Language, ethnicity and separatism
ReadMe and Codebook

Kyle L. Marquardt

September 25, 2021

This dataset contains replication materials for "Language, ethnicity and separatism." Specifically, it contains two .R files to replicate all graphics and tables in the article, as well as to run the corresponding analyses; two .RData files containing all relevant variables; nine .txt model files for the Bayesian analyses (in *model.zip*); and 10 .RData files containing posterior draws from the analyses (in *posterior.zip*). Running replication code—excluding the run time for for the Bayesian models—takes approximately 1 minute (OS: Ubuntu 20.04.03, 64 bit); Bayesian models can take multiple hours to run and require four cores. Analyses require JAGS and the following R libraries:

- ggplot2
- runjags
- parallel
- MASS
- rjags
- stargazer

1 Replication materials

1.1 Replication files and data

gagauzia_analyses.R and *pridnestrovie_analyses.R* contain the code necessary to run all Bayesian analyses and replicate all graphics and tables in the article, for the regions of Gagauzia and Pridnestrovie respectively. Note that Bayesian model code is hashed out so as to facilitate quick replication of graphics and tables, as well as to prevent over-writing posterior files. All Gagauzia and Pridnestrovie analyses require the variables in the files *gatudata.RData* and *pmrdata.RData*, respectively; as well as the relevant posterior data files.

1.2 Model files and posterior data files

gF2.txt contains code for the main experimental and observational analyses of Gagauzia, as well as analyses of nonresponse (Appendix K); *gF2.RData* contains posterior draws from these analyses. *pF2.txt* and *pF2.RData* contain equivalent code/data for Pridnestrovie. *gF2_exp.txt* contains code for the experimental analysis robustness check (Appendix J.6) and *gF1_exp.RData* contains posterior draws from this analysis. *pF2_exp.txt* and *pF1_exp.RData* contain equivalent code/data for Pridnestrovie. *gF2r.txt* contains code for the experimental and observational analyses of media consumption in Gagauzia (Appendix M), while *gF2r.RData* contains posterior draws from these analyses. *pF2r.txt* and *pF2r.RData* contain equivalent code/data for Pridnestrovie. *gF2c.txt* contains code for the experimental and observational analyses of latent language proficiency in Gagauzia (Appendix N.2), while *gF1xi.RData* contains posterior draws from these analyses. *pF2c.txt* and *pF1xi.RData* contain equivalent code/data for Pridnestrovie.

lv_sp.txt is the model code for the latent variable analysis of linguistic proficiency (Appendix N.1); the code is the same for analyses of both regions. *gf_sp.RData* contains posterior draws from this analysis from Gagauzia, and *pf_sp.RData* for Pridnestrovie.

2 Codebook

This section describes the variables contained in *gatudata.RData* and *pmrdata.RData*, the respondent-level datasets for Gagauzia and Pridnestrovie, respectively. Note that there is some variation in the ethnic and linguistic variables in the region due to different demographics; in general, there is data for Gagauz in Gagauzia and Ukrainian in Pridnestrovie.

2.1 Outcomes

2.1.1 Experiment

"DV1" represents responses to the experimental question, the respondents' preferred outcome for relations between their region and Moldova. "1" indicates belief that the region should be part of Moldova, without autonomy; "2" that the region should be part of Moldova with autonomy; "3" a confederation with Moldova; and "4" separate statehood. "DV1na" indicates an NA response to this question; "DV1_dk" indicates a "don't know" response.

2.1.2 Observational

All observational responses are on a five-point scale representing level of agreement with statements about potential separatist outcomes. The scale ranges from "1," "Fully disagree," to "5," "Fully agree." "S2d" represents responses to the statement "Gagauzia/Pridnestrovie should be an independent state;" "S2a" the statement "Gagauzia/Pridnestrovie should be a region with no special status within Moldova;" "S2b" the statement "Gagauzia/Pridnestrovie should have autonomous status within Moldova;" and "S2c" the statement "Gagauzia/Pridnestrovie should have confederal status with Moldova." "R1a-c" replace "Moldova" in "S2a-c" with "Russia" and "U1a-c" with "Ukraine" (Ukraine questions only included

in Pridnestrovie survey). As with the experimental outcome, the suffix "na" indicates NA responses to a given question, while "_dk" indicates "don't know" responses.

2.2 Experimental condition

"tg" indicates a respondent's assigned experimental condition. Control is self-explanatory. "T1" represents $T_{Moldovan}$, "T2" represents $T_{Russian}$ and "T3" represents $T_{Gagauz/Ukrainian}$.

2.3 Language

2.3.1 Dichotomous spoken fluency

"rf," "mf," "gf" and "uf" are the dichotomous indicators of self-reported spoken fluency in Russian, Moldovan, Gagauz and Ukrainian. "rf" indicates a *lack of* spoken fluency, while the remaining indicators indicate fluency.

2.3.2 "Can do" questions

Dichotomous indicators of tasks respondents self-report being able to perform. "sr1," "sr2" and "sr3" indicate an ability to respectively 1) introduce oneself, 2) talk about day with friends, and 3) carry on in-depth discussions about complicated issues in Russian; "sm1," "sm2" and "sm3" the same for Moldovan; "sg1," "sg2" and "sg3" for Gagauz; and "su1," "su2" and "su3" for Ukrainian. "ur1," "ur2," "ur3" and "ur4" indicate an ability to respectively 1) understand basic directions about locations, 2) understand basic discussions, 3) understand people speaking quickly with slang and 4) understand tv and radio programs on unfamiliar topics in Russian; "um1," "um2," "um3" and "um4" the same for Moldovan; "'ug1," "ug2," "ug3" and "ug4" for Gagauz; and "uu1," "uu2," "uu3" and "uu4" for Ukrainian.

2.3.3 Spoken fluency scale

"rsf" provides a five-point scale of self-reported spoken proficiency in Russian, "msf" in Moldovan, "gsf" in Gagauz and "usf" in Ukrainian. "1" corresponds to "I cannot speak at all;" "2" "I can speak with extreme difficulty;" "3" "I speak it with difficulty;" "4" "I am not fluent, but speak it with ease;" and "5" "I am fluent."

2.3.4 Spoken comprehension scale

"ruf" provides a five-point scale of self-reported spoken comprehension in Russian, "muf" in Moldovan, "guf" in Gagauz and "uuf" in Ukrainian. "1" corresponds to "I do not understand anything;" "2" "I understand with extreme difficulty;" "3" "I understand with difficulty;" "4" "I understand a good deal with ease;" and "5" "I understand almost everything with ease."

2.3.5 Age of acquisition

"aa_r" reports the age at which a respondent began learning the Russian language, "aa_m" the Moldovan language, "aa_g" the Gagauz language, and "aa_u" the Ukrainian language.

2.3.6 Latent spoken proficiency

"xiR" is the posterior median estimate of a respondent's latent spoken proficieny in Russian, "xiM" in Moldovan, "xiG" in Gagauz, and "xiU" in Ukrainian.

2.3.7 Reason for learning languages

"reason_r" reports the reason a respondent learned the Russian language. If the respondent learned Russian due to personal desire, "preason_r" further specifies the reason. If the respondent learned Russian because of a parental or family decision, "freason_r" further specifies the reason. "reason_m," "preason_m" and "freason_m" provide the same information for the Moldovan language; "reason_g," "preason_g" and "freason_g" for Gagauz; and "reason_u," "preason_u" and "freason_u" for Ukrainian.

2.3.8 Perceived government support for languages

"ggs_r" reports the level at which a respondent believes the Gagauz government supports the Russian language ("1" represents "Completely opposes" and "5" "Completely supports"), "ggs_m" the Moldovan language, and "ggs_g" the Gagauz language. "pgs_r," "pgs_m" and "pgs_u" reports the same information, but regarding the Pridnestrovien government (with Ukrainian replacing Gagauz). "mgs_r," "mgs_m" and "mgs_g/u" reports this information regarding the Moldovan government (with Gagauz in Gagauzia data and Ukrainian in Pridnestrovie data).

2.3.9 Perceived frequency of language use

"glf_r" reports how frequently a respondent believes the Russian language is used in Gagauzia ("1" represents "Almost never" and "5" "Almost all the time"), "glf_m" the Moldovan language, and "glf_g" the Gagauz language. "plf_r," "plf_m" and "plf_u" reports the same information, but regarding language use in Pridnestrovie (with Ukrainian replacing Gagauz). "mlf_r," "mlf_m" and "mlf_g/u" reports this information regarding language use in Moldova proper (with Gagauz in Gagauzia data and Ukrainian in Pridnestrovie data).

2.4 Ethnic identity

2.4.1 Dichotomous indicators

"Russian," "Moldovan," "Ukrainian," "Gagauz," and "Bulgarian" are dichotomous indicators for individuals who primarily identify with the relevant ethnic group. "Other" indicators respondents who identify with a different, non-specified group. "NGagauz" and "NRussian" represent respondents who do not primarily identify as Gagauz or Russian, respectively.

2.4.2 Ethnic membership indicators

"N3a" indicates if a respondent considers a mother's nationality important for membership in their ethnic group (note: 0 represents either a negative response or NA), "N3b" their father's nationality, "N3c" following their nationality's cultural traditions, "N3d" knowledge

of their national language, "N3e" belief in the religion of their group, "N3f" their physical appearance, "N3g" their self-identification, "N3h" their location of residence, "N3i" their citizenship, "N3j" a common history with other members of their nationality, "N3k" following the professional traditions of their nationality, "N3l" marrying someone of the same nationality, and "N3m" educating their children in the language of their nationality.

2.5 Demographics

"male" is a dichotomous indicator for respondents who identify as male. "age" is log-transformed respondent age. "highered" is a dichomous indicator for respondents who reported some experience with higher education. "urban" is an indicator for respondents who live in an urban locality. "married" is a dichotomous indicator for respondents who report being married. "prof" is a categorical variable representing the respondent's current form of employment. "resp" is a continuous variable representing the number of viable respondents in a household, while "house" represents the total number of household residents.

2.6 Income

Household income is estimated within each Bayesian model using a variety of demographic covariates. It represents the natural logarithm of estimated household income, divided by the number of household residents; estimated household income is restricted to be within the bounds of the self-reported minimum and maximum household income ("ymin" and "ymax"). "income" provides a rough estimate of household income by using "ymin" and "ymax" in place of model-estimated income.

2.7 Locality

Locality in which a respondent resides.

2.8 Media

"rf_m," "mf_m," "gf_m" and "uf_m" are dichotomous indicators of the media which a respondent consumes most frequently: Russian, Moldovan, Gagauz or Ukrainian. "rf_m" indicates a respondent who does *not* consume Russian media most frequently, while the remaining indicators indicate most frequent consumption.

2.9 Time in Russia

"Russia" reports the amount of time a respondent had spent in Russia (in years).