Caveat of missing data in EM-DAT

Data are incomplete for disaster events attributed to natural and technological hazards in EM-DAT. There is a high proportion of missing data on the economic impacts of a disaster event. The EM-DAT data only presents reported data and does not include imputed data. Users are free to apply imputation methods based on their specific objectives. This document offers insights that may assist in achieving these objectives.

For disaster events attributed to natural hazards occurring between 1990 and 2020, proportions of missing data on the human impacts of a disaster event were found to range from 1.3% - 22.3% (Figure 1) [1]. Proportions of missing data were much greater on the economic impacts, ranging from 41.5% - 96.2%. The probability of missingness on the variables: number of people affected, number of deaths and total estimated damages (in US$) were partially explained by observed predictors of missingness: disaster type, income status of the country, disaster severity and the year the disaster occurred. For this reason, such missing data are unlikely to be missing completely at random (MCAR) [2]. In this case, methods to handle missing data that rely on the assumption of MCAR, are inappropriate and could bias study results. Instead, more advanced missing data methods are recommended.

We urge users of EM-DAT to consider the presence and potential mechanisms of missing data in their analyses and handle missing data accordingly. Users considering imputation methods should be aware that the different imputation approaches will lead to different results and increase variables uncertainty. Transparency in documenting the chosen imputation approach is essential for reproducibility. A glossary of conventional and advanced missing data methods is provided in Table 1; advanced methods such as multiple imputation or similar ensemble approaches, should be considered preferably as they often yield more reliable outcomes. Users are advised to test and identify, e.g., through cross-validation, the best approaches for their purpose and case study. Hereafter, we provide useful resources for this purpose.

Useful resources

Overview of missing data:

- Little, R. J. A. & Rubin, D. B. Statistical analysis with missing data (Wiley, 2002).

Missing data in empirical research:

- Brooks N, Adger NW. Country Level Risk Measures of Climate-Related Natural Disasters and Implications for Adaptation to Climate Change*. Tyndale Centre Working Paper*. **26**; (2003).
- Faria, R., Gomes, M., Epstein, D. & White, I. R. A Guide to Handling Missing Data in Cost-Effectiveness Analysis Conducted Within Randomised Controlled Trials. *Pharmacoeconomics*. **32**, 1157–1170 (2014).

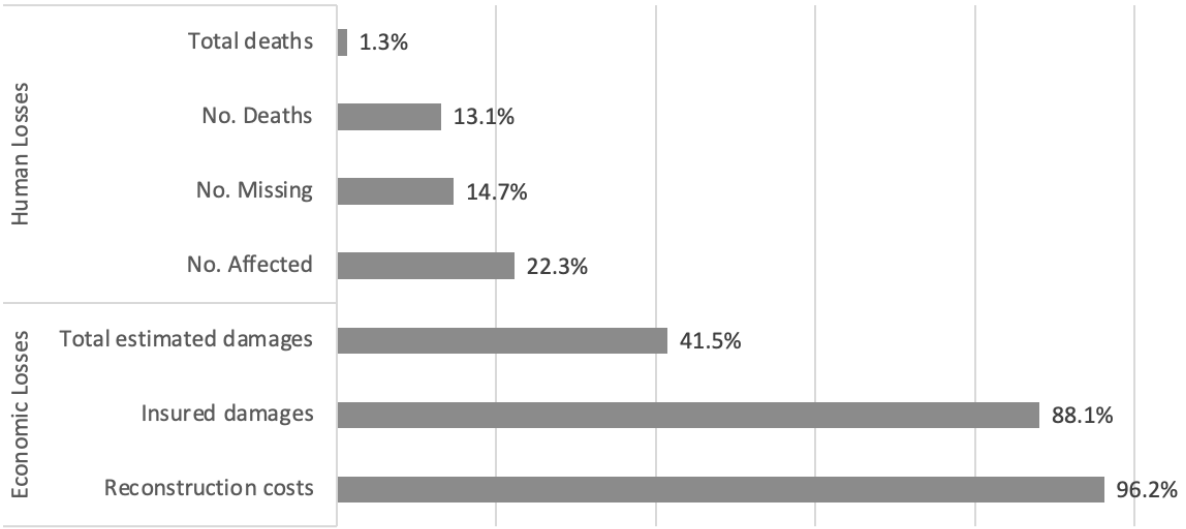
Methods to handle missing data:

- Schafer, J. L. & Graham, J. W. Missing Data: Our View of the State of the Art. *Psychological Methods*. **7**, 147–177 (2002).
- Allison, P. D. In The Sage handbook of quantitative methods in psychology (ed. Millsap, R. E.) Ch. 4 (Springer Publications Ltd., 2009).
- Graham, J. W., Cumsille, P. E. & Shevock, A. E. In Handbook of Psychology Vol. 2 (ed. Weiner) Ch. 4 (Wiley, 2012).

Overview of missing data in EM-DAT:

- Jones R. L., Guha-Sapir D., Tubeuf S. Human and economic impacts of natural disasters: can we trust the global data? *Scientific Data*. **9**; 1–7 (2022).
- Jones, R. L., Kharb, A., and Tubeuf, S.: The untold story of missing data in disaster research: a systematic review of the empirical literature utilising the Emergency Events Database (EM-DAT), *Environmental Research Letters*. **18**, 103006 (2023).

Figure 1. Proportions of missing data on disaster impacts in EM-DAT (Jones et al., 2022).



Data is restricted to disaster events attributed to natural hazards occurring between 1990 and 2020 (n = 11,124).

Table 1. Glossary of conventional and advanced missing data methods (Jones et al., 2023) [3]

Method	Description	Notes
Conventional methods		
Column deletion	Deleting variables which have a high proportion of missing data. A threshold of greater than 60% missing data is commonly suggested.	This method should only be considered for variables which are not necessary to the analysis.
Complete Case Analysis (CCA) (Listwise deletion)	Also referred to as row deletion. Observations with missing data on at least one variable of interest are excluded.	CCA is used by default in most statistical software programmes. It yields a complete dataset which facilitates the use of conventional data analysis methods. When a dataset contains a large proportion of missing data, CCA excludes a large fraction of the original data and reduces the statistical power of analyses. CCA relies on the assumption that missing data are MCAR or MAR if all predictors of missingness are included in the analysis.
Aggregating data	Compiling and expressing individual-level data into summary forms for statistical analysis.	Missing data are masked within summary statistics, minimising their relative impact. However, the precision of analyses are substantially reduced.
Dummy variable adjustment	For continuous variables, a dummy variable is created to indicate if data is missing on that variable. For categorical variables, an additional category is created to hold cases with missing data.	This method allows the entire dataset to be used in data analysis, maximising the sample size and statistical power. However, dummy variable adjustment has been shown to yield biased parameter estimates.
Available Case Analysis (ACA) (Pairwise deletion)	All observed values for each variable or pair of variables are utilised to calculate sample 'moments' (population mean, variance etc.). In other words, only missing data for the variable, or pairs of variables of interest are excluded. Sample moments are then included in the data analysis in place of population parameters.	Like CCA, this method yields a complete dataset which facilitates the use of conventional data analysis methods. As ACA uses all the data available for each analysis, it does not skew summary statistics. For bivariate and multivariate analyses, ACA requires sufficient correlation between variables to yield consistent parameter estimates. However, as different subsets of the data are used to calculate sample moments, there is no guarantee of this. ACA relies on the assumption of MCAR.
Mean imputation	Missing values are substituted with a single unconditional mean of the observed values.	Single imputation methods yield a complete dataset and facilitate the use of conventional data analysis methods independently of missing data methods. As with most single imputation methods, mean imputation yields biased parameter estimates. Predicted values do not contain random error, so sample variation is reduced. This can lead to an underestimation of standard errors and optimistic significance values. This issue is magnified with higher proportions of missing data.

Table 1. Glossary of conventional and advanced missing data methods (Jones et al., 2023) [3]

Method	Description	Notes
Regression-based imputation	Missing values are substituted with a single, predicted value estimated using regression methods, conditional on observed predictors of missingness.	Single imputation methods yield a complete dataset and facilitate the use of conventional data analysis methods independently of missing data methods. Relies on the assumption that missing data are MAR. As with mean imputation, regression-based imputation yields biased parameter estimates and uncertainty in the predicted value is not adequately reflected. Predicted values do not contain random error, so sample variation is reduced. This can lead to an underestimation of standard errors and optimistic significance values. This issue is magnified with higher proportions of missing data.
Data merging	Merging data sources, or data subsets by integration or aggregation to supplement existing data.	Data merging by conditional merging is most appropriate when merging incomplete datasets. This involves filling missing data gaps with observed values found in other source datasets. Data loss and file-matching errors can occur if there is heterogeneity in the coding of data across datasets, or if there is heterogeneity in the number and type of variables. Hence, datasets need to be standardised prior to merging. Data matching is also necessary to prevent the duplication of data across datasets. This method can therefore be timeconsuming.
Advanced methods		
Inverse probability weighting (IPW)	'Complete cases' are weighted by the inverse probability of being observed. Weights are calculated using a binary regression model conditional on observed predictors of missingness.	IPW rebalances the data so complete cases better represent the entire sample. By adjusting for missing data without manipulating the full dataset, IPW does not create issues of incompatibility with subsequent data analysis. Relies on the assumption that missing data are MCAR or MAR, if all predictors of missingness are included in the binary regression model.
Maximum-likelihood	Uses all observed data to generate the parameter estimates most likely to result from the available data. Likelihoods are computed separately for observations with complete and incomplete data on the variables of interest. The product of the individual likelihoods is then maximised to give the maximum-likelihood parameter estimates.	Maximum likelihood yields asymptotically unbiased and efficient parameter estimates. Missing data and parameter estimation are handled in a single step. However, this requires all predictors of missingness to be specified in the intended analysis model. Relies on the assumption that missing data are MAR but can be modified for missing data which are MNAR. For each variable with missing data, parametric models for the joint distributions need to be specified. This is potentially difficult and parameter estimates may be sensitive to the choice of model. Maximum-likelihood is limited to only linear models and requires specialist statistical software packages.

Table 1. Glossary of conventional and advanced missing data methods (Jones et al., 2023) [3]

Method	Description	Notes
Multiple imputation	An extension of regression-based single imputation. Multiple imputation involves 3 steps: 1. Imputation using regression methods is performed several times, generating m imputed datasets. Each dataset contains a different, randomly drawn, imputed value for all missing values. 2. Datasets are analysed separately using standard methods. 3. The parameter estimates and standard errors obtained from each are combined using Rubin's Rules to generate a single set of parameter estimates and standard errors.	Multiple imputation yields asymptotically unbiased and efficient parameter estimates. By generating multiple, randomly drawn imputed values, multiple imputation adequately accounts for uncertainty in the predicted value. Makes no assumptions about the missing data mechanism; can be modified for missing data which is MNAR. Requires several decisions to be made on: the type of imputation model, the number of imputations (m), the number of iterations between imputations and the choice of prior distribution. With larger proportions of missing data, a greater number of imputations are required. Generally, m = 20 is considered sufficient. Potentially computationally difficult with a large number of variables and/or observations.
Other advanced methods		
Hot deck imputation	Each missing value is replaced with a plausible, observed value taken from similar observations within the same classification. Imputed values may be selected at random, or by using distance metrics, such as nearest neighbour matching.	This method yields a complete dataset and facilitates the use of conventional data analysis methods independently of missing data methods. Hot deck imputation does not require missing values to be modelled. Therefore, parameter estimates are less sensitive to model misspecifications. If there are large proportions of missing data, only a small sample of observations may be used to impute missing values, leading to replication of values and reduced sample variation. This can lead to an underestimation of standard errors and optimistic significance values.
Bayesian simulation	An extension of multiple imputation. Missing data are treated as additional, unknown variables for which posterior predictive distributions can be calculated by specifying a missing data model and Bayesian priors. Algorithms, such as Monte Carlo Markov Chain are then used to yield parameter estimates from the posterior predictive distributions.	Missing data and parameter estimation are handled in a single step. Bayesian analysis can be easily adapted for incomplete data. Bayesian priors may be based on expert opinion which can improve the reliability of results. As with multiple imputation, Bayesian simulation adequately accounts for uncertainty due to the missing values. Can be modified to account for any assumption on the mechanism of missing data Parameter estimates may be sensitive to model misspecifications. Requires specialist software and can be highly complex.

ACA, available case analysis; CCA, complete case analysis; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random.

Bibliography

1 Jones R. L., Guha-Sapir D., Tubeuf S. Human and economic impacts of natural disasters: can we trust the global data? *Scientific Data*. **9**; 1–7 (2022).

2 Rubin D. B. Inference and Missing Data. Oxford Univ Press. **63**; 581–92 (1976).

3 Jones, R. L., Kharb, A., Tubeuf, S. The untold story of missing data in disaster research: a systematic review of the empirical literature utilising the Emergency Events Database (EM-DAT). *Environmental Research Letters*. **18** (10), 103006. (2023).