Read Me: Replication Directory Overview

Human Rights Violations in Space: Assessing the External Validity of Machine Geo-coded Vs. Human Geo-Coded Data

24 August, 2021

README OVERVIEW

Author	Logan Stundal, stund005@umn.edu
Date	July 26, 2021
Article	Human Rights Violations in Space: Assessing the External Validity of Machine Geo-coded Vs. Human Geo-coded Data
Journal	Political Analysis
Co-Authors	Bagozzi, Benjamin; Freeman, John; Holmes, Jennifer
Purpose	This document explains script execution order to replicate all results in the paper / appendix as well as produce all figures and tables. Data imports and data exports are noted for each script.

Directory structure

- `Data/` - contains original data used as inputs for all models
- `Scripts/` - contains all scripts to reproduce published results
- `Results/` - contains folders to store replication script outputs
  - `Replication-Estimates/` - stores exported estimated quantities
  - `Replication-Figures/` - stores exported figures
  - `Replication-Tables/` - stores exported tables
  - `Published-Models` - stores published models as compressed `.Rdata` files

Replication files are located at:

- GitHub repository
- Harvard Dataverse

Hardware and Software

All models and analysis were conducted using R version 4.0.4 (2021-02-15) running on a PC using Windows 10 x64 (build 19043) with an Intel Core i5-6200U CPU (2.30GHz) and 8Gb of DDR3 memory (1867 MHz).

The primary analysis in this paper was conducted using R-INLA version 21.02.23 compiled on Feb 22, 2021. The tar.gz installation file for this version of INLA is provided here:
- https://inla.r-inla-download.org/R/stable/src/contrib/

- https://inla.r-inla-download.org/R/stable/src/contrib/INLA_21.02.23.tar.gz

The following packages are called in scripts 1-7. Code to install these packages is located in `Scripts/0-master.R`.

Package	**Version**
caret	6.0.86
cowplot	1.1.1
dplyr	1.0.7
forcats	0.5.1
foreach	1.5.1
ggplot2	3.3.3
ggrepel	0.9.1
kableExtra	1.3.4
Matrix	1.3.2
parallel	4.0.4
ProbitSpatial	1.0
pROC	1.17.0.1
purrr	0.3.4
raster	3.4.5
sandwich	3.0.0
scales	1.1.1
sf	0.9.7
sp	1.4.5
spdep	1.1.5
stringr	1.4.0
tibble	3.0.6
tidyr	1.1.3
viridis	0.5.1

Data

All data for this project are stored in the `Data/` subdirectory which contains three files:

- **`/farc_events.Rdata`** contains all data necessary to replicate the analysis. This file contains an r-object called `dat` which is a named list comprising of 4 data frames: `2002-2004`, `2005-2007`, `2008-2009`, and `2002-2009` each corresponding to a temporal aggregation of the data implied by the name. These data frames all contain the following variables:

Variable name	Description
department	Colombia Admin. level-1, Department name
municipality	Colombia Admin. level-2, Municipality name
yr_grp	Year periodization
cinep	FARC activity, count (CINEP source)
cinep_bin	FARC activity, binary-indicator (CINEP source)
icews	FARC activity, count (ICEWS source)
icews_bin	FARC activity, binary-indicator (ICEWS source)
icews_cinep_under	FARC activity, ICEWS underreporting relative to CINEP, binary-indicator
ged	FARC activity, count (GED source)
ged_bin	FARC activity, binary-indicator (GED source)
ged_cinep_under	FARC activity, GED underreporting relative to CINEP, binary-indicator
area_km2	Municipality area
area_km2_ln	Municipality area, logged
distance_bogota_km_ln	Distance (km) between municipality centroid and Bogota, logged
distance_bogota_km	Distance (km) between municipality centroid and Bogota
pop_sum	Population count at municipality level
pop_sum_ln	Population count at municipality level, logged
terrain_ri_mean_m	Terrain roughness indicator, municipality average (meters)
centroid_mun_lat	Municipality centroid, latitude
centroid_mun_long	Municipality centroid, longitude

- **`colombia.Rdata`** - an unprojected simple feature of Colombia's international border used to provide a sharply contrasting boundary in maps containing posterior mean, probability, and standard deviation estimates of the Gaussian random field.
- **`colombia2.Rdata`** - an unprojected simple feature object of Colombia's level-2 administrative boundaries (Municipalities) used for mapping purposes.

Replication results

The R-scripts stored in the `Scripts/` directory generate outputs to produce all figures or tables presented in the main article or appendix. These exports are sored in one of three sub-directories in the `Results/` folder:

- **`Replication-Estimates/`** - contains `.Rdata` files with model estimates used to reproduce all tables and figures
- **`Replication-Figures/`** - contains `.png` files of all figure exports
- **`Replication-Tables/`** - contains outputs used to create all tables in main article and appendix.
- **`Published-Models/`** - contains compressed `.Rdata` files with pre-estimated models from executing code in `2-models-spde.R` or `3-models-spem.R`

R-INLA models with random fields have large file sizes (~130Mb * 5 models * 4 time cuts = ~2.5Gb). To reduce file sizes, after estimating models in the primary `2-models-spde.R` script, quantities of interest necessary for figures or tables are extracted or computed and exported as vectors rather than complete model objects. The compressed full model results are stored as an Rdata object in `Published-Models/`

Scripts

Note - All scripts assume the top-level `Replication/` folder is set as the working directory.

The following tables provide an overview of each script in the `Scripts/` directory. Each table indicates what data sources are imported - either raw data from `Data/` or estimated quantities from `Replication-Estimates/`. Each table also indicates what component of the paper the script exports (either a table or figure) as well as where that element is stored within the Replication directory.

Script Name	**0-master.R**
Purpose	This top-level script contains code to install all packages used in scripts 1-7 as well as code that automates the execution of scripts 1-7.
Runtime	42.6 minutes - total run time for all scripts

Script Name	**1-descriptive.R**
Purpose	Reproduces the descriptive data overview and comparison presented in main paper Section 3.2 including Figures 1 & 2 and Tables 1 & 2. This script also produces Appendix Table 1.
Imports	Data/farc_events.Rdata Data/colombia.Rdata Data/colombia2.Rdata
Exports	Replication-Figures/figure_main_1.png - Figure 1, observed FARC Events map Replication-Figures/figure_main_2.png - Figure 2, selected remote and non-remote municipalities map Replication-Tables/table_main_1.txt - Table 1, full cross-section confusion matrix Replication-Tables/table_main_2.txt - Table 2, selected remote vs. non-remote municipalities confusion matrices Replication-Tables/table_appendix_1.txt - Table A1, selected remote and non-remote municipalities table
Dependencies	dplyr, tidyr, forcats, stringr, ggplot2, ggrepel, sf, purrr, kableExtra, caret
Runtime	20 seconds

Script Name	**2-models-spde.R**
Purpose	Reproduces INLA models with SPDEs reported graphically in the main draft as well as in table format in the Appendix.
Imports	Data/farc_events.Rdata
Exports	Replication-Estimates/parameter-data.Rdata - All model parameter and HPD estimates Replication-Estimates/field-data.Rdata - Projected estimates of Gaussian field mean and SD Replication-Estimates/range-data.Rdata - Spatial field decay and range estimates Replication-Estimates/pred-data.Rdata - Model predicted outcomes on probability scale Published-Models/published-models-spde.Rdata - Compressed INLA model estimates presented in the published article
Dependencies	dplyr, tibble, INLA
Runtime	27 minutes

Script Name	**3-models-spem.R**
Purpose	Reproduces discrete spatial probit error models (spem) presented in Appendix Tables A3 and A4 as well as Figure A3 (spem ROC comparison) and Figure A4 (spem predicted probabilities). This script and associated export also provides estimates for non-spatial probits presented in the Appendix.
Imports	Data/farc_events.Rdata
Exports	Published-Models/published-models-spem.Rdata - Estimated SPEM and non-spatial probit models Replication-Figures/figure_appendix_4.png - Figure A4, SPEM mapped predicted outcome probabilities Replication-Tables/table_appendix_3.txt - Table A3, Probit (non-spatial) 2002-2009 Replication-Tables/table_appendix_4.txt - Table A4, SPEM 2002-2009
Dependencies	dplyr, tidyr, tibble, ggplot2, stringr, purrr, forcats, sf, ProbitSpatial, sandwich, spdep, kableExtra, viridis
Runtime	13 minutes

Script Name	**4-figures_coefplots.R**
Purpose	Reproduces Figures 3 and 4 from the main paper which present INLA model parameter estimates and credibility intervals for included regressors as well as GRMF parameters.
Imports	Replication-Estimates/parameter-data.Rdata
Exports	Replication-Figures/figure_main_3.png - Figure 3, Observed FARC SPDE model estimates Replication-Figures/figure_main_4.png - Figure 4, Underreporting FARC SPDE model estimates
Dependencies	dplyr, tidyr, stringr, forcats, ggplot2, cowplot
Runtime	6 seconds

Script Name	**5-figures-ROCs.R**
Purpose	Produces figure 5 which presents ROCs and AUC estimates comparing ICEWS and GED model performance against CINEP ground truth observations.
Imports	Data/farc_events.Rdata Replication-Estimates/pred-data.Rdata Published-Models/published-models-spem.Rdata
Exports	Replication-Figures/figure_main_5.png - ROC SPDE (Observed FARC) Replication-Figures/figure_appendix_3.png - Figure A3, ROC SPEM (Observed FARC) Replication-Figures/figure_appendix_5.png - Figure A4, ROC SPEM (Underreporting FARC) Replication-Figures/figure_appendix_6.png - Figure A6, ROC SPDE (Underreporting FARC) Replication-Figures/figure_appendix_7.png - Figure A7, ROC SPDE (Observed and Underreporting 2008-2009 period)
Dependencies	dplyr, tidyr, ggplot2, stringr, purrr, pROC, cowplot
Runtime	33 seconds

Script Name	**6-tables.R**
Purpose	Produces Appendix Tables A5-A8 which present INLA SPDE model estimates in table format with median and 95% HPD estimates.
Imports	parameter-data.Rdata
Exports	Replication-Tables/table_appendix_5.txt - Table A5, SPDE 2002-2009 Replication-Tables/table_appendix_6.txt - Table A6, SPDE 2002-2004 Replication-Tables/table_appendix_7.txt - Table A7, SPDE 2005-2007 Replication-Tables/table_appendix_8.txt - Table A8, SPDE 2008-2009
Dependencies	dplyr, tidyr, kableExtra
Runtime	2 seconds

Script Name	**7-figures-field_range-estimates.R**
Purpose	Produces figures related to posterior Gaussian Markov Random Field. Figure 8 - posterior field maps and Figure 9 - Posterior spatial error range and decay.
Imports	Replication-Estimates/field-data.Rdata Replication-Estimates/range-data.Rdata Data/colombia.Rdata
Exports	Replication-Figures/figure_appendix_8.png - Figure 8, SPDE 2002-2009 model GMRF estimate maps Replication-Figures/figure_appendix_9.png - Figure 9, SPDE 2002-2009 model GMRF error correlation range and decay
Dependencies	INLA, sf, dplyr, tidyr, ggplot2, purrr, raster, cowplot, scales
Runtime	1.2 minutes