique requires three different steps. In the first stage, m values are imputed for each missing values, resulting into the creation of m different data sets. Once the imputed datasets are created, these are used in data analysis (the second

step). The results of the analyses performed for each of the *m* datasets are then saved and used in the third phase, which requires the aggregation of the results using the formulas proposed by Rubin.

At the imputation stage, the uncertainty implied by missing data imputation is reflected in the imputation of more than one value for the missing data. It should be noted that, although the multiple imputation procedure assumes that the data are jointly multivariate normal, this assumption is very robust to departures from normality. Another advantage of this method is that the number of imputed datasets is relatively low: "the relative efficiency of estimators with *m* as low as 5 or 10 is nearly the same as with $m = \infty$, unless missingness is exceptionally high" (King *et al*., 2001, 56). The remaining two steps, while time consuming, are easy to implement and do not require more sophisticated analyses than those performed in a regular statistical analysis. It should be noted that there are several software packages devoted to multiple imputation; among them *Amelia II* (available at http://gking.harvard.edu/amelia/), *IveWare* (requires SAS, available at http://www.isr.umich.edu/src/smp/ive/), and *Norm* (available at http://www.stat.psu.edu/~jls/misoftwa.html). For reviews of some of these packages, see Horton and Lipsitz (2001) and Horton and Kleinman (2007).

I test in this paper a theoretical model used by Reisinger *et al*. in the study of political values in Russia, at the beginning of the post-communist transition. I use three different methods of data imputation (mean imputation, regression imputation, and multiple imputation) and then I compare the results obtained using these methods to the original results which used listwise deletion.

MISSING DATA – APPLICATION

Reisinger *et al*. (1994) study political values in Russia, Ukraine, and Lithuania at the beginning of the post-communist transition, testing three competing hypotheses about the source of political values in post-Soviet societies: political culture, regime indoctrination, and societal modernization. The data used in their analyses come from the *New Soviet Citizen Survey, 1992: Monitoring Political Change* (Miller, Reisinger, and Hesli, 1992). The sample sizes for the three countries are 1301 (Russia), 900 (Ukraine), and 500 (Lithuania).

Based on their analyses, Reisinger *et al*. conclude that, out of the three competing hypotheses, only the modernization theory is supported by the data. Their results for the modernization theory are not generally accepted, however. Finifter and Mickiewicz (1992), using data from 1989, obtain different results for education and gender. Which of these results are closer to the true relationships in the population? Since both studies use listwise deletion for dealing with missing data, it is difficult to offer an answer to this question.

Table 1 presents the percentages of missing data for each of the variables included in analysis. Most of the variables have missing data on more than 10% of the cases (from a low of 1.6% – TRUST in Lithuania – to a high of 28.6% – EI in

Lithuania). While this may not be a significant problem in univariate analyses, the percentages add up in multivariate analyses, resulting in significant reductions of the sample sizes (this is a problem especially for Lithuania where the initial sample size is 500). Two variables seem to be especially problematic: the index of economic indoctrination (approximately 25% of the respondents refused to answer this question in all three countries) and the index of democratic values (around 20% missing data). Overall, the missingness rates presented in *Table 1* paint a fairly common picture for attitudinal surveys.

Table 1

Missing data in the dataset used by Reisinger *et al.* (1994) – percentages

Variable	Russia (N=1301)	Ukraine (N=900)	Lithuania (N=500)
Index of Desire for Strong Leadership (DSL)	13.5	11.1	11.6
Index of Desire for Order (DO)	15.4	12.6	11.8
View of Stalin (STALIN)	17.9	13.6	20.2
Index of Economic Indoctrination (EI)	25.1	23.6	28.6
Interpersonal Trust (TRUST)	3.0	2.9	1.6
Party Competition (COMPETE)	10.5	9.8	6.0
Opposition to the Government (OPPOSE)	10.0	9.7	6.4
Postmaterial Values (PM)	6.1	6.2	9.2
Index of Rights Orientation (RO)	14.0	13.6	11.4
Index of Democratic Values (DV)	22.6	19.7	17.6

I retest in this paper the models proposed by Reisinger *et al*. using three methods of missing data imputation: mean imputation, regression imputation, and multiple imputation. Mean imputation was performed in SPSS v.11.0, replacing the missing data with the mean value of the variables within strata defined by the three countries. Regression imputation was performed using the impute command in Stata v.8.0. The impute command uses the best-subset regression technique, being thus a combination of hot-deck and regression imputation. This command takes into account the patterns of missing data for a more efficient estimation of the regression equations. Finally, the multiple imputation was performed using Norm v.2.03. I have imputed five different datasets, I performed the statistical analyses in SPSS, and then I aggregated the results using Norm again.

Table 2 presents the efficiency of the estimators for the variables included in analysis. In computing the efficiency of the estimators I used the formula proposed by Rubin (1987): $E = (1 + \gamma / m) ^ (-1)$, where E is the efficiency of the estimators, γ is the rate of missing data, and *m* is the number of imputations. As the results indicate, the efficiency of the estimators ranges between 0,95 (in the case of the index of economic indoctrination, which had a high proportion of missing data) and 1,00 (in the case of interpersonal trust, which had a very small proportion of

missing data). Overall, the results indicate that the estimators have high levels of efficiency, even in the case of a relatively small number of imputations (m = 5).

Table 2

Efficiency of estimators for m = 5 imputations

Variable	Russia (N=1301)	Ukraine (N=900)	Lithuania (N=500)
Index of Desire for Strong Leadership (DSL)	0.97	0.98	0.98
Index of Desire for Order (DO)	0.97	0.98	0.98
View of Stalin (STALIN)	0.97	0.97	0.96
Index of Economic Indoctrination (EI)	0.95	0.95	0.95
Interpersonal Trust (TRUST)	0.99	0.99	1.00
Party Competition (COMPETE)	0.98	0.98	0.99
Opposition to the Government (OPPOSE)	0.98	0.98	0.99
Postmaterial Values (PM)	0.99	0.99	0.98
Index of Rights Orientation (RO)	0.97	0.97	0.98
Index of Democratic Values (DV)	0.96	0.96	0.97

The first analysis presented by Reisinger *et al*. is the mean comparison between the three countries included in their sample. By comparing the means on different variables for the three countries, the authors test for the political culture and the regime indoctrination hypotheses. In Table 3, I present the original results and the results of my replications using different methods of dealing with missing data.

There are only two sign changes. The index of desire for strong leadership had a negative sign in the original model for the Russia – Ukraine pair, which changes into a positive sign in all the models with imputed data. The index of economic indoctrination had a positive sign in the original model for the Russia – Lithuania pair, which changes into a negative sign in the model using regression imputation. Given that all these comparisons are not significant, the sign changes are not a significant source for concern.

There are more differences between the original model and the imputed data models in terms of the significance associated with the t-test for the equality of means. This result was expected, given that listwise deletion usually has a more significant effect on the standard errors of the coefficients than on the coefficients themselves. In comparing the significance levels, there are two types of differences. The first type represents changes in the significance levels of the mean differences (e.g. the significant coefficients remain significant, but at different levels of significance). The second type is more important: variables that were significant in the original model lose their significance, while other variables may become significant. There are four such cases in my analysis.

Table 3

Mean comparisons

	Russia – Ukraine				Russia – Lithuania				Ukraine – Lithuania			
	Original	**Mean**	**RI**	**MI**	**Original**	**Mean**	**RI**	**MI**	**Original**	**Mean**	**RI**	**MI**
DSL	-0,010	**0,020**	**0,010**	**0,027**	0,060	0,090	0,070	0,092	0,070	0,070	0,060	0,065
	(0,669)	(0,370)	(0,576)	(0,377)	(0,131)	***(0,013)***	(0,053)	***(0,016)***	(0,078)	(0,086)	(0,151)	(0,120)
DO	-0,050	-0,050	-0,050	-0,067	0,040	0,080	0,050	0,059	0,090	0,130	0,100	0,126
	(0,258)	(0,222)	(0,251)	(0,157)	(0,437)	(0,105)	(0,293)	(0,274)	(0,096)	***(0,010)***	(0,054)	***(0,026)***
STALIN	0,160	0,160	0,140	0,134	-0,050	-0,050	-0,050	-0,047	-0,210	-0,210	-0,190	-0,181
	(0,005)	***(0,001)***	(0,004)	***(0,025)***	(0,563)	(0,463)	(0,439)	(0,489)	(0,004)	***(0,001)***	(0,002)	***(0,011)***
EI	-0,270	-0,270	-0,330	-0,263	0,270	0,270	**-0,030**	0,102	0,540	0,540	0,300	0,365
	(0,211)	(0,097)	(0,056)	(0,167)	(0,308)	(0,167)	(0,873)	(0,678)	(0,026)	***(0,002)***	***(0,152)***	***(0,205)***
TRUST	0,120	0,120	0,120	0,116	0,030	0,030	0,030	0,037	-0,090	-0,090	-0,090	-0,078
	(0,000)	(0,000)	(0,000)	(0,000)	(0,196)	(0,190)	(0,209)	(0,126)	(0,000)	(0,001)	(0,001)	***(0,002)***
COMPETE	0,130	0,130	0,120	0,140	-0,100	-0,100	-0,080	-0,070	-0,230	-0,230	-0,200	-0,210
	(0,006)	(0,002)	(0,007)	(0,007)	(0,119)	(0,088)	(0,136)	(0,240)	(0,000)	(0,000)	(0,000)	(0,001)
OPPOSE	0,130	0,130	0,130	0,108	0,300	0,300	0,300	0,270	0,170	0,170	0,170	0,162
	(0,008)	(0,003)	(0,004)	***(0,030)***	(0,000)	(0,000)	(0,000)	(0,000)	(0,003)	***(0,001)***	***(0,001)***	(0,005)
PM	0,040	0,030	0,040	0,037	-0,170	-0,260	-0,204	-0,190	-0,210	-0,290	-0,244	-0,228
	(0,106)	(0,180)	(0,088)	(0,124)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)
RO	-0,080	-0,060	-0,090	-0,089	-0,300	-0,230	-0,262	-0,261	-0,220	-0,170	-0,172	-0,172
	(0,054)	(0,089)	***(0,021)***	***(0,048)***	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,001)
DV	-0,110	-0,060	-0,124	-0,119	-0,380	-0,190	-0,290	-0,280	-0,270	-0,130	-0,166	-0,162
	(0,087)	(0,289)	(0,029)	(0,091)	(0,000)	(0,001)	(0,000)	(0,000)	(0,001)	***(0,029)***	***(0,012)***	***(0,023)***

Note: Entries in the table show the mean difference and the level of significance for the difference (in parentheses). **Bolded** figures indicate estimates that have a different sign than the estimates in the original model. ***Bolded and italicized*** figures indicate a level of significance that is different from the level of significance in the original model.

For the comparisons between Russia and Ukraine, the difference between the means of the index of rights orientation is not significant in the original model, but it becomes significant in the multiple imputation model, indicating that Ukrainian respondents have more respect for human rights than the Russian respondents. For the Russia – Lithuania pair, Lithuanians have a significantly lower score on the index of desire for strong leadership in the multiple imputation model, whereas in the initial model, the difference was not significant. The same result is observed for the index of desire for order in the Ukraine – Lithuania pair: in the multiple imputation model, the Lithuanians have a significantly lower score. In the case of the index of economic indoctrination, while Lithuanians seemed to be significantly different from the Ukrainians in the original model, once the missing data are imputed using multiple imputation, the difference loses its significance.

The results of this part of the analysis indicate that even for simple analyses, like means comparisons, different results are obtained using different methods of dealing with missing data. Since previous studies suggest that multiple imputation offers unbiased results in comparison with listwise deletion, some of the results reported by Reisinger *et al*. are incorrect.

In the second part of their analysis, Reisinger *et al*. report a series of regressions explaining attitudes toward strong leadership and order, political and

economic indoctrination, and democratic values. The results of my replications (as well as the original results) are presented in *Table 4* through *Table 8*. The regression analyses indicate significant differences between the original model and the models with imputed missing data.

In *Table 4*, which explains attitudes toward strong leadership, four out of the ten coefficients either become significant or lose their significance, if missing data are imputed. Age becomes significant in both Russia and Lithuania, while education loses its significance in Russia and Ukraine. Two more coefficients (urban in Russia and education in Lithuania), while remaining significant, change their significance levels. In the regression equation explaining attitudes towards order (*Table 5*) there is only one significant change: the coefficient for urban residency becomes significant, once the missing data are imputed. Three additional coefficients (education in Russia and age in Ukraine and Lithuania) change their significance levels, while remaining significant. The changes in the model explaining the respondents' views of Stalin (*Table 6*), there are no coefficients that change significance and there are only four coefficients (age and education in Russia and Ukraine) that show minor changes in their significance level. In *Table 7* (economic indoctrination), Lithuania also presents significant changes: the coefficients for education and Russian nationality, which were not significant in the original model, become significant in the models using imputed data. Finally, in the equation explaining democratic values there are three important changes in the case of Lithuania, and one in the case of Ukraine. In Lithuania, the coefficients for age, urban residency, and Russian nationality, although significant in the original model, are not significant anymore in any of the models that use missing data imputation. In Ukraine, the coefficient for attitudes towards order becomes significant when missing data are imputed.

By comparing all the regression models by country, it can be seen that, in the case of Russia, out of the 19 coefficients estimated in the five models, two change significantly (10%) and another seven change their significance levels (36%). In the case of Ukraine, two coefficients present significant changes (10%) while another eight change their significance levels (42%). Finally, in Lithuania, six out of the 24 estimated coefficients change significantly (25%) and an additional four coefficients change their significance levels (16%).

In discussing these results I have focused mainly on comparing the results of the original model to the results from the model using multiple imputation for solving the missing data problems. While some of the changes indicated by these comparisons are also captured by the models using mean imputation and regression imputation to replace the missing data, there are still changes that appear only in the multiple imputation model. Taking into account the problems associated with missing data replacement by mean imputation or by regression imputation, the use of multiple imputation is even more justified.

Table 1

Regression model for dependent variable "Desire for strong leadership"

	Russia				Ukraine				Lithuania			
	Model	**Mean**	**RI**	**MI**	**Model**	**Mean**	**RI**	**MI**	**Model**	**Mean**	**RI**	**MI**
Age	-0,002	**-0,003****	**-0,004*****	**-0,003***	-0,001	-0,002	**-0,003***	-0,003	-0,004	**-0,005***	**-0,006****	**-0,005***
		(0,001)	**(0,001)**	**(0,001)**		(0,001)	**(0,001)**	(0,002)		**(0,002)**	**(0,002)**	**(0,002)**
Education	0,047*	**0,016**	**0,017**	**0,020**	0,054*	**0,022**	**0,027**	**0,025**	0,103*	**0,050****	**0,056****	**0,057****
		(0,009)	**(0,009)**	**(0,010)**		**(0,013)**	**(0,014)**	**(0,014)**		**(0,019)**	**(0,020)**	**(0,021)**
Urban	0,318**	**0,089***	**0,193*****	**0,127***	0,359***	0,306***	0,329***	0,336***	0,016	**-0,165***	-0,101	-0,139
		(0,042)	**(0,045)**	**(0,053)**		(0,045)	(0,046)	(0,050)		**(0,069)**	(0,071)	(0,077)
Russian									-0,052	-0,065	-0,070	-0,081
										(0,093)	(0,096)	(0,103)
R^2	0,050	0,012	0,026	0,015	0,070	0,058	0,066	0,063	0,020	0,021	0,019	0,019

Note: *** p < 0,001, ** p < 0,01, * p < 0,05. **Bolded** entries represent coefficients whose significance level is different from the original model.

Table 2

Regression model for dependent variable "Desire for order"

	Russia				Ukraine				Lithuania			
	Model	**Mean**	**RI**	**MI**	**Model**	**Mean**	**RI**	**MI**	**Model**	**Mean**	**RI**	**MI**
Age	-0,019***	-0,017***	-0,021***	-0,020***	-0,007*	**-0,009*****	**-0,010*****	**-0,011*****	-0,014**	**-0,014*****	**-0,016*****	**-0,015*****
		(0,002)	(0,002)	(0,002)		**(0,002)**	**(0,002)**	**(0,002)**		**(0,003)**	**(0,003)**	**(0,003)**
Education	-0,050*	**-0,040****	**-0,059*****	**-0,050****	-0,004	-0,016	-0,019	-0,015	0,051	0,033	0,025	0,017
		(0,014)	**(0,014)**	**(0,016)**		(0,021)	(0,021)	(0,022)		(0,024)	(0,025)	(0,026)
Urban	-0,025	-0,048	-0,115	-0,068	0,083	0,101	0,082	0,092	-0,018	**-0,178***	**-0,216***	**-0,214***
		(0,066)	(0,069)	(0,078)		(0,068)	(0,071)	(0,074)		**(0,088)**	**(0,091)**	**(0,095)**
Russian									0,178	0,116	0,168	0,071
										(0,119)	(0,123)	(0,128)
R^2	0,090	0,090	0,122	0,109	0,010	0,025	0,028	0,031	0,050	0,058	0,076	0,066

Note: *** p < 0,001, ** p < 0,01, * p < 0,05. **Bolded** entries represent coefficients whose significance level is different from the original model.

Table 3

Regression model for dependent variable "View of Stalin"

	Russia				Ukraine				Lithuania			
	Model	**Mean**	**RI**	**MI**	**Model**	**Mean**	**RI**	**MI**	**Model**	**Mean**	**RI**	**MI**
Age	-0,018**	**-0,015***	**-0,018***	**-0,017***	-0,012**	**-0,011***	**-0,012***	**-0,012***	-0,009*	-0,007*	**-0,010****	-0,011*
		(0,002)	**(0,002)**	**(0,002)**		**(0,002)**	**(0,002)**	**(0,003)**		(0,003)	**(0,003)**	(0,005)
Education	-0,149**	**-0,089***	**-0,103***	**-0,096***	-0,073	-0,044	**-0,052***	**-0,054***	0,012	0,015	-0,007	-0,014
		(0,017)	**(0,017)**	**(0,018)**		(0,022)	**(0,023)**	**(0,027)**		(0,031)	(0,032)	(0,037)
Urban	-0,034	-0,038	-0,038	-0,085	-0,117	-0,097	-0,105	-0,120	-0,265	-0,182	**-0,246***	-0,229
		(0,079)	(0,080)	(0,091)		(0,074)	(0,076)	(0,095)		(0,114)	**(0,116)**	(0,150)
Russian									0,000	-0,003	0,048	0,039
										(0,154)	(0,156)	(0,190)
R^2	0,090	0,069	0,092	0,078	0,040	0,036	0,046	0,043	0,010	0,008	0,024	0,023

Note: *** p < 0,001, ** p < 0,01, * p < 0,05. **Bolded** entries represent coefficients whose significance level is different from the original model.

Table 4

Regression model for dependent variable "Economic indoctrination"

	Russia				Ukraine				Lithuania			
	Model	**Mean**	**RI**	**MI**	**Model**	**Mean**	**RI**	**MI**	**Model**	**Mean**	**RI**	**MI**
Age	-0,105***	-0,073***	-0,099***	-0,098***	-0,065**	**-0,047***	**-0,064***	**-0,065***	-0,064**	**-0,043***	**-0,071***	**-0,072***
		(0,007)	(0,007)	(0,009)		**(0,007)**	**(0,007)**	**(0,009)**		**(0,009)**	**(0,011)**	**(0,011)**
Education	-0,422**	**-0,194***	**-0,389***	**-0,335***	-0,349*	**-0,185****	**-0,343***	**-0,287****	-0,309	-0,113	**-0,292****	**-0,261***
		(0,055)	**(0,060)**	**(0,069)**		**(0,069)**	**(0,076)**	**(0,089)**		(0,083)	**(0,100)**	**(0,126)**
Urban	-0,965*	-0,631*	**-1,007***	**-0,909****	-0,899*	**-0,759***	**-0,917***	**-0,860****	-1,009*	**-0,574**	-0,936*	-1,046*
		(0,264)	**(0,288)**	**(0,324)**		**(0,228)**	**(0,252)**	**(0,300)**		**(0,304)**	(0,365)	(0,436)
Russian									1,110	**0,815***	**1,320****	0,805
										(0,410)	**(0,493)**	(0,521)
R^2	0,160	0,107	0,179	0,164	0,120	0,083	0,133	0,116	0,110	0,069	0,143	0,145

Note: *** p < 0,001, ** p < 0,01, * p < 0,05. **Bolded** entries represent coefficients whose significance level is different from the original model.

Table 5

Regression model for dependent variable "Democratic values"

	Russia				Ukraine				Lithuania			
	Model	**Mean**	**RI**	**MI**	**Model**	**Mean**	**RI**	**MI**	**Model**	**Mean**	**RI**	**MI**
Age	-0,001	-0,002	-0,004	-0,004	-0,004	-0,003	**-0,005***	-0,006	-0,012*	**-0,004**	**-0,006**	**-0,005**
		(0,002)	(0,002)	(0,003)		(0,003)	**(0,003)**	(0,003)		**(0,003)**	**(0,003)**	**(0,003)**
Education	0,183***	0,092***	0,113***	0,118***	0,031	0,045	0,041	0,046	0,018	0,042	0,046	0,044
		(0,016)	(0,018)	(0,021)		(0,025)	(0,027)	(0,028)		(0,028)	(0,030)	(0,032)
Urban	0,121	0,045	0,053	0,071	0,343*	**0,334*****	**0,342*****	**0,339****	-0,286*	**-0,076**	**-0,161**	**-0,164**
		(0,078)	(0,084)	(0,098)		**(0,086)**	**(0,092)**	**(0,105)**		**(0,103)**	**(0,109)**	**(0,121)**
DSL	0,064	0,014	0,031	0,081	0,222*	**0,179****	**0,204****	**0,188****	0,029	**0,148***	**0,174***	0,136
		(0,051)	(0,052)	(0,058)		**(0,064)**	**(0,067)**	**(0,070)**		**(0,068)**	**(0,070)**	(0,073)
DO	-0,126*	-0,094*	-0,087*	**-0,125****	-0,046	**-0,115****	**-0,123****	**-0,119***	-0,220**	**-0,184*****	**-0,195*****	-0,177**
		(0,037)	(0,039)	**(0,041)**		**(0,043)**	**(0,045)**	**(0,049)**		**(0,054)**	**(0,056)**	(0,060)
STALIN	-0,041	-0,031	-0,055	-0,051	-0,178**	**-0,190*****	**-0,224*****	**-0,201*****	-0,062	-0,033	-0,066	-0,026
		(0,029)	(0,032)	(0,032)		**(0,039)**	**(0,042)**	**(0,042)**		(0,041)	(0,043)	(0,042)
EI	-0,076***	-0,069***	-0,087***	-0,078***	-0,074**	**-0,046*****	**-0,048*****	**-0,055*****	-0,058**	**-0,039***	-0,040**	**-0,068*****
		(0,009)	(0,009)	(0,010)		**(0,013)**	**(0,012)**	**(0,014)**		**(0,015)**	(0,014)	**(0,016)**
Russian									-0,335*	**-0,166**	**-0,168**	**-0,178**
										(0,138)	**(0,147)**	**(0,158)**
R^2	0,170	0,115	0,180	0,161	0,120	0,106	0,125	0,123	0,090	0,045	0,067	0,087

Note: *** p < 0,001, ** p < 0,01, * p < 0,05. **Bolded** entries represent coefficients whose significance level is different from the original model.

CONCLUSIONS

What can be concluded from these comparisons? As the results of the means comparisons indicate, 13% of the differences change significance and an additional 20% change significance levels. Similar results are obtained in the regression analyses 16% of all coefficients change significance and an additional 30% change significance levels. These differences in the results obtained using the same data but different treatments of missing data may explain, in part, why some debates in the literature continue for a long period of time. In the examples presented here, for instance, the coefficients for education gain or lose significance, depending on what model is used. Education was at the center of a debate between Finifter and Mickiewicz and Reisinger *et al*. that started in 1992 and was still going on in 2003. It seems thus that the way social scientists usually treat missing data may have significant effects on the results they obtain.

The statisticians tell us that listwise deletion is not a good solution for the missing data problems that characterize survey research. They also tell us that other traditional methods of dealing with missing data (i.e., mean imputation, hot-deck imputation, and regression imputation), while better than just deleting the cases, still fail to obtain unbiased estimators. The solution advocated by the statisticians is to use a maximum likelihood method of imputing the missing data or multiple imputation. The advantage of multiple imputation over the maximum likelihood methods is given by its wider applicability on different statistical models: while maximum likelihood methods cannot offer any help beyond linear and log-linear models, multiple imputation can be used with any type of statistical models.

Until recently, technical difficulties prevented the widespread use of multiple imputation methods. In the last years, however, software for imputing the missing data has become readily available. Moreover, for those who do not like learning a new statistical package, the SAS, Stata, and SPSS have all implemented (with more or less success – see, for instance, von Hippel, 2004) a multiple imputation procedure. Under these circumstances, it becomes clear that social scientists will have to adopt multiple imputation as the conventional way of solving the missing data problems.

REFERENCES

1. Allison, P., *Missing Data*, "Sage University Papers Series on Quantitative Applications in the Social Sciences", series no. 07–136, Thousand Oaks, CA: Sage, 2002.

2. Arbuckle, J., Wothke W., *Amos user's guide, version 4.0*., Chicago: Marketing Division SPSS Inc.: SmallWaters Corporation, 1999.

3. Brehm, J., *The Phantom Respondents: Opinion Surveys and Political Representation*, Ann Arbor, University of Michigan Press, 1993.

4. Comşa, M., *O analiză a ratei de răspuns în anchetele de opinie naţionale*, „Sociologie Românească", 3–4: 1–32, 2002.

5. Comşa, M., *O analiză a ratei de răspuns la itemi în anchetele de opinie naţionale*, "Sociologie Românească", 4: 56–72, 2003.

6. Finifter, A. W., Mickiewicz, E., *Redefining the Political System of the USSR: Mass Support for Political Change*, "American Political Science Review", 86: 857–874, 1992.

7. Horton, N. J., Lipsitz, S. R., *Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables*, "The American Statistician", 55: 244–254, 2001.

8. Horton, N. J., Kleinman, K. P., *Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models*, "The American Statistician", 61: 79–90, 2007.

9. Jones, M. P., *Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression*, "Journal of the American Statistical Association", 91: 222–230, 1996.

10. King, G. *et al*., *Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation*, "American Political Science Review", 95: 49–69, 2001.

11. Little, R. J., *Regression with Missing X's: A Review*, "Journal of the American Statistical Association", 87: 1227–1237, 1992.

12. Little, R. J., Rubin, D. B., *Statistical Analysis with Missing Data*, New York, John Wiley and Sons, 1987

13. Mackelprang. A. J., *Missing Data in Factor Analysis and Multiple Regression*, "Midwest Journal of Political Science", 14: 493–505, 1970.

14. Miller, A. H., Reisinger, W. M., and Hesli, V., *New Soviet Citizen Survey, 1992: Monitoring Political Change*, Computer file, ICPSR version, Iowa City, IA: A. H. Miller, W. Reisinger, and V. Hesli, Iowa Social Science Institute [producers], 1992, Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2000.

15. Reisinger, W. M. *et al*., *Political Values in Russia, Ukraine, and Lithuania: Sources and Implications for Democracy*, "British Journal of Political Science", 24: 183–223, 1994.

16. Rubin, D., *Inference and Missing Data*, "Biometrika", 63: 581–592, 1976.

17. Rubin, D., *Multiple Imputation for Nonresponse in Surveys*, New York, John Wiley and Sons, 1987.

18. Rubin, D., *Multiple Imputation After 18+ Years*, "Journal of the American Statistical Association" 91: 473–489, 1996.

19. Schafer, J. L., *NORM: Multiple imputation of incomplete multivariate data under a normal model*, version 2, Software for Windows 95/98/NT, available online at http://www.stat.psu.edu/~jls/misoftwa.html, 1999.

20. Schafer, J. L., Olsen, M. K., *Multiple Imputation for Multivariate Missing Data Problems: A Data Analyst's Perspective*, "Multivariate Behavioral Research", 33: 545–571, 1998.

21. Voicu, B., *Despre măsurarea intenţiei de vot în sondajele de opinie*, "Sociologie Românească", 4: 48–76, 1999.

22. Von Hippel, P. T., *Biases in SPSS 12.0 Missing Value Analysis*, "The American Statistician", 58: 160–164, 2004.

23. Wang, R., Sedransk, J., and Jinn, J. H., *Secondary Data Analysis When There Are Missing Observations*, "Journal of the American Statistical Association", 87: 952–961, 1992.

24. Wothke, W., Longitudinal and multi-group modeling with missing data, in Little, T., Schnabel, K, and Baumert, J. (eds.), *Modeling Longitudinal and Multilevel Data: Practical Issues, Applied Approaches, and Specific Examples*, Mahwah, NJ, Lawrence Erlbaum Associates, 2000.

În acest articol analizez o serie de tehnici dezvoltate pentru înlocuirea non-răspunsurilor. Am ales drept exemplu unul dintre cele mai citate modele din literatura care analizează valorile politice în ţările post-comuniste, modelul propus în Reisinger et al. (1994). Acest model a adus o importantă contribuţie literaturii de specialitate, prin analiza valorilor politice din Rusia la începutul tranziţiei. Analizele din acest model, însă, sunt afectate de soluţia folosită de autori pentru a rezolva problema non-răspunsurilor: eliminarea cazurilor din analiză. Teoria statistică oferă soluţii alternative pentru această problemă, soluţii ce duc la obţinerea unor rezultate nebiasate. Pornind de la aceste alternative, estimez acest model folosind trei metode diferite pentru tratamentul non-răspunsurilor (imputarea la valoarea medie, imputarea prin regresie şi imputarea multiplă), pentru a testa dacă rezultatele din Reisinger et al. sunt valabile. Rezultatele re-estimărilor din acest articol arată că rezultatele iniţiale nu îşi menţin validitatea şi sugerează imputarea multiplă ca soluţie pentru problema estimării valide a coeficienţilor în prezenţa non-răspunsurilor.

***Cuvinte cheie**: non-răspunsuri, imputare multiplă, metodologie, software statistic.*