The Journal of Open Source Software

verdata: An R package for analyzing data from the Truth Commission in Colombia

Maria Gargiulo [1¶], **María Juliana Durán** [1*], **Paula Andrea Amado** [1*], **and Patrick Ball**[1]

1 Human Rights Data Analysis Group ¶ Corresponding author * These authors contributed equally.

DOI: 10.21105/joss.05844

Software

- Review
- Repository
- Archive

Editor: Nikoleta Glynatsi

Reviewers:

- @jamesmbaazam
- @JosiahParry

Submitted: 16 August 2023
Published: 12 December 2023

Summary

In 2016, the Colombian Government and the guerrilla group *Fuerzas Armadas Revolucionarias de Colombia – Ejército del Pueblo* (FARC-EP) signed a peace agreement ending decades of hostilities between the two groups. As part of the agreement, the *Comisión para el Esclarecimiento de la Verdad, la Convivencia y la No Repetición* (CEV; hereafter the Truth Commission) was created, a temporary institution charged with investigating what happened during the armed conflict, clarifying violations of international humanitarian law, and explaining the conflict in all its complexity to Colombian society. As part of the truth and reconciliation process, the Truth Commission collaborated with the *Jurisdicción Especial para la Paz* (JEP), a temporary judicial body, and the Human Rights Data Analysis Group (HRDAG) to produce official statistics about the magnitudes and patterns of five human rights violations—forced displacement, enforced disappearance, homicide, kidnapping, and forced recruitment of minors (Amado et al., 2022). These analyses made use of over 100 databases compiled by over 40 organizations, including governmental institutions, civil society organizations, and victims' collectives.

The data compiled by the joint JEP-CEV-HRDAG project are publicly available from the *Departamento Administrativo Nacional de Estadística* (DANE). The data published by DANE is available in a format that may not be familiar to researchers who have not previously worked with statistical imputation methods. Recognizing this, `verdata` was created to support researchers in responsibly and correctly using the data despite the potential unfamiliarity of its structure. Researchers can use `verdata` to verify that the data files they are using in their analyses have not been altered, to replicate the main findings of the technical appendix, and to design new analyses of the conflict in Colombia.

Statement of need

Collecting data on human rights abuses in conflict settings is a difficult and often dangerous task. Organizations collecting such data may have constrained resources, lack physical access to areas where violence is occurring, or may be unable to collect data due to security concerns or community mistrust, among other challenges (e.g., Gargiulo, 2022; Price & Ball, 2014, 2015). As a result, the data produced by these data collection efforts is seldom a complete enumeration of all violence that occurred nor a statistically representative sample. Furthermore, instances of violence that are documented may be missing key information about victims, perpetrators, or contextual details about the violent events. The incompleteness of the data is not a critique of the data itself nor the organizations that courageously document human rights violations, but rather it is an empirical reality that must be addressed in quantitative work analyzing conflict-related violence.

The data analyzed in the joint JEP-CEV-HRDAG project was no exception to this empirical

reality and was subject to two types of missing data: missing fields and underreporting. Related to missing fields, some records were missing socio-demographic information about victims such as their age, sex, or ethnicity, information identifying armed groups thought to be responsible for the violence, or precise information about the date and location of a particular violent event. These gaps in the data pose challenges for analyses seeking to stratify the data based on any fields containing missing values. With respect to underreporting, some instances of violence were not documented by any of the databases we received, leaving some victims' stories untold (Amado et al., 2022). Moreover, this missingness is unlikely to be randomly distributed among members of the victim population, meaning that inferences drawn from samples of documented victims alone could result in erroneous conclusions about patterns of violence.

The joint JEP-CEV-HRDAG project employed two statistical methods to address the two types of missingness. To address missing fields within records of documented victims, the project used the `R` package `mice` (van Buuren & Groothuis-Oudshoorn, 2011) to perform multiple imputation (e.g., Murray, 2018), probabilistically filling in missing values at the record level multiple times. Multiple systems estimation (e.g., Bird & King, 2018; Chao, 2001), performed on the imputed replicate files, was then used to estimate the number of missing observations, that is, the number of the victims never documented by any of the data sources used in the project.[1] To estimate the number of missing observations, we used a Bayesian latent class multiple-capture model (Manrique-Vallier, 2016) implemented in the `R` package `LCMCR`. The analyses presented in the technical appendix of the joint project combine these two methods to examine patterns of enforced disappearance, homicide, kidnapping, and forced recruitment of minors in the armed conflict.[2]

DANE has published 100 imputed replicate files with missing values filled in at the record level available for each of these four violations. This data format where there is no single file representing "the data" may be unfamiliar to researchers who have not worked with multiple imputation methods in the past and researchers may be tempted to select a single imputed replicate file to conduct their analyses rather than computing their analyses on multiple replicate files and combining the results using standard practices based on the laws of total expectation and total variance. The `verdata` package aims to support researchers in using the data from the Colombian Truth Commission responsibly and correctly despite the potential unfamiliarity of its structure. Software packages have not historically been created to facilitate the access and use of data published by past truth commissions. To date, the `pinochet` package (Freire et al., 2019), which facilitates access to data about killings and disappearances published in the Chilean Truth Commission, is the only other example of a software package created for this purpose.

To complement the package, we have also created a repository of examples of basic function use, replications of main findings from the technical appendix, and applications to other studies of interest not examined in the technical appendix. We have also published a series of pre-calculated estimates that researchers can opt to use to reduce the computational costs of multiple systems estimation. These pre-calculated estimates are available from the Colombian Truth Commission website.

We hope that `verdata` will play a role in expanding the use of statistical methods to address the two types of missing data in research on the conflict in Colombia, and armed conflicts more generally, so that the statistical biases apparent in individual data sources are not reproduced in future research on the conflict.

[1] Multiple systems estimation is also called capture-recapture in some disciplines.

[2] While the join JEP-CEV-HRDAG project also examined forced displacement due to the armed conflict, we were unable to provide multiple systems estimation estimates of forced displacements because nearly all documented victims were registered on only one list, the *Registro Único de Víctimas*. As a result, we did not have sufficient overlap with other sources to construct estimates using multiple systems estimation, which generally requires three or more sources in the case of applications to human rights questions.

Acknowledgements

We thank Tarak Shah, Valentina Rozo-Angel, and Dr. Megan Price for their helpful comments and Micaela Morales for her thoughtful beta testing.

Support for this project was provided by the British Embassy in Colombia Conflict, Stability and Security Fund, the Swiss Embassy in Colombia, Human Security Division, Justus Liebig University Gießen, with funds awarded by the German Foreign Federal Office, and Filecoin Foundation for the Decentralized Web.

References

Amado, P., Acero, W., Argoty, C., Babativa, G., Bernal, L. K., Castro, A., Durán, M. J., Espinosa, E., Gómez, V., González, A., Ortíz, M. A., Ball, P., Gargiulo, M., Rozo, V., Shah, T., Dueñas, J. G., Jaimes, L. E., Lozano, C., Murillo, M. J., & Rodríguez, H. (2022). *Informe metodológico del proyecto conjunto JEP-CEV-HRDAG de integración de datos y estimación estadística*. https://hrdag.org/wp-content/uploads/2022/08/20220818-fase4-informe-corrected.pdf

Bird, S. M., & King, R. (2018). Multiple systems estimation (or capture-recapture estimation) to inform public policy. *Annual Review of Statistics and Its Application*, *5*, 95–118. https://doi.org/10.1146/annurev-statistics-031017-100641

Chao, A. (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, *6*(2), 158–175. https://doi.org/10.1198/108571101750524670

Freire, D., Skarbek, D., Meadowcroft, J., & Guerrero, E. (2019). *Deaths and disappearances in the pinochet regime: A new dataset*. SocArXiv. https://doi.org/10.31235/osf.io/vqnwu

Gargiulo, M. (2022). Statistical biases, measurement challenges, and recommendations for studying patterns of femicide in conflict. *Peace Review*, *34*(2), 163–176. https://doi.org/10.1080/10402659.2022.2049002

Manrique-Vallier, D. (2016). Bayesian population size estimation using dirichlet process mixtures. *Biometrics*, *72*(4), 1246–1254. https://doi.org/10.1111/biom.12502

Murray, J. S. (2018). Multiple Imputation: A Review of Practical and Theoretical Findings. *Statistical Science*, *33*(2), 142–159. https://doi.org/10.1214/18-STS644

Price, M., & Ball, P. (2014). Big data, selection bias, and the statistical patterns of mortality in conflict. *The SAIS Review of International Affairs*, *34*(1), 9–20. https://doi.org/10.1353/sais.2014.0010

Price, M., & Ball, P. (2015). The limits of observation for understanding mass violence. *Canadian Journal of Law and Society/La Revue Canadienne Droit Et Société*, *30*(2), 237–257. https://doi.org/10.1017/cls.2015.24

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03