Readme File for Replication Package of

"Cultural Distance and Conflict-Related Sexual Violence"

Eleonora Guarnieri and Ana Tur-Prats

Quarterly Journal of Economics

1. Folder Contents

The replication folder includes data files in .dta format, one Stata do file, shapefiles (.shp), and a QGIS Project file to reproduce the main figures and tables:

Folder	*Subfolder*	*Content*
Working Files		Guarnieri_Tur-Prats_QJE_Index.dta Guarnieri_Tur-Prats_QJE_DHS.dta Guarnieri_Tur-Prats_QJE_Main_Analysis.dta Guarnieri_Tur-Prats_QJE_RSVAC.dta Guarnieri_Tur-Prats_QJE_SFE.dta
Do Files		Guarnieri_Tur-Prats_QJE_Analysis.do
Maps	Country borders	World_countries.shp (plus identically named auxiliary files with the following extensions: .cpg, .dbf, .prj, .qmd, .shx)
	EPR groups – CRSV	GeoEPR_CRSV.shp (plus identically named auxiliary files with the following extensions: .cpg, .dbf, .prj, .qmd, .shx)
	Murdock groups – MDI	MDI_map.shp (plus identically named auxiliary files with the following extensions: .cpg, .dbf, .prj, .qmd, .shx)
	QGIS Project	QGIS_Figures.qgz

2. Software

Stata 17.0 SE-Standard Edition

To run the do file Guarnieri_Tur-Prats_QJE_Analysis.do, the `reghdfe` command needs to be installed (`ssc inst reghdfe`). The results in the paper are obtained using the command's version 5.7.3 13nov2019.

QGIS Desktop 3.20.3

The QGIS_Figures.qgz file provides the user with a QGIS project that employs all layers necessary to reproduce Figure 1 and Figure 3 in the paper (World_countries.shp, GeoEPR_CRSV.shp, and MDI_map.shp). QGIS may require the user to re-specify the paths from which to retrieve the three layers.

3. Data Sources and Instructions for Replication

We hereby provide a description of how intermediate data sets were employed to create each of the final data sets. Additional details on the dataset construction can be found in the paper and in the Online Appendix. Full replication files are available upon request.

3.1. Guarnieri_Tur-Prats_QJE_Index.dta

This dataset is at the Murdock ethnic group level. It includes groups present in the Ethnographic Atlas and in the Murdock Map, their ancestral ethnic characteristics, and the Male Dominance Index (MDI).

Input data employed to construct the final dataset:

- Murdock Ethnographic Atlas (a-d) and Murdock Map (e). Sources:
  a) https://scholar.harvard.edu/files/nunn/files/ethnographic_atlas_final.dta_.zip
  b) https://scholar.harvard.edu/files/nunn/files/easternmost_europe_final.dta_.zip
  c) https://scholar.harvard.edu/files/nunn/files/siberia_final.dta_.zip
  d) https://scholar.harvard.edu/files/nunn/files/wes_final.dta__0.zip
  e) https://worldmap.maps.arcgis.com/home/item.html?id=d495496a57c1428eaee838f785a163c3

For details on the variables coding and the index construction, refer to section F.3 in the Online Appendix and to section 4.2 in the paper.

3.2 Guarnieri_Tur-Prats_QJE_DHS.dta

This dataset is at the survey respondent level. It includes information on contemporaneous proxies for gender (in)equality and each respondent is assigned to an MDI based on their reported ethnicity or language.

Input data employed to construct the final dataset:

- Demographic and Health Survey (DHS) survey waves. For a list of countries included in the sample, see table A-26 in section F.1 in the Online Appendix. Source:
  - https://dhsprogram.com/data/available-datasets.cfm
- To link self-reported ethnicity or language in the DHS survey to the Murdock Atlas we employed the Linking Ethnic Data for Africa (LEDA) R package (Müller-Crepon et al. 2020) for African countries. For non-African countries, we employ the Giuliano and Nunn's (2018) database. Sources:
  - https://github.com/carl-mc/LEDA
    - When employing the LEDA package in R, we specified `link.level = "language"` to match African groups. Details on how the LEDA package works can be found in the "R Package Documentation" section in Müller-Crepon et al. (2020).
  - https://scholar.harvard.edu/files/nunn/files/ancestral_characteristics_database_language_level.zip
    - To merge the DHS survey to the Giuliano and Nunn's (2018) database, we manually assigned a language code to self-reported languages in the DHS survey. Language codes come from the Ethnologue database (https://www.ethnologue.com/).

For details on how we coded DHS survey variables, refer to the paper's Appendix F.3.

3.3 Guarnieri_Tur-Prats_QJE_Main_Analysis.dta

This is a dyadic and bidirectional dataset at the perpetrator-victim-conflict-year level (1989-2019).

Input data employed to construct the final dataset:

- For the main dependent variable: Sexual Violence in Armed Conflict (SVAC) dataset and geo-SVAC datasets. The SVAC dataset has as unit of observation an actor (either rebel or government force) in a conflict in a year. Our analysis requires a dyadic and bidirectional structure (see below for more details on how we obtained it). In the first submission, this paper only covered the African continent from 1989 to 2009. For this geographical coverage and time period, the geo-SVAC dataset provided a dyadic structure (perpetrator-victim-conflict-year). Sources:
  - https://www.prio.org/download/datasetfile/47/GEO-SVAC.zip

    - http://www.sexualviolencedata.org/wp-content/uploads/2021/02/SVAC_3.0_complete.xlsx
- In the revised version of the paper, we expanded the temporal and geographical coverage of our dataset. For these additional years, since the geo-SVAC dataset is no longer available, we relied on the dyadic version of the UCDP dataset to obtain a dyadic structure. This dataset can be merged to the SVAC dataset using conflict id, actor id, and year. For additional details on the dyadic structure of the dataset see section 3.4 in the paper and table A-2 in the Online Appendix. Source:
    - https://ucdp.uu.se/downloads/ged/ged221-dta.zip
- To assign ethnic identities to conflict actors (government forces and rebels) we used the Ethnic Power Relations (EPR) dataset family (EPR for governments and ACD2EPR for rebels) and the GrowUp platform. For details on how we assigned ethnic identities to actors, refer to section F.4 in the Online Appendix. Sources:
    - Ethnic identities of rebel forces (ACD2EPR): https://icr.ethz.ch/data/epr/acd2epr/ACD2EPR-2021.csv
    - Ethnic identities of governments (EPR): https://icr.ethz.ch/data/epr/core/EPR-2021.csv
    - We used additional information and narratives from the GrowUp platform to confirm the validity of the matches: https://growup.ethz.ch/rfe
- To match EPR ethnic groups to the Ethnographic Atlas, we used the concordance table by Michalopoulos and Papaioannou (2016) (for Africa, our initial sample) and Giuliano and Nunn's (2018) database (for the remaining countries in our extended and final sample). For details on how we matched EPR groups to the Atlas, refer to section F.5 in the Online Appendix. Sources:
    - Concordance table by Michalopoulos and Papaioannou (2016): https://www.openicpsr.org/openicpsr/project/112954/version/V1/view?path=/openicpsr/112954/fcr:versions/V1/codes_aer_journal/EPRtoMurcdockMap_Concordance.dta&type=file
    - Giuliano and Nunn's (2018) database: https://scholar.harvard.edu/files/nunn/files/ancestral_characteristics_database_language_level.zip
- We merged Murdock ethnic groups to information on their ethnic characteristics and the MDI using the file Guarnieri_Tur-Prats_QJE_Index.dta (we link groups by name). We then aggregated ethnic characteristics up to the actor level. For details on the aggregation procedure, refer to section F.5 in the Online Appendix.
- Information on languages comes from the EPR Ethnic Dimensions dataset. Since the EPR Ethnic Dimensions belongs to the EPR dataset family, we can directly merge this data to our dataset via the shared EPR ethnic identity of conflict actors. Source:
    - https://icr.ethz.ch/data/epr/ed/ED-2021.csv.
- To compute linguistic distance between actors, we collected information on linguistic trees from the Ethnologue database. Section F.6 in the Online Appendix provides details on how we computed linguistic distance between actors. Source for linguistic trees:
    - https://www.ethnologue.com/browse/names
- Finally, to add information on battle-related deaths: UCDP Battle-Related Deaths dataset. This dataset is dyadic and can be merged using dyads, conflict-id, and year. For details on how we coded battle-related deaths refer to section 7.1 in the paper and section F.3 in the Online Appendix. Source:
    - https://ucdp.uu.se/downloads/index.html#battlerelated

3.4 Guarnieri_Tur-Prats_QJE_RSVAC.dta

Dyadic and bidirectional dataset at the perpetrator-victim-conflict-year level (1989-2015), which includes information on different types of sexual violence employed by conflict actors.

Input data employed to construct the final dataset:

- The Repertoires of Sexual Violence in Armed Conflict (RSVAC) dataset. Source:
    - https://github.com/thegargiulian/RSVAC

This dataset can be directly merged to our main dataset using conflict id, perpetrator's id, and year. For details on how we coded different types of sexual violence, refer to section F.3 in the Online Appendix and section 8.2 in the paper.

3.5 Guarnieri_Tur-Prats_QJE_SFE.dta

Dyadic and bidirectional dataset at the perpetrator-victim-conflict-year level (1989-2013, only Africa and Middle East), with an alternative coding of the ethnic identity of government forces.

Input data employed to construct the final dataset:

- The Security-Force Ethnicity (SFE) dataset for North Africa and the Middle East. Source:
    - https://journals.sagepub.com/doi/full/10.1177/0738894217709012
- The SFE dataset for sub-Saharan Africa. Source:
    - Shared confidentially with the authors by Dr. Ches Thurber

The unit of observation of the SFE dataset is a country-time period. We matched our dataset to the SFE dataset using country and year and we assigned an ethnic identity to government forces distinguishing between officers and rank-and-file soldiers. Conveniently, the SFE dataset assigns the ethnic composition of the military using the EPR definition of ethnicity. To assign the MDI to the newly-coded government forces (officers and rank and file), we use the same procedure we adopted for the main dataset (see section 3.3 above). For details on the coding procedure of the ethnicity for the rank-and-file and officers and the SFE variables employed, refer to section F.3 in the Online Appendix.

3.6 World_countries.shp

This is a world map with national boundaries. Source:

- https://international.ipums.org/international/resources/gis/IPUMSI_world_release2020.zip

3.7 GeoEPR_CRSV.shp

This is a shapefile including ethnic settlements from the geoEPR database. We add information on each group's use of conflict-related sexual violence through a variable ranging between 0 and 1 (see notes to figure 1 for additional details). In the shapefile, this variable is called "CRSV." A value of -99 indicates that the group did not participate in conflict in the period considered; a value of 0 indicates that the group was involved in conflict but did not use CRSV. Information on the use of sexual violence comes from the SVAC dataset (see section 3.1 above). We merged ethnic groups to the SVAC dataset as described above in section 3.1. Source of the shapefile on group's ethnic settlements:

- geoEPR 2021: https://icr.ethz.ch/data/epr/geoepr/

3.8 MDI_map.shp

This is a shapefile including Murdock ethnic groups for which we have information on the MDI. Each dot denotes an ethnic group. Each ethnic group can be located on the map using the Ethnographic Atlas' information on latitude (variable v104) and longitude (variable v106). The sources for the Ethnographic Atlas can be found in section 3.1. Information on groups' MDI comes from the file Guarnieri_Tur-Prats_QJE_Index.dta. In the shapefile, the MDI variable takes the name "MDI" and ranges between 0 and 1.

References

Giuliano, P. and Nunn, N. (2018). Ancestral characteristics of modern populations. *Economic History of Developing Regions*, **33** (1), 1–17.

Michalopoulos, S. and Papaioannou, E. (2016). The long-run effects of the scramble for Africa. *American Economic Review*, **106** (7), 1802–48.

Müller-Crepon, C., Pengl, Y. and Bormann, N.-C. (2020). Linking ethnic data from Africa.