**DOI 10.15407/sociology2024.03.062**
**UDC 316.4**

**ANDRII GORBACHYK,**
*Candidate of Sciences in Mathematics, Associate Professor, Faculty of Sociology, Taras Shevchenko National University of Kyiv (64/13, Volodymyrska St., Kyiv, Ukraine, 01601)*
*a.gorbachyk@knu.ua*
*https://orcid.org/0000-0003-1944-435X*

**YAROSLAV KOSTENKO,**
*PhD student, Faculty of Sociology, Taras Shevchenko National University of Kyiv (64/13, Volodymyrska St., Kyiv, Ukraine, 01601)*
*yarosl.kostenko@gmail.com*
*https://orcid.org/0009-0001-7878-5034*

# Using Paradata for Imputation of Missing Values in Sociological Survey Data: Results of Statistical Modeling (Case of Croatia and Slovakia)

## *Introduction*

The scarcity of complete data sets is a common issue for the quantitative sociological researches. Traditionally, researchers have relied on complete case analysis, a method that, while prevalent, is often criticized for introducing significant biases into study results or reducing the sample size. The alternative solution is data imputation, which has been outlined specifically for the social sciences in the works of Rubin (1977), Little (1989), McKnight (2007), etc. Basic methods of data imputation commonly introduce heavy biases, as shown by numerous works (e.g. Lee, 2011), and thus can't be recommended for use in cases other than minimal amount of missing data. More sophisticated approaches to handling missing data begin with its classification, with most common classification proposed by Rubin in 1977 and further explored in numerous works, e.g. Graham (2009), Newman (2014), Mirzaei et al. (2022). Beyond classification, understanding of the dataset's structure is essential, as is selecting the correct imputation technique for the data at hand.

Paradata is a relatively new type of data associated with the digitalization of data collection, with the term proposed by Couper (1998) as an additional, by-product, auxiliary data which was collected through a computer system. Such data has a variety of uses such as research robustness or data evaluation and assessment. For instance, European Statistician System has proposed improving survey quality through the

*Citation*: Gorbachyk, A., Kostenko, Ya. (2024). Using Paradata for Imputation of Missing Values in Sociological Survey Data: Results of Statistical Modeling (Case of Croatia and Slovakia). *Sociology: Theory, Methods, Marketing*, 3, 62–82, https://doi.org/10.15407/sociology2024.03.062.

analysis of paradata (Aitken et al., 2004). Since paradata can hold the respondent information that's not obtainable through the survey means, it can also be used be used to construct more accurate predictive models. Brunton-Smith and Tarling (2017) investigated the use of multilevel multiple imputation and paradata in managing missing data for the longitudinal Surveying Prisoner Crime Reduction study. Their approach utilized advanced imputation methods and paradata analysis to effectively address both unit and item nonresponse issues. Another of applications of paradata during longitudal researches is described by Skafida (2022), where paradata is used as a predictor for non-response for longitudal study regarding domestic violence in order to construct a model closer to the would-be answers, if respondents have provided them. One more example on how paradata can be used during imputation is shown by Mathiowetz (1998), while discussing utilizing the expressions of uncertainty during imputation by comparing two imputation models: one uses the entire pool to fill in the missing data, other uses only the 'successful probe reporters' subgroup—one that expressed some unsureness during questioning, but gave the answers in the end, under the assumption that they might possess similar qualities to those that did not give a successful answer in the end.

The aim of this article is to estimate the process of incorporating paradata into the imputation process, in a way that can be further evaluated. To explore this approach, we start with an ideal dataset as a subset of the initial dataset with no missing values. In order to generate missing data in a realistic way, we propose a novel method of using an algorithm based on clusterization of respondents with regards to their non-response pattern. The goal of this approach towards production of missing values is to create more real-like missing data, one that's MAR. This approach simulates real-life scenarios of data collection where some questions may be more likely to have missing responses than others, with the fraction of missing values for each item reflecting this, and some respondents in particular are less likely to respond to such questions, with our clusters reflecting different patterns of non-response. As a result, we get various datasets with different fractions of missing data, all of the MAR nature.

We then apply several imputation methods, including those that utilize paradata, to address these gaps. In this case, we are working with quasimetric scales and using regression models, one of the models employing paradata to provide an additional predictor for imputation process.

To evaluate the effectiveness of these approaches, we compare the models using several key statistical metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared ($R^2$). The RMSE and MAE metrics help us understand the average magnitude of errors in our imputed values—essentially, how much the data we've filled in deviates from what was originally there. A lower score in these metrics indicates better accuracy of the imputed data. On the other hand, the R-squared ($R^2$) metric offers insight into the proportion of variance in the original data that our model can replicate. A higher $R^2$ value suggests that our imputed dataset closely reflects the original datasets variability, indicating a more accurate and reliable imputation process. These metrics collectively allow us to assess how closely our imputed datasets mirror the original data, providing a thorough evaluation of our imputation techniques effectiveness. This comparison is designed to assess not only

the direct benefits of incorporating paradata but also to identify any potential limitations or challenges associated with use of imputation.

### *Dataset Description*

The dataset of our choice is the European Social Survey (ESS), wave 10. ESS is a multinational survey that includes a variety of European countries, allowing for comparative analysis of various countries. In addition, ESS provides an extensive amount of paradata, collected as a 'Contact Form Data', which provides a wide selection of data that can be used, for example, to validate the quality of data collection, or, in our case, to work with missing data. 'Interviewers Questionnaire'—part of a survey that's filled by an interviewer can also be interpreted as paradata, however, for the purpose of working with missing data our focus will be on the 'Contact Form Data'.

In the process of preparing our data for analysis, we integrated three distinct arrays of data from the ESS, wave 10. These include the main dataset containing the survey results, a dataset derived from contact forms, and a dataset compiled from interviewer questionnaires. The integration was based on two key variables: idno, the unique identifier for each respondent, and country, indicating the respondents country. Together, these variables formed a unique ID for every participant, ensuring precise merging of the datasets.

We merged the data for respondents present across all three sources. This approach allowed for analysis that incorporated survey responses and the paradata, allowing for further usage of paradata. While 'Contact Form Data' also contains records for the respondents for which interview hasn't been conducted for a variety of reasons (such as: inability to get in touch, respondent's refusal for various reasons, etc.), for the purpose of this research, our dataset was comprised only of those individuals who participated in the survey, excluding non-respondents.

Deciding to take advantage of the multinational factor of the ESS, we've decided to work with two different countries for our experiment. The choice of countries themselves is tied to the questions of interest—the questions must be those where non-responses can be an issue, and also ones where missingness may have a non-random pattern. We've chosen three variables related to the LGBT issues, and will discuss them in more detail in the next block.

As such, countries have to be the ones where LGBT issues may be a controversial topic—namely those that are more socially conservative, suggesting Southern or Eastern Europe. At the same time, comparing countries from different regions may produce extended insights compared to two similar ones.

As such, we've chosen two fairly socially conservative, yet with significant differences countries of Europe—Croatia and Slovakia. Both of these countries have shared a similar historical trajectory being part of larger federations before gaining independence in the early 1990s. In addition, religion plays a significant role in both countries, with Catholicism being predominant in Croatia and Slovakia, unlike some of their less religious neighbors like Slovenia or Czech Republic. Overall, these countries both have enough similarities for a meaningful comparison and enough differences for a broader spectrum of insights.

### *Picking the Questions of Interest*

As our questions of interest, we've selected three ordinal variables from the survey, each related to attitudes toward LGBT issues:

LGBT-Related Familial Shame (hmsfmlsh): This variable measures the respondent's potential shame regarding having a gay or lesbian close family member. The response scale ranges from 1 (agree strongly) to 5 (disagree strongly), indicating the degree of shame or acceptance. Values higher than 5 indicate missing values. This variable has 4.46% of missing values for Croatia and 10.3% missing values for Slovakia.

Rights of Gay and Lesbian Couples to Adopt (hmsacld): This question evaluates the respondents support for the right of gay and lesbian couples to adopt children. Responses are scaled from 1 (agree strongly) in support of these rights to 5 (disagree strongly) against them. Values higher than 5 indicate missing values. This variable has 3.89% of missing values for Croatia and 7.19% missing values for Slovakia.

Freedom for Gays and Lesbians to Live Openly (freehms): This variable measures the extent to which respondents believe gays and lesbians should be free to live their lives as they wish, without societal constraints or discrimination. The scale is from 1 (agree strongly) for full support of these freedoms to 5 (disagree strongly) for opposition. Values higher than 5 indicate missing values. This variable has 3.64% of missing values for Croatia and 5.85% missing values for Slovakia.

*Table 1*

**Proportion of valid responses for Croatia and Slovakia**

| Variable | HR Missing Value Percentage | SK Missing Value Percentage |
|---|---|---|
| hmsfmlsh | 4.46% | 10.30% |
| hmsacld | 3.89% | 7.19% |
| freehms | 3.64% | 5.85% |

The rationale behind choosing these specific questions is based on the following considerations:

1. LGBT issues, in Eastern and Southern Europe specifically, remain a sensitive question with socially desirable answers, and thus these questions contain a significant fraction of non-responses.
2. Correlation with Paradata Variables: We identified a statistically significant relationship between these questions and the paradata variables we've selected. This indicates that the context of data collection may have a connection to the respondent's answers to these questions. In the paradata bloc, we'll explain the significance of these correlations.
3. Mutual Intercorrelation: These questions are not only individually significant but also interrelated, suggesting that attitudes towards one aspect of LGBT rights may be associated with attitudes towards others. This mutual correlation allows us to explore patterns of attitudes within the dataset more thoroughly.
4. Suitability as Quasi-Metric Variables: These questions are ordinal, but can be interpreted in a quasi-metric manner for the purposes of our analysis. This means we can treat them as if they were metric (continuous) variables, which allows for more nuanced statistical analysis.

These questions are chosen not only for their relevance to societal attitudes but also because they are indicative of areas where paradata may reveal significant insights. By examining how various factors, including socio-economic status and the context of data collection, influence responses to these sensitive topics, we aim to uncover patterns that can improve our strategies for imputing missing data. This focus allows us to investigate the potential of paradata to provide a deeper understanding of the complexities involved in survey responses, especially in areas where respondents may hesitate to provide full information.

### *Picking the Paradata Variables of Interest*

ESS dataset contains a vast amount of paradata variables, such as information about the contact attempts, refusal reasons, respondent's level of understanding of questions, observations regarding respondent's dwelling, and so on. To effectively harness paradata for our analysis, we've started with an initial phase of correlational analyses, specifically focusing on ordinal-scaled questions. Given the ordinal nature of these questions, we utilized Spearman's rank correlation coefficient as our tool of choice. This approach was crucial in identifying paradata variables that exhibit meaningful relationships with survey responses, thereby highlighting those with substantial predictive potential for filling in missing values. Our analysis was applied to various variables, ultimately selecting three that not only demonstrated significant correlations with survey questions but also had a theoretical foundation supporting their predictive value.

1. Physical Condition of Building/House (physa): Part of the ESSs Contact Form, it assesses the buildings condition, indirectly indicating material well-being. Its external assessment, rather than self-reporting by the respondent, is likely to enhance its validity. The values range from 1 (Very good) to 5 (Very bad), with values higher than 5 indicating missing values.
2. Amount of Litter and Rubbish (littera): Also from the Contact Form, this variable measures the cleanliness of the immediate vicinity, correlating with both socio-economic status and survey responses more strongly than the physa variable. The values range from 1 (Very large amount) to 4 (None or almost none), with values higher than 4 indicating missing values.
3. Presence of Vandalism and Graffiti (vandaa): This variable complements the previous two by adding a cultural dimension to the socio-economic indicators derived from the abodes condition. The values range from 1 (Very large amount) to 4 (None or almost none), with values higher than 4 indicating missing values.

These variables can be used as indirect indicators of the respondent's living conditions, which allows us to use this aspect in our predictive models, enriching them with another aspect that's ought to improve their predictive capabilities. In addition, there are some differences between Croatian and Slovakian datasets, with Croatian respondents having, on average, better physical condition of their housing, but also a slightly higher presence of vandalism and graffiti.

### *Constructing the Predictive Model*

In our analysis to develop a robust predictive model for both Slovakia (SK) and Croatia (HR), we've focused on variables with significant correlations to our primary LGBT-related questions, aligning with themes of social liberalism. We've decided to use model with 3 main aspects that represent different aspects that correlate with attitude towards LGBT:

1. Immigration Attitudes
2. Religiousness
3. Social Responsibility and Values

Then, we further enrich our model by implementing the fourth, paradata-based aspect of "Living Conditions". After filtering out variables with low item-level response rates, we identified the following as relevant for our model:

Immigration Attitudes:

- Allowance for Immigrants of Different Race/Ethnic Group (imdfetn): This variable measures respondents attitudes towards allowing immigrants of a different race or ethnic group from the majority to enter the country. It reflects broader societal views on racial and ethnic diversity among immigrants.
- Allowance for Immigrants from Poorer Countries (impcntr)—This question captures views on permitting immigrants from poorer countries outside Europe to settle. It gauges the level of openness towards economic migrants and refugees.
- Impact of Immigrants on Country (imwbcnt)—This variable assesses perceptions of whether immigrants make the country a worse or better place to live, offering insights into the perception of social and cultural impacts of immigration.
- Economic Effects of Immigration (imbgeco)—This question evaluates opinions on whether immigration is beneficial or detrimental to the country's economy, addressing economic dimensions of immigration debates.

Religious Practices and Beliefs:

- Religious Service Attendance (rlgatnd)—Measures how often respondents attend religious services, apart from special occasions, indicating the role of organized religion in their lives.
- Self-Reported Religiosity (rlgdgr)—This variable captures how religious respondents consider themselves to be, reflecting personal faith intensity.
- Frequency of Prayer (pray)—Assesses how often respondents pray outside of religious services, highlighting personal religious practices.

Social Responsibility and Values:

- Personal Responsibility for Climate Change (ccrdprs)—Measures the extent to which respondents feel personally responsible for reducing climate change, indicating environmental attitudes.
- Importance of Understanding Different People (ipudrst)—This variable gauge the value placed on understanding people from diverse backgrounds, reflecting attitudes towards social diversity and inclusivity.
- Value of Traditions and Customs (imptrad)—Assesses the importance attributed to following traditions and customs, revealing attitudes towards cultural preservation versus modernization.

- Equality and Equal Opportunities (ipeqopt)—Measures the emphasis on ensuring that people are treated equally and have equal opportunities, indicating views on equality, both of opportunities and attitudes.
- Helping and Caring for Others (iphlppl)—Captures the importance of helping and caring for others well-being, reflecting altruistic values and social empathy.

We also add our three paradata variables we've discussed in a previous paragraph: Physical Condition of Building or House, Amount of Litter and Rubbish, and Presence of Vandalism and Graffiti.

For the predictive model, it's very important that the predictor variables have correlations with the predicted ones, because otherwise they can't be employed in order to improve predictions using a regression model. All the variables from our list of predictors have statistically significant correlation coefficient (Spearman's R) with at the very least 2 out of 3 variables representing attitudes towards LGBT and have meaningful reasoning behind their choice as a part of a predictor model, for both the HR and SK datasets.

*Table 2*

**Spearman's R with significance: HR Dataset**

| | hmsfmlsh | hmsacld | freehms |
|---|---|---|---|
| imdfetn | –0.27 * | 0.29 * | 0.22 * |
| impcntr | –0.23 * | 0.26 * | 0.21 * |
| rlgatnd | 0.12 * | –0.26 * | –0.20 * |
| imwbcnt | 0.21 * | –0.27 * | –0.24 * |
| imbgeco | 0.16 * | –0.23 * | –0.18 * |
| ccrdprs | 0.17 * | –0.14 * | –0.15 * |
| pray | 0.10 * | –0.21 * | –0.12 * |
| rlgdgr | –0.17 * | 0.30 * | 0.18 * |
| ipudrst | –0.16 * | 0.10 * | 0.21 * |
| imptrad | 0.19 * | –0.29 * | –0.21 * |
| ipeqopt | –0.12 * | 0.08 * | 0.20 * |
| iphlppl | –0.08 * | 0.00 | 0.09 * |
| physa | –0.05 | 0.00 | 0.10 * |
| littera | 0.06 * | –0.09 * | –0.10 * |
| vandaa | 0.01 | –0.09 * | –0.09 * |

*Note: Coefficients marked with * are statistically significant at $p < 0.05$.*

These variables together form a comprehensive predictive model designed to estimate values for hmsfmlsh, hmsacld, and freehms. To ensure the integrity of our predictive model, we included only entries with complete data for the selected paradata variables, our three variables of interest, and all predictor variables. After filtering, the dataset used constitutes 73.82% of the original dataset, maintaining its representative nature and suitability for our imputation experiments.

*Table 3*

**Spearman's R with significance: SK Dataset**

| | hmsfmlsh | hmsacld | freehms |
|---|---|---|---|
| imdfetn | –0.16 * | 0.29 * | 0.29 * |
| impcntr | –0.16 * | 0.30 * | 0.26 * |
| rlgatnd | 0.19 * | –0.29 * | –0.24 * |
| imwbcnt | 0.21 * | –0.25 * | –0.27 * |
| imbgeco | 0.22 * | –0.28 * | –0.25 * |
| ccrdprs | 0.09 * | –0.11 * | –0.19 * |
| pray | 0.18 * | –0.28 * | –0.23 * |
| rlgdgr | –0.12 * | 0.32 * | 0.22 * |
| ipudrst | –0.04 | 0.04 | 0.12 * |
| imptrad | 0.09 * | –0.17 * | –0.09 * |
| ipeqopt | –0.08 * | –0.02 | 0.10 * |
| iphlppl | –0.05 | 0.01 | 0.04 |
| physa | –0.16 * | 0.02 | 0.02 |
| littera | 0.08 * | 0.15 * | 0.04 |
| vandaa | 0.07 * | 0.14 * | 0.05 |

*Note: Coefficients marked with * are statistically significant at $p < 0.05$.*

In the next phase of our analysis, we focus on a subset of the data, applying linear regression models to test the predictive power of our chosen variables on the three key questions of interest: attitudes towards LGBTQ+ individuals and issues. Linear regression is a statistical method that helps us understand how well our selected predictor variables can estimate responses to these questions. We treat our predictors as quasimetric, meaning that, despite being ordered categories, they can be analyzed similarly to continuous numerical data for this purpose.

Here is what we found from our models:

For HR (Croatia) the average mean squared error (MSE) for three LGBT-related variables is 1.22, and the average coefficient of determination (R-squared) is 0.22.

For SK (Slovakia) the average mean squared error (MSE) for three LGBT-related variables is 1.44, and the average coefficient of determination (R-squared) is 0.18.

The mean squared error tells us, on average, how much the predicted values deviate from the actual responses, with lower numbers indicating better accuracy. The coefficient of determination, or R-squared, measures how well the predictor variables explain the variation in responses to our questions of interest. In social sciences, an R-squared value with values around 0.2 is considered significant, indicating that our model does a fair job of predicting attitudes towards LGBT issues based on the variables we selected for Croatia and Slovakia. Therefore, these results suggest that our model is capturing important trends effectively enough to be used for further analysis.

### *Generating the Missing Values*

In the next phase of our study, we artificially introduce missing values into our variables of interest to closely mimic the real-world scenarios of data collection. Introducing missing data that's not Missing Completely at Random (MCAR) and resembles real-world scenarios always poses a challenge, as it requires introducing a realistic pattern to missing data. In addition, for our experiment, we can't use the same pattern as we do for the imputation procedure, as it would be 'overfit' for the task, restoring the data in the same pattern it is missing.

Therefore, another approach is required. In order to produce a scenario of Missing at Random (MAR) data that's aligned with the real-world scenarios, we propose a novel approach: usage of clusterization as a method to classify respondents based on their response patterns. Since some respondents tend to be significantly more likely to refuse responding a question, the goal of this approach is to identify respondents with different response patterns and account for that during the production of missing values. Then, for each of the three questions of interest, we'll produce missing values proportional to the missing fraction for each of the clusters. For example, if in the original dataset, for question A, cluster 1 has 5% of missing values, and cluster 2 has 10%, for the question A, we'll produce twice as much missing values for the respondents from cluster 2 compared to ones from cluster 1.

Our clusterization has to be done separately for both the Croatian and Slovakian dataset, in order to capture different response patterns. The fraction of missing values is significantly higher in the Slovakian dataset, meaning the item response rate is overall lower. Our clusterization has to reflect that, therefore we perform clusterization separately for each country.

***Table 4***

**Proportion of item-level response**

| Variable | HR Valid Frequency | SK Valid Frequency |
|---|---|---|
| hmsfmlsh | 0.955 | 0.897 |
| hmsacld | 0.961 | 0.928 |
| freehms | 0.967 | 0.942 |

The basis of our clusterization would be a set of dichotomous variables that capture the non-responses of the respondents. Such approach first requires constructing two separate datasets (one for Croatia, one for Slovakia) where our non-dichotomous variables are re-coded as dichotomous. If respondent has provided an answer to the question, it'll be re-coded as "1", otherwise it will be re-coded as "0". We re-code an extensive set of questions this way for the purpose of this clusterization. Included questions are the ones tied to the aspects of our topic of interest—politics, religiousness and social responsibility and values. While we include the questions we use for the regression models for the purpose of re-coding and being used as a basis of clusterization, we also include various other questions, with a full list provided in the appendix.

Since we're employing full case analysis for the purpose of experimentation, we can't limit ourselves to the list of variables employed in the initial predictive model, as

we'll work with a filtered dataset, and thus cluster distribution will be heavily one-sided. Therefore, a variety of other questions that don't necessarily correlate to the views regarding LGBT, but could be useful for exploring the pattern of non-responses on questions that involve politics, religion and social views.

After choosing the variables of interest, we create two separate datasets—one for Croatia, one for Slovakia, each having these questions re-coded into dichotomous. We keep IDs of each respondent, so after clusterization we can use them to assign a cluster value for each respondent from the original datasets.

For our choice of clusterization algorithm, we employ hierarchical clustering, using the Hamming distance to calculate the distance matrix. We use Ward's minimum variance method for calculating the distance between clusters. This approach is directed towards creation of flexible amount of clusters (the amount which we can specify after seeing the dendrograms) based on the dichotomous data. Ward's minimum variance method is aimed towards low variability within clusters, and tends to produce clusters of a more similar sizes, making it an optimal choice for our task.

While there's no single way to interpret these clusters, our goal for this clusterization is to generate realistic missing data, for which, ideally, we want a decent number of clusters with a significant size of items—our respondents—within them. Having most respondents belong to a single cluster could reduce the complexity of MAR data generation, and thus our goal is to find a minimum number of clusters that distribute the respondents between a variety of them. This proved to be harder to achieve for the Croatian dataset, where item-level response rate was higher, and thus the majority of the respondents were converging on a single cluster of respondents that's best interpreted as a cluster of respondents with overall high item-level response rate.

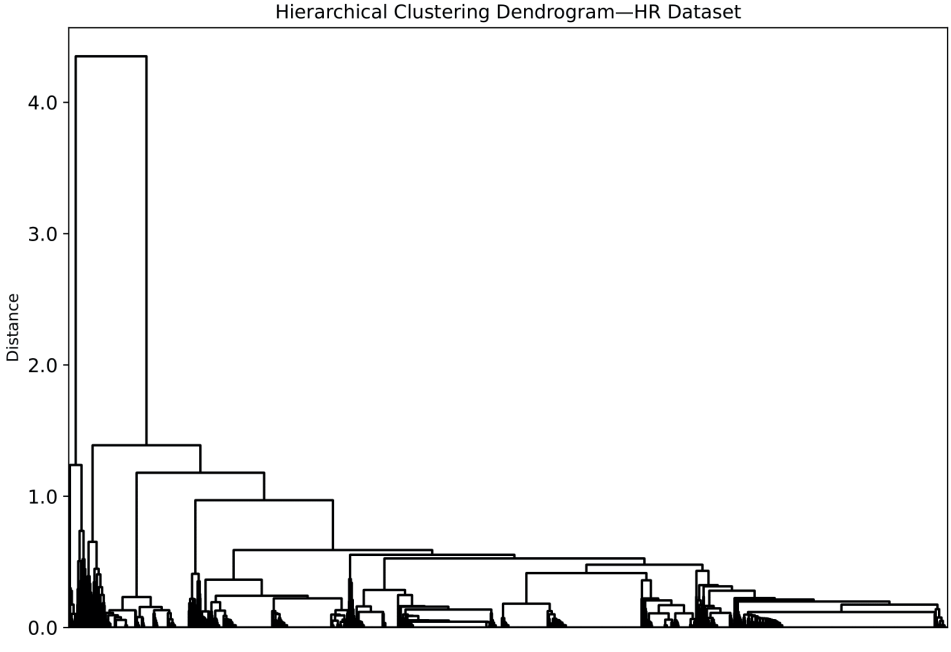


*Figure 1. Dendrogram Clusterization of HR dataset*

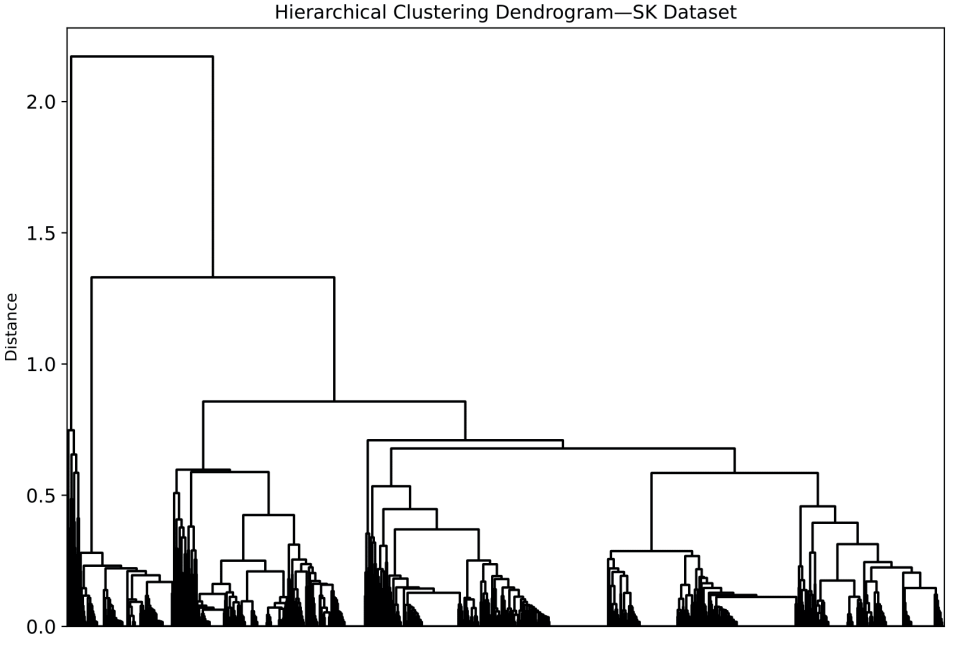


*Figure 2. Dendrogram Clusterization of SK dataset*

***Table 5***

**HR Cluster Percentages**

| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Percentage | 0.94% | 0.57% | 0.50% | 0.75% | 1.38% | 9.48% | 1.38% | 16.83% | 68.15% |

***Table 6***

**SK Cluster Percentages**

| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Percentage | 0.35% | 1.13% | 10.65% | 21.79% | 0.63% | 27.08% | 38.36% |

After experimenting with various counts of clusters, we've settled on 9 clusters from HR dataset and 7 clusters for SK dataset, with primary goal being clusters of comparable size. The interpretation of each of these clusters can be summed up as a 'nonresponse pattern', with each cluster representing a certain pattern towards answering (or non-answering) questions from the pool we've settled on previously. While differences between some of these clusters could be insignificant, overall they contribute towards the goal of generating missing data in a realistic way.

*Table 7*

**Fraction of valid values for HR clusters**

| C | hmsfmlsh | hmsacld | freehms |
|---|---|---|---|
| 1 | 0.13 | 0.07 | 0.07 |
| 2 | 0.78 | 0.89 | 0.78 |
| 3 | 0.25 | 0.38 | 0.50 |
| 4 | 0.67 | 0.75 | 0.83 |
| 5 | 0.91 | 0.86 | 0.86 |
| 6 | 0.98 | 0.99 | 0.98 |
| 7 | 0.86 | 0.95 | 1.00 |
| 8 | 0.93 | 0.94 | 0.93 |
| 9 | 0.98 | 0.98 | 0.99 |

*Table 8*

**Fraction of valid values for SK clusters**

| C | hmsfmlsh | hmsacld | freehms |
|---|---|---|---|
| 1 | 0.00 | 0.00 | 0.00 |
| 2 | 0.56 | 0.44 | 0.62 |
| 3 | 0.96 | 0.97 | 0.99 |
| 4 | 0.92 | 0.95 | 0.97 |
| 5 | 0.11 | 0.00 | 0.33 |
| 6 | 0.93 | 0.93 | 0.92 |
| 7 | 0.87 | 0.94 | 0.96 |

Fractions of valid values for our variables of interest will be used as a basis for generating missing values. It's important to note that these fractions are based off the entire dataset. Since we'll be working with a complete case scenario where all the predictor values and variables of interest will be present, we'll filter some of the array values, and this will affect these clusters disproportionately, with ones that have low item response rate to be much more likely to get filtered. Some of the clusters, ones that correspond to the very low item response rate, will entirely absent in our filtered datasets. For instance, for Croatia, 100% of respondents belonging to clusters 1 and 3 will be absent in the filtered dataset. Similarly, for Slovakia, 100% of respondents belonging to clusters 1, 2, and 5 will be absent in the filtered dataset.

For the respondents in our filtered dataset, we introduce missing values proportionally to the fraction of missing values in each cluster, for each question. For example, for those that belong to Croatian dataset's cluster 9, there's a 0.98 fraction of valid values for the question coded as 'hmsfmlsh'. This means that we'll proportionally introduce missing values with the goal of having 2 percent of Croatian's cluster 9 having missing values for this question, and so on for each question and each cluster.

This will be our first dataset with generated missing values. However, to explore various missing data scenarios, we generated 10 different datasets, each with an

incrementally higher fraction of missing values, with *i* = number of the dataset being the multiplier of the fraction of missing values. Going back to our previous example, it means that for *i* = 2, Croatian dataset's cluster 9, for 'hmsfmlsh' question, we'll introduce about 4% of missing values, for *i* = 3—6%, and so on. For some of the less frequent clusters with low item response rate, at high value of *i* there may be no values for these variables in cluster at all (100% missing values).

The overall fraction of missing data in these datasets ranges from minimal (about 5%) to more significant (up to 50%), allowing us to analyze the effects of different fractions of missing values. This approach enables us to assess the robustness of our predictive model across a spectrum of scenarios reflecting both minor and significant data missingness, following the realistic missing data scenario.

### *Imputation of Missing Data*

Subsequently, we employ Multivariate Imputation by Chained Equations (MICE) to impute the missing values in each of the 20 generated datasets, 10 for each country.

To address the missing values across the 20 datasets we created, we applied the MICE technique, using linear regression as a choice of model. This method allows us to impute missing data by creating multiple imputations, reflecting the uncertainty about the right values to impute. To evaluate the impact of paradata on the imputation process, we designed two different models for each dataset:

1. Base Model: This model uses only our original set of predictors, which include attitudes towards immigration, religious practices and beliefs, and social responsibility and values. Specifically, the predictors are:
   - Immigration Attitudes: imdfetn, impcntr, imwbcnt, imbgeco
   - Religious Practices and Beliefs: rlgatnd, rlgdgr, pray
   - Social Responsibility and Values: ccrdprs, ipudrst, imptrad, ipeqopt, iphlppl
2. Paradata Model: This model incorporates the original predictors along with three variables related to the respondents living conditions, namely:
   - physa: Overall assessment of physical condition of building/house
   - littera: Amount of litter and rubbish in the immediate vicinity
   - vandaa: Amount of vandalism and graffiti in the immediate vicinity

This approach allows us to assess the impact of an additional aspect—living conditions—on imputation accuracy.

To minimize the effects of randomness and ensure that our findings are reproducible, we generated missing values 10 times for each dataset, using a different seed for each iteration, starting from a 'starting_seed' of 10001. This systematic approach ensures that our imputation results can be consistently replicated and verified.

### *Methodology of Comparison*

To ensure the integrity of our imputed datasets, we conduct a comparative analysis by measuring the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared ($R^2$) metrics. These metrics allow us to evaluate the accuracy and reliability of our two predictive models—Baseand Paradata—against our original dataset, which has no missing values. The process is as follows:

1. For each predictive model, we undertake a detailed comparison with our base dataset—the original dataset that contains no missing values. We use three key metrics for this comparison: the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the R-squared ($R^2$). These metrics serve dual purposes: RMSE and MAE help us quantify the average errors in our imputed data, giving us a clear measure of accuracy, while the $R^2$ metric tells us how much of the variance in our original data is captured by the imputed data, indicating the imputations overall effectiveness.
2. To thoroughly assess the performance of our imputation models under varying conditions, we apply this comparative analysis across 20 distinct datasets—10 for Croatia, and 10 for Slovakia. Each of these datasets is denoted by an *i* variable (ranging from 1 to 10), where *i* represents a dataset with an incrementally higher fraction of missing values. Essentially, the *i* variable helps us systematically increase the missing data challenge, allowing us to observe how well each model performs as the complexity of imputation increases. For each *i*, from 1 to 10, we calculate the RMSE, MAE, and $R^2$ metrics for both Croatian and Slovakian datasets, enabling us to evaluate and compare the models performance across a spectrum of scenarios with varying degrees of missing data.
3. To account for variability and ensure robustness, we repeat this comparative analysis for each of the 10 seeds used to generate missing values. This step addresses the potential randomness in how values are imputed.
4. After conducting the analyses across all seeds, we calculate the average values for RMSE, MAE, and $R^2$ for each model and each *i* value. Averaging these metrics provides for a more stable and reliable measure of our models performance than just comparing them within a single seed.

By comparing these averaged metrics across our predictive models, we can assess how models perform while imputing missing data. This approach allows us to determine the effectiveness of incorporating paradata (variables responsible for living conditions) into the imputation process, for both Croatian and Slovakian datasets.

## *Results*

### *Comparison for Croatia*

***Table 9***

**Results for Croatian dataset**

| *i* | Base RMSE | Paradata RMSE | Base MAE | Paradata MAE | Base $R^2$ | Paradata $R^2$ | Diff RMSE | Diff MAE | Diff $R^2$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.2113 | 0.2282 | 0.0232 | 0.0258 | 0.9686 | 0.9641 | 0.0169 | 0.0025 | –0.0045 |
| 2 | 0.3441 | 0.3286 | 0.0566 | 0.0532 | 0.9188 | 0.9258 | –0.0155 | –0.0033 | 0.0070 |
| 3 | 0.4183 | 0.4166 | 0.0858 | 0.0848 | 0.8801 | 0.8811 | –0.0017 | –0.0010 | 0.0011 |
| 4 | 0.4620 | 0.4660 | 0.1076 | 0.1078 | 0.8538 | 0.8508 | 0.0040 | 0.0003 | –0.0030 |
| 5 | 0.5473 | 0.5407 | 0.1451 | 0.1414 | 0.7945 | 0.7998 | –0.0066 | –0.0038 | 0.0052 |
| 6 | 0.5993 | 0.5960 | 0.1724 | 0.1735 | 0.7544 | 0.7566 | –0.0032 | 0.0010 | 0.0022 |

| *i* | Base RMSE | Paradata RMSE | Base MAE | Paradata MAE | Base $R^2$ | Paradata $R^2$ | Diff RMSE | Diff MAE | Diff $R^2$ |
|---|---|---|---|---|---|---|---|---|---|
| 7 | 0.6460 | 0.6552 | 0.2020 | 0.2059 | 0.7153 | 0.7063 | 0.0092 | 0.0039 | –0.0090 |
| 8 | 0.6882 | 0.6792 | 0.2306 | 0.2265 | 0.6762 | 0.6843 | –0.0090 | –0.0040 | 0.0081 |
| 9 | 0.7291 | 0.7323 | 0.2569 | 0.2606 | 0.6371 | 0.6343 | 0.0032 | 0.0037 | –0.0029 |
| 10 | 0.7673 | 0.7728 | 0.2853 | 0.2888 | 0.5977 | 0.5916 | 0.0056 | 0.0035 | –0.0061 |

As shown by RMSE and MAE For Croatia, there's no significant difference between imputation quality of base and paradata-enhanced predictive models. This suggests that 'Living Conditions' is not a significant predictor for the LGBT-related questions of interest for this country.

The models demonstrated reasonable performance at lower values of *i*, corresponding to lower fractions of missing values. For example, at $i = 1$, the R-Squared value exceeded 0.96, indicating that our predictive model accounts for 96% of the variance in the missing values. Given that Croatia exhibited a lower proportion of missing values relative to Slovakia, the statistical coefficients remained robust even as *i* increased, albeit with a gradual decline observed. This trend underscores the efficacy of the models in handling datasets with varying degrees of missing information, particularly in scenarios characterized by minimal data omission.

*Comparison for Slovakia*

***Table 10***

**Results for Slovakian dataset**

| *i* | Base RMSE | Paradata RMSE | Base MAE | Paradata MAE | Base $R^2$ | Paradata $R^2$ | Diff RMSE | Diff MAE | Diff $R^2$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.4152 | 0.4172 | 0.0809 | 0.0814 | 0.8875 | 0.8862 | 0.0019 | 0.0005 | –0.0013 |
| 2 | 0.5830 | 0.5841 | 0.1624 | 0.1601 | 0.7769 | 0.7782 | 0.0011 | –0.0023 | 0.0012 |
| 3 | 0.7034 | 0.6872 | 0.2372 | 0.2323 | 0.6760 | 0.6894 | –0.0163 | –0.0049 | 0.0133 |
| 4 | 0.7986 | 0.8034 | 0.3113 | 0.3140 | 0.5828 | 0.5782 | 0.0048 | 0.0028 | –0.0047 |
| 5 | 0.9135 | 0.8968 | 0.3992 | 0.3880 | 0.4544 | 0.4752 | –0.0167 | –0.0113 | 0.0207 |
| 6 | 0.9871 | 0.9823 | 0.4731 | 0.4710 | 0.3613 | 0.3671 | –0.0049 | –0.0021 | 0.0058 |
| 7 | 1.0637 | 1.0620 | 0.5480 | 0.5498 | 0.2605 | 0.2607 | –0.0017 | 0.0017 | 0.0002 |
| 8 | 1.1446 | 1.1363 | 0.6335 | 0.6274 | 0.1434 | 0.1545 | –0.0083 | –0.0061 | 0.0111 |
| 9 | 1.1913 | 1.1937 | 0.6873 | 0.6891 | 0.0765 | 0.0709 | 0.0025 | 0.0018 | –0.0056 |
| 10 | 1.2454 | 1.2372 | 0.7449 | 0.7406 | –0.0055 | 0.0041 | –0.0082 | –0.0043 | 0.0097 |

Contrary to the Croatian dataset, the Slovakian dataset exhibited improvements, albeit modest, when paradata variables were incorporated into the imputation process. Notably, at $i = 5$, the R-Squared value increased by 0.02, indicating that the paradata-enhanced model accounts for an additional 2% of the variance compared to the model

without paradata. This increment suggests that 'Living Conditions' may play a non-negligible role in predicting responses to LGBT-related questions.

Although the overall performance of the model for the Slovakian dataset may appear inferior relative to the Croatian dataset, it is important to consider the significantly higher fraction of missing values within the Slovakian dataset. Notably, even at $i = 1$, the missing value rate surpasses 10% for certain questions, highlighting the challenges posed by the dataset's sparsity.

## *Conclusions*

In our study, we aimed to assess the impact of incorporating paradata into imputation models on the accuracy and reliability of the imputed data, with a particular focus on whether the inclusion of 'Living Conditions' paradata enhances the predictive capabilities of our imputation model. We've proposed a novel method of producing MAR missing values using clusterization based on respondents response patterns, which aims to create realistic missing data for the purpose of this experiment and could be incorporated in further similar researches.

Before discussing role of paradata in our study, it is crucial to address the efficiency of the imputation method employed and provide general guidelines for the data imputation procedure, which consists of a couple key steps:

1. Missing data considerations. Classifying the missing data is important understanding how to approach the task. Generally, missing data in social sciences is either MAR or MNAR, which means that probability of missing data is not the same for all observation. This requires application of other variables in order to improve on quality of data imputed. Another point worth paying attention to is the fraction of missing data. Even more robust data imputation techniques struggle with accurate representation of missing data when fraction of missing data is high–which depends on the complexity of the missing pattern, however, generally speaking, 20% of data being missing already tends to make data imputation significantly less accurate, as been shown in our experiments.

2. Construction of predictive model. After identifying the missing data, the next step should be selection of variables that'll be applied for the imputation procedure. In our case–dealing with LGBT-related questions–we've picked variables that represent 3 different aspects that have a meaningful connection with our variables of interest: Immigration Attitudes, Religiousness, and Social Responsibility and Values. In addition, for one of the models, we've used paradata variables that reflect abode condition, suggesting it could reflect the socioeconomical status of an individual. The choice of variables used during imputation should always be tied to the ones that are both meaningfully connected and have statistically significant correlations.

3. Choice of a method. While there is a large variety of methods used for data imputations, most of the commonly used ones tend to produce biased results, both on the level of distributions and on the level of intervariable connections. One of the few regularly used methods that does not negatively impact them is Multiple Imputation by Chained Equations (MICE) with linear regression, which is the one we recommend

for the general use in social sciences, if variables can be treated as quasimetric. Utilizing MICE with linear regression as the backbone for our imputation strategy provided promising results, particularly in datasets where the fraction of missing values remained relatively low. For instance, for the Croatian dataset, at $i = 1$, which is roughly equal to 5% of missing data the R-Squared value exceeded 0.96, whereas MAE was at 0.023.

Overall, usage of linear regression for the datasets demonstrated robust capability in accurately estimating missing data under scenarios where fraction of missing value wasn't too high. Based on these results, we can overall recommend the usage of regression models with Multiple Imputation with Chained Equations when dealing with missing data that may be treated as quasimetric and constructing a viable predictive model is deemed feasible.

Our analyses indicate that the inclusion of paradata may or may not be viable depending on a country. In the context of the Croatian dataset, the imputation quality—assessed through RMSE and MAE metrics—revealed no considerable difference between the base and paradata-enhanced predictive models. This suggests that within this dataset, 'Living Conditions' does not significantly influence the predictive accuracy for LGBT-related questions of interest. Conversely, the Slovakian dataset demonstrated slight improvements upon the integration of paradata variables into the imputation process, which hints at the potential significance of 'Living Conditions' in predicting responses to LGBT-related questions for the Slovakian dataset.

Based on these observations, our conclusion is that while paradata's inclusion does not universally improve imputation model performance, it may offer marginal benefits under specific conditions. When paradata may cover some aspect which might be deemed important for the predictive model, yet not recorded through the traditional means of surveying, it could be a valid addition to the predictive model that will improve the quality of imputations. While the enhancements observed were minor, they suggest the possibilities for future research to further investigate and possibly expand the utility of paradata in enhancing the quality of imputed datasets.

## APPENDIX

Variables used for the clusterization procedure:

- nwspol—News about politics and current affairs, watching, reading or listening, in minutes
- netusoft—Internet use, how often
- netustm—Internet use, how much time on typical day, in minutes
- ppltrst—Most people can be trusted, or you can't be too careful
- pplfair—Most people try to take advantage of you, or try to be fair
- pplhlp—Most of the time people helpful or mostly looking out for themselves
- polintr—How interested in politics
- psppsgva—Political system allows people to have a say in what government does
- actrolga—Able to take active role in political group
- psppipla—Political system allows people to have influence on politics

- cptppola—Confident in own ability to participate in politics
- trstprl—Trust in country's parliament
- trstlgl—Trust in the legal system
- trstplc—Trust in the police
- trstplt—Trust in politicians
- trstprt—Trust in political parties
- trstep—Trust in the European Parliament
- trstun—Trust in the United Nations
- trstsci—Trust in scientists
- vote—Voted last national election
- prtdgcl—How close to party
- lrscale—Placement on left right scale
- stflife—How satisfied with life as a whole
- stfeco—How satisfied with present state of economy in country
- stfgov—How satisfied with the national government
- stfdem—How satisfied with the way democracy works in country
- stfedu—State of education in country nowadays
- stfhlth—State of health services in country nowadays
- gincdif—Government should reduce differences in income levels
- freehms—Gays and lesbians free to live life as they wish
- hmsfmlsh—Ashamed if close family member gay or lesbian
- hmsacld—Gay and lesbian couples right to adopt children
- euftf—European Union: European unification go further or gone too far
- lrnobed—Obedience and respect for authority most important virtues children should learn
- loylead—Country needs most loyalty towards its leaders
- imsmetn—Allow many/few immigrants of same race/ethnic group as majority
- imdfetn—Allow many/few immigrants of different race/ethnic group from majority
- impcntr—Allow many/few immigrants from poorer countries outside Europe
- imbgeco—Immigration bad or good for country's economy
- imueclt—Country's cultural life undermined or enriched by immigrants
- imwbcnt—Immigrants make country worse or better place to live
- atchctr—How emotionally attached to [country]
- atcherp—How emotionally attached to Europe
- rlgblg—Belonging to particular religion or denomination
- rlgdnm—Religion or denomination belonging to at present
- rlgdgr—How religious are you
- rlgatnd—How often attend religious services apart from special occasions
- pray—How often pray apart from at religious services
- ccnthum—Climate change caused by natural processes, human activity, or both
- ccrdprs—To what extent feel personal responsibility to reduce climate change
- fairelc—National elections are free and fair
- dfprtal—Different political parties offer clear alternatives to one another
- medcrgv—The media are free to criticise the government
- rghmgpr—The rights of minority groups are protected

- votedir—Citizens have the final say on political issues by voting directly in referendums
- cttresa—The courts treat everyone the same
- gptpelc—Governing parties are punished in elections when they have done a bad job
- gvctzpv—The government protects all citizens against poverty
- grdfinc—The government takes measures to reduce differences in income levels
- viepol—The views of ordinary people prevail over the views of the political elite
- wpestop—The will of the people cannot be stopped
- keydec—Key decisions are made by national governments rather than the European Union
- fairelcc—In country national elections are free and fair
- dfprtalc—In country different political parties offer clear alternatives to one another
- medcrgvc—In country the media are free to criticise the government
- rghmgprc—In country the rights of minority groups are protected
- votedirc—In country citizens have the final say on political issues by voting directly in referendums
- cttresac—In country the courts treat everyone the same
- gptpelcc—In country governing parties are punished in elections when they have done a bad job
- gvctzpvc—In country the government protects all citizens against poverty
- grdfincc—In country the government takes measures to reduce differences in income levels
- viepolc—In country the views of ordinary people prevail over the views of the political elite
- wpestopc—In country the will of the people cannot be stopped
- keydecc—In country key decisions are made by national governments rather than the European Union
- chpldm—Best for democracy: government changes policies in response to what most people think
- chpldmi—Important for democracy: government changes policies in response to what most people think
- chpldmc—In country government changes policies in response to what most people think
- stpldmi—Important for democracy: government sticks to policies regardless of what most people think
- stpldmc—In country government sticks to policies regardless of what most people think
- implvdm—How important for you to live in democratically governed country
- accalaw—Acceptable for country to have a strong leader above the law

***References***

Aitken, A., Hörngren, J., Jones, N., Lewis, D., & Zilhro, M.J. (2004). *Handbook on improving quality by analysis of process variables.* Eurostat.

Brunton-Smith, I. & Tarling, R. (2017). Harnessing paradata and multilevel multiple imputation when analysing survey data: A case study. *International Journal of Social Research Methodology*, 20(6), 709-720. https://doi.org/10.1080/13645579.2017.1287842

Couper, M.P. (1998). *Measuring Survey Quality in a CASIC Environment*. Survey Research Center, University of Michigan.

Graham, J.W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549-576.

Lee, J. H. & Huber Jr., J. (2011). Multiple imputation with large proportions of missing data: How much is too much? In: *Proceedings of the 23rd United Kingdom Stata Users' Group Meetings*. Stata Users Group.

Little, R.J.A. & Rubin, D.B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, 18(2-3), 292-326. https://doi.org/10.1177/0049124189018002004

Mathiowetz, N.A. (1998). Respondent expressions of uncertainty: Data source for imputation. *Public Opinion Quarterly*, 62(1), 47-56.

McKnight, P.E., McKnight, K.M., Sidani, S., & Figueredo, A.J. (2007). *Missing Data: A Gentle Introduction*. Guilford Press.

Newman, D.A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, 17(4), 372-411. https://doi.org/10.1177/1094428114548590

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316696

Skafida, V., Morrison, F., & Devaney, J. (2022). Answer refused: Exploring item non-response on domestic abuse questions in a social survey affects analysis. *Survey Research Methods*, 16(2), 227-240. https://doi.org/10.18148/srm/2022.v16i2.7823

*Received 06.05.2024*

**ANDRII GORBACHYK, YAROSLAV KOSTENKO**

## Using Paradata for Imputation of Missing Values in Sociological Survey Data: Results of Statistical Modeling (Case of Croatia and Slovakia)

*Missing values are a common issue in quantitative social researches. One of the ways to handle missing data is by data imputation. This article outlines the challenges of traditional data imputation methods, which often introduce biases, and presents an advanced approach that features integration of paradata—auxiliary information collected during surveys—into the imputation process, using the European Social Survey (ESS) as its dataset. It is proposed that the usage of paradata could enhance predictive models used for imputation. It discusses the practical applications of data imputation, particularly through the lens of sensitive topics such as LGBT issues in socially conservative countries, where missingness could be heavily skewed due to social inacceptability of certain answers. To evaluate the effectiveness of the proposed approach towards imputation, the research employs the approach of using the 'ideal dataset', which is a subset of the original dataset with no missing vales, and then introduces artificial missing values that are not MCAR (Missing Completely at Random) to simulate the real case of missing data. Having artificial missingness allows for evaluation of the imputation procedure by comparing it with the original dataset. The study uses a novel approach towards creation of realistic missing data patterns through clustering based on response patterns. The research uses advanced statistical methods to handle missing data, and incorporates paradata from the survey process to improve the accuracy of predictive models. By comparing statistical metrics such as RMSE, MAE, and R-squared, the article evaluates the effectiveness of these methods in mimicking the original dataset's variability.*

***Keywords:*** *missing data; item non-response; data imputation; multiple imputation; paradata; missing data patterns; modelling of missing data*

**АНДРІЙ ГОРБАЧИК, ЯРОСЛАВ КОСТЕНКО**

# Використання параданих для імпутації пропущених даних в соціологічних дослідженнях: результати статистичних експериментів (кейси Хорватії та Словаччини)

*Відсутні дані — це поширена проблема у кількісних соціологічних дослідженнях. Одним із способів розв'язання цієї проблеми є імпутація даних. У статті описуються проблеми традиційних методів імпутації даних, які часто викривляють дані, і представлено інноваційний підхід, який включає інтеграцію параданих — додаткової інформації, зібраної під час опитувань, — у процес імпутації, з використанням результатів European Social Survey (ESS) як масиву даних. У статті припускається, що використання параданих може підвищити якість предиктивних моделей, застосовуваних для імпутації. Обговорюються практичні застосування імпутації даних, особливо стосовно сенситивних тем, таких як питання ЛГБТ у соціально консервативних країнах, де може бути значна частка відсутніх даних через соціальну прийнятність певних відповідей. Для оцінки ефективності запропонованого підходу до імпутації дослідження використовує підхід з 'ідеальним набором даних', який є підмножиною оригінального набору даних без відсутніх значень, а потім вводить штучні відсутні значення, що не є повністю випадковими (MCAR), для імітації реального кейсу відсутніх даних. Наявність штучно згенерованих пропущених даних дозволяє оцінити процедуру імпутації, порівнюючи її з оригінальним набором даних. Дослідження використовує інноваційний підхід до створення реалістичних патернів відсутніх даних через кластеризацію на підставі патернів не-відповідей респондентів. Дослідження застосовує передові статистичні методи для роботи з відсутніми даними й інтегрує парадані для підвищення точності предиктивних моделей. Порівнюючи статистичні метрики, такі як RMSE, MAE та R2, автори статті оцінюють ефективність цих методів у відтворенні варіативності оригінального набору даних.*

***Ключові слова:*** *пропущені дані; не-відповідь; імпутація даних; множинна імпутація; парадані; патерни пропущених даних; моделювання пропущених даних*