Imputing Missing Answers in the World Values Survey

Arsen Matej Golubovikj[1], Branko Kavšek[1,2] and Marko Tkalčič[1]

[1] University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies, Glagoljaška 8, SI-6000 Koper, Slovenia

[2] Jožef Stefan Institute, Department for Artificial Intelligence, Jamova 39, SI-1000 Ljubljana, Slovenia

Abstract

Questionnaire surveys are useful for many areas of science, in particular social sciences. Such surveys are often the prime means of gathering data directly from participants, however, they are prone to missing data, which could be caused by many reasons: (i) an error by survey administrators, (ii) participants not responding to certain questions, (iii) acts of nature and, (iv) etc. In order to keep the full survey sample, researchers must often use imputation to deal with the missing data problem. Methods for imputation can sometimes offer reasonable estimates for the missing data, however, in the case of the survey: (i) imputation can add high noise to the data, (ii) imputation becomes unreliable when more than 40% of the data is missing. This work attempts to address these issues by evaluating if the usage of matrix completion methods stemming from collaborative filtering (CF) in recommender systems can yield more accurate imputations of survey data. The rationale for the usage of these methods is (i) the similarity between the problem framing, methods and data representation used in CF and survey imputation; (ii) the effectiveness of CF-based methods in recommender systems. We use data from the World Values Survey, a valuable dataset in social science of high volume and veracity, to compare (i) one simple approach to imputation, (ii) two established imputation approaches (iii) two CF matrix completion techniques. The results show that our chosen CF matrix completion techniques perform overall comparable, but not better than existing imputation techniques for the case of survey imputation. The matrix completion techniques, however, might prove useful in niche situations, such as in the imputation of non-ordinal question answers. The right technique for imputation often depends on the problem, these results beckon the consideration of CF-based techniques in future research on survey imputation.

Keywords

imputation, survey, matrix completion, collaborative filtering

1. Introduction

For many areas of science, in particular social sciences, questionnaires are an essential tool for gathering data. The process of collecting data through questionnaires, called a survey [4], has advantages, such as getting data directly from the participants, but also downsides, such as missing data values [16]. There are many reasons why data might be missing, (i) the survey administrator/s made an error [12], (ii) participants might not answer all questions i.e. item non-response [12], (iii) other reasons, e.g. acts of nature. No matter their cause, missing values in questionnaire-acquired data must be dealt with before researchers can make inferences from the data [2].

Human-Computer Interaction Slovenia 2022, November 29, 2022, Ljubljana, Slovenia

matej.golubovik@gmail.com (A. M. Golubovikj); branko.kavsek@famnit.upr.si (B. Kavšek); marko.tkalcic@famnit.upr.si (M. Tkalčič)

A common approach to dealing with missing values is to delete all entries which contain them [10]. The advantage of this deletion is its simplicity [10], however, it forces the researcher to operate on a partial dataset, which might produce misleading results [10]. To operate on the whole data, missing values must often be imputed i.e. filled in with replacement values. Often used techniques for imputation in surveys include: (i) simple imputation [12], which replaces missing data in a variable with its average or most frequent value (ii) hot-deck imputation [12], which exploits the similarities between entries in the data to find suitable replacements (iii) model based approaches [12], which model each variable based on the available data and fill in missing values using the model for each variable.

Existing imputation techniques have advantages, such as, allowing the user to operate on the full data, however, they can have the issues of: (i) introduction of high noise to the data [10] and, (ii) in the survey case, ineffectiveness when more than 40% of the data is missing [12] (high missingness). In our work, we address these issues by evaluating if the usage of alternative imputation methods that are commonly used in recommender systems (RS), can yield more accurate imputations of missing values, both in the case of low and high missingness. The rationale for the usage of these methods is (i) the similarity of the problem framing between questionnaires and RS, and (ii) the effectiveness of these methods in recommender systems.

These similarities in problem framing, are most noticeable in collaborative filtering (CF) for recommender systems. CF operates on a user-to-item ratings matrix that stores the opinion of human users about given items, usually expressed as a scalar value called a rating (ex. 1-5 Likert scale, where 1 is a very negative and 5 is a very positive ratting). Due to the large volume of items in such systems, users are usually familiar with only a fraction of the items, consequently, much of the entries in the ratings matrix are empty [1], i.e. missing. The recommendation is then done by filling these missing entries using solely data from this matrix, through a process called matrix completion [1], items with high predicted ratings, i.e. opinion, are then recommended to the user. If we represent the questionnaire data as a matrix, where rows represent participants, and columns represent questions, the problem of filling missing data is now similar to the problem of matrix completion.

This paper focuses on the comparison between matrix completion techniques and classical survey imputation techniques, in the task of filling in missing answers in the World Values Survey [6] - a highly valued dataset in the field of social science [11].

2. Related Work

A number of studies have utilized matrix completion and collaborative filtering outside of the field of recommender systems. Some of the fields which have used these techniques include medicine[7], bioinformatics[14], image processing[5], infrastructure[9] and security[13]. Many of these fields find favorable results in the use of collaborative filtering for their specific problems, especially when large amounts of data is missing. Moreover, the specific works of Saha et al.[14] and Li et al.[9] have successfully utilized matrix completion in the imputation of DNA and highway traffic-related data, respectfully.

Two works examine the use of matrix completion in a broad imputation scenario: (i) Wang et al. produce an ensemble-based imputation method, which includes an item-to-item collaborative technique in the ensemble, they show that their ensemble method outperforms k-nearest neighbors (KNN) imputation, on common datasets from the UCI (University of California Irvine) data repository, however, do not evaluate the performance of the item to item collaborative technique on its own; (ii) Chi and Li [3] examines the use of low-rank matrix completion for the general role of imputation, they use synthetic data to show that low-rank matrix completion techniques can operate under the statistical assumptions for missing data, utilized in imputation.

In the case of survey imputation, the use of matrix completion is also highlighted in some cases. Vozalis et. al.[17] test the usage of a user-based collaborative filtering technique in the imputation of a small transportation survey consisting of univariate question answers on the Likert (1-5) scale. They report a MAE (Mean Absolute Error) of 0.846 for this technique when imputing data with 20% missing answers. Similarly, Oliveira et al. [17] compare matrix factorization and item-to-item collaborative filtering techniques for the purpose of predicting univariate Likert scale questionnaire responses in a large company survey. They find that, on 20% missing data, these techniques can distinguish between a positive and negative response with an Area Under the Curve (AUC) score of at least 0.80 on the given data.

Although there has been research using matrix completion on survey data, to the best of our knowledge, there have been no attempts to compare the effectiveness of matrix completion techniques and classical survey imputation techniques. In this work, we fill this gap by directly comparing both approaches on the scenario of World Values Survey data.

3. Data Overview

For the purpose of comparing the effectiveness of matrix completion and classical imputation techniques in the case of missing survey data, we utilize data from the World Values Survey (WVS). The WVS is an international research program devoted to the scientific and academic study of social, political, economic, religious, and cultural values of people in the world [6]. In our testing, we use a subset of the data from the WVS's 7th wave (7th iteration) of the survey, conducted across 57 countries in the years 2017-2021. This subset used in our testing contains the answers of 84638 participants to 274 survey questions, covering topics such as: (i) ethical values (ii) social values and perceptions (iii) political values, (iv) stances on various social and political questions, (v) etc.

The questions used in the WVS are closed questions, meaning that the participants respond using a list of provided answers rather than articulating the answers themselves. Responses are recorded as a number which denotes the participant's choice from the list. The ranges of the numbers used to record responses in the WVS are from 1 to the number of answers, e.g. for five answers the answer range is 1 to 5. Among the questions in our subset, we find 8 question answer ranges: ”1-2”, ”1-3”, ”1-4”, ”1-5”, ”1-7”, ”1-8”, ”1-10” and ”1-11”. Based on their range, the questions answers in our subset can be divided into three categories: (i) Dichotomous - questions with binary (e.g. Yes/No) answers, questions in the ”1-2” range fall into this category, (ii) Nominal-Polytomous - questions with a set of more than two answers with no inherent ordering, in our subset questions on the ”1-3” range fall into this category, and (iii) Ordinal-Polytomous, questions with a set of more than two answers which in themselves contain an ordering, in our subset all other ranges (”1-4” and up) fall into this category.

For reasons mentioned in the methodology section of this paper, we also retain data on the participant's country of origin in our testing subset.

4. Methodology

The flow of our methodology from the data preparation step to the final comparisons between approaches is presented in Figure 1.

The data utilized in our testing is described in Section 3. Among the three types of survey question answers, observed in section 3, i.e. (i) Dichotomous, (ii) Nominal-Polytomous, and (iii) Ordinal-Polytomous questions, we find two imputation tasks, namely, a regression task and a classification task. Ordinal-polytomous answers are handled using regression, while classification is used to handle answers to dichotomous and nominal-polytomous questions.

In both tasks, we compare the effectiveness of matrix completion and classical imputation approaches on the testing data for the specific task. The approaches remain the same in both tasks, only they are adjusted to fit the problem (classification or regression).

Three classical imputation approaches are considered: (i) simple imputation, which serves as a baseline, it imputes the mean value in the regression case and the mode value in the classification case; (ii) k-nearest neighbors (KNN) imputation, a hot-deck approach, in the regression case it uses weighted mean resolution to impute from the neighborhood, while mode resolution is used in the classification case; (iii) model based imputation, which performs initial simple imputation then imputes utilizing one regressor per feature, it uses linear regression with initial mean imputation for the regression task and a bayesian ridge regressor with initial mode imputation in the classification task.

The classical imputation approaches are compared to two matrix completion techniques: (i) item-to-item CF, which, similarly to KNN, uses weighted mean resolution among similar items in the regression task and mode resolution among similar items in the classification case; (ii) non-negative matrix factorization, refined by a Decision Tree regressor and classifier in the regression and classification tasks respectively.

The non-negative matrix factorization is refined by a decision tree in the following way. Let Q be a column of the original matrix and let Q' be the estimation of Q in the resulting matrix from matrix factorization, for each pair of Q and Q' we use the available data in Q to train a Decision Tree which predicts Q from Q' and use this model to predict the remaining missing answers in Q from Q'.

To compare the five approaches described above, for each task, we simulate varying degrees of missingness in the data, from 10% to 50%, and evaluate their performance in imputing the data. For regressors, we evaluate Mean Absolute Error (MAE) and Mean Squared error (MSE), while classifiers are evaluated with their accuracy, precision, recall, and F1 scores. The simulation and evaluation is done through an augmented cross-validation technique. A comparison between ordinary and augmented cross-validation is given in Figure 2.

To cater our imputation to the data, hence producing more robust results, we perform all imputations per country separately. We impute per country since, if the survey data contains clusters, such as those born of demographics, better imputation results are achieved if data is imputed for each cluster separately [8, 15], moreover, in international surveys, an often taken

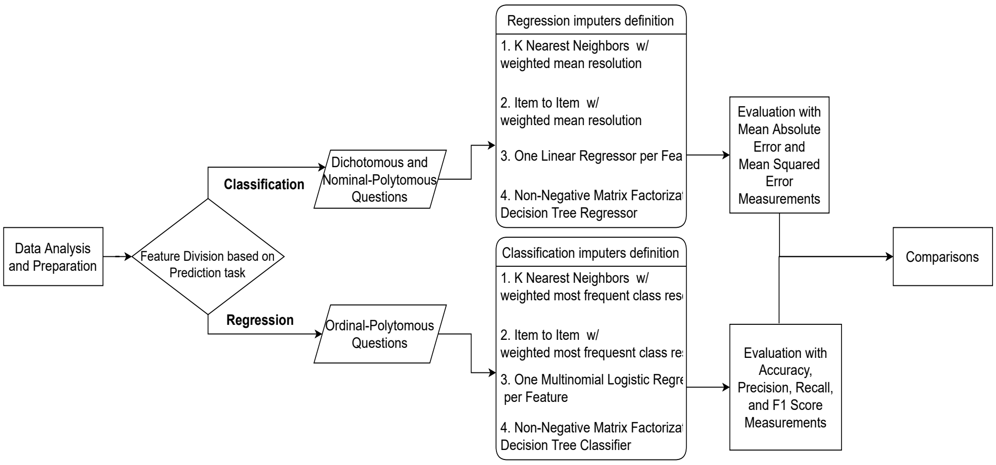
Data Analysis and Preparation
Feature Division based on Prediction task
Classification
Regression
Dichotomous and Nominal-Polytomous Questions
Ordinal-Polytomous Questions
Regression imputers definition
1. K Nearest Neighbors w/ weighted mean resolution
2. Item to Item w/ weighted mean resolution
3. One Linear Regressor per Fea
4. Non-Negative Matrix Factoriza
Decision Tree Regressor
Classification imputers definition
1. K Nearest Neighbors w/ weighted most frequent class res
2. Item to Item w/ weighted most frequesnt class re
3. One Multinomial Logistic Regre per Feature
4. Non-Negative Matrix Factoriza
Decision Tree Classifier
Evaluation with Mean Absolute Error and Mean Squared Error Measurements
Evaluation with Accuracy, Precision, Recall, and F1 Score Measurements
Comparisons

Figure 1: Flow of the imputation and evaluation procedures

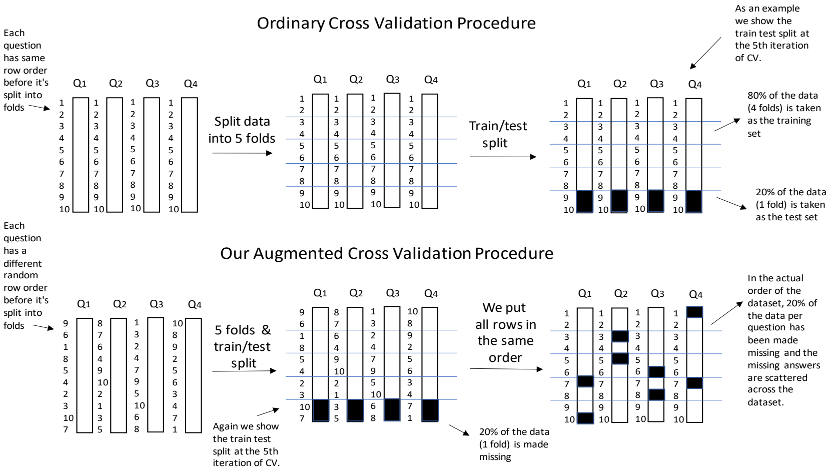
Ordinary Cross Validation Procedure
Each question has same row order before it's split into folds
Q1
Q2
Q3
Q4
Split data into 5 folds
Train/test split
As an example we show the train test split at the 5th iteration of CV.
80% of the data (4 folds) is taken as the training set
20% of the data (1 fold) is taken as the test set
Our Augmented Cross Validation Procedure
Each question has a different random row order before it's split into folds
5 folds & train/test split
Again we show the train test split at the 5th iteration of CV.
We put all rows in the same order
20% of the data (1 fold) is made missing
In the actual order of the dataset, 20% of the data per question has been made missing and the missing answers are scattered across the dataset.

Figure 2: Differences between our augmented cross-validation procedure and ordinary cross-validation

and effective approach is imputing answers for each country separately [18].

5. Results

5.1. Regression

Table 1 shows the Mean Absolute Error and Mean Squared error of our regression approaches. The final errors are calculated by taking the average across all questions, all countries, and finally across all magnitudes of missingness tested (from 10% to 50%). For the task of regression, all values are scaled from to the 1 to 3 scale. We keep this scale in our final results to make sense of average error in the multivariate scenario. The best-performing imputer is marked in bold.

Metric per %	Matrix Factorization (w/ Decision Tree)	Item to item CF	KNN	Regression Based Imputation	Mean Imputation (whole dataset)	Mean Imputation (per country)
MAE	0.4120	0.3809	0.3933	**0.3602**	0.5076	0.4488
MSE	0.3713	0.2677	0.2591	**0.2302**	0.3758	0.3171

Table 1
MAE and MSE scores for each regression imputation method considered, calculated by taking the average across all questions, all countries, and across all magnitudes of missingness tested (10%, 20%, 30%, 40% and 50%). The errors are presented on a scale of 1 to 3 (we would achieve similar answers on the -1 to 1 scale, as well). The best imputer is marked in bold.

5.2. Classification

The results for Accuracy, F1 Score, Precision, and Recall in the classification task are given in Table 2. Similarly, as in the regression case, the evaluation statistics presented are the average across all questions that fall under this task, as well as over all countries and magnitudes of missingness tested. The best-performing technique for each evaluation statistic is marked in bold.

From Table 2 we can see that the mode per country is a powerful predictor in the case of the classification task. This implies that the data is unbalanced, hence, the F1 score, Precision, and Recall are better indicators of the performance in this imputation task. Since the F1 score is a balanced measure between Precision and Recall, we will use it as the prime metric for comparison in the case of classification.

Metric per %	Matrix Factorization (w/ Decision Tree)	Item to item CF	KNN	Regression Based Imputation	Mode Imputation (whole dataset)	Mode Imputation (per country)
Accuracy	0.6940	0.6701	**0.7344**	0.6175	0.6680	0.7077
F1	**0.4745**	0.4116	0.4391	0.3483	0.3261	0.3413
Precision	**0.5345**	0.4599	0.4893	0.4427	0.2761	0.2985
Recall	**0.4952**	0.4180	0.4612	0.3574	0.4082	0.4133

Table 2
Accuracy, F1, Precision, and Recall scores for each classification imputation method considered, calculated by taking the average across all questions, all countries, and across all magnitudes of missingness tested (10%, 20%, 30%, 40% and 50%). For each score, the best imputer is marked in bold.

6. Discussion and Conclusion

The results show that our chosen CF matrix completion techniques perform overall comparable, but not better than existing imputation techniques for the case of survey imputation. The matrix completion techniques, however, might prove useful in niche situations highlighted in the results. Item-to-item collaborative filtering performs comparable to the KNN technique in both imputation tasks, only failing to match it on high ratios of missing data in the classification case. On the other hand, item-to-item fails to compare to model-based imputation in the regression, however, performs better than it in the classification task. Moreover, the results show that the matrix factorization technique offers poor performance in terms of MSE in the regression case, failing to match both existing imputation techniques, however, in the case of classification it outperforms all techniques tested with its F1 performance on unbalanced data.

In comparison with our related work, we achieve similar results to Vozalis et. al. [17] for MAE in terms of matrix completion, his MAE of 0.846 on univariate 1 to 5 data is comparable to our 0.40 MAE on the scale of 1 to 3, achieved under multivariate data. This raises the question of whether the scale affects the matrix completion techniques, collaborative filtering techniques in recommender systems usually operate on ratings all on the same scale. Can alterations of these techniques to fit multivariate data, be more beneficial in future work in survey imputation?

We also note that the nature of the data might affect the results, for example, the model-based imputer performs initial mean imputation before building its models, therefore the high performance of the model-based imputer in the regression task may be due to the power of mean imputation in our data. Future work might compare matrix completion and classical imputation techniques on a larger range of survey data.

Future work on this subject should also consider these techniques in different scenarios, as well as, examine the effects that these techniques have on the statistical inference. Moreover, our study included only simple techniques for matrix completion, CF techniques are vast and varied, and other techniques might succeed where we have failed, the considerations of such techniques in the study of imputation may also prove fruitful in future work.

References

[1] C. C. Aggarwal. *Recommender Systems*. Springer International Publishing, 2016. isbn: 978-3-319-29657-9. doi: 10.1007/978-3-319-29659-3.

[2] S. Buuren. *Flexible Imputation of Missing Data*. 2nd ed. New York: Chapman and Hall/CRC, 2018. isbn: 978-0-429-49225-9. doi: 10.1201/9780429492259.

[3] E. C. Chi and T. Li. "Matrix completion from a computational statistics perspective". In: *WIREs Comp Stat* 11.5 (2019). issn: 1939-5108, 1939-0068. doi: 10.1002/wics.1469.

[4] R. M. Groves, ed. *Survey methodology*. 2nd ed. Wiley series in survey methodology. Wiley, 2009. 461 pp. isbn: 978-0-470-46546-2.

[5] N. Gupta K.and Goyal and H. Khatter. "Optimal reduction of noise in image processing using collaborative inpainting filtering with Pillar K-Mean clustering". In: *The Imaging Science Journal* 67.2 (2019), pp. 100–114. issn: 1368-2199, 1743-131X. doi: 10.1080/13682199.2018.1560958.

[6] C. Haerpfer et al. *World Values Survey Wave 7 (2017-2022) Cross-National Data-Set*. In collab. with K. Kizilova et al. Version Number: 4.0.0 Type: dataset. 2022. doi: 10.14281/18241.18.

[7] F. Hao and R. H. Blair. "A comparative study: Classification vs. user-based collaborative filtering for clinical prediction". In: *BMC Medical Research Methodology* 16.1 (2016). issn: 1471-2288. doi: 10.1186/s12874-016-0261-9.

[8] N. Karmitsa et al. “Missing Value Imputation via Clusterwise Linear Regression”. In: *IEEE Transactions on Knowledge and Data Engineering* 34.4 (2022). Conference Name: IEEE Transactions on Knowledge and Data Engineering, pp. 1889–1901. ISSN: 1558-2191. DOI: 10.1109/TKDE.2020.3001694.

[9] L. Li et al. “Missing value imputation for traffic-related time series data based on a multi-view learning method”. In: *IEEE Transactions on Intelligent Transportation Systems* 20.8 (2019), pp. 2933–2943. ISSN: 1524-9050. DOI: 10.1109/TITS.2018.2869768.

[10] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* 3rd. John Wiley & Sons, 2019.

[11] S. G. Ludeke and E. G. Larsen. “Problems with the Big Five assessment in the World Values Survey”. In: *Personality and Individual Differences* 112 (2017), pp. 103–105. ISSN: 0191-8869. DOI: 10.1016/j.paid.2017.02.042.

[12] A. Mirzaei et al. “Missing data in surveys: Key concepts, approaches, and applications”. In: *Research in Social and Administrative Pharmacy* 18.2 (2022), pp. 2308–2316. ISSN: 15517411. DOI: 10.1016/j.sapharm.2021.03.009.

[13] R. M. Rodríguez et al. “Using collaborative filtering for dealing with missing values in nuclear safeguards evaluation”. In: *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems* 18.4 (2010), pp. 431–449. ISSN: 0218-4885. DOI: 10.1142/S0218488510006635.

[14] S. Saha et al. “Missing value imputation in DNA microarray gene expression data: A comparative study of an improved collaborative filtering method with decision tree based approach”. In: *International Journal of Computational Science and Engineering* 18.2 (2019), pp. 130–139. ISSN: 1742-7185. DOI: 10.1504/IJCSE.2019.097954.

[15] J. Shao and H. Wang. “Sample Correlation Coefficients Based on Survey Data Under Regression Imputation”. In: *Journal of the American Statistical Association* 97.458 (2002), pp. 544–552. DOI: 10.1198/016214502760047078.

[16] G. N. Singh et al. “Some imputation methods for missing data in sample surveys”. In: *Hacettepe Journal of Mathematics and Statistics* 45.6 (2016), pp. 1865–1880. ISSN: 2651-477X. DOI: 10.15672/HJMS.20159714095.

[17] M. Vozalis, S. Basbas, and I. Politis. “Applying Collaborative Filtering Techniques In Transportation Surveys”. In: *1st International Conference on Engineering and Applied Sciences Optimization.* 2014, pp. 1630–1638.

[18] M. Weber and M. Denk. *Imputation of Cross-Country Time Series: Techniques and Evaluation.* 2010.