Imputing Race/Ethnicity Data with Bayesian Improved Surname Geocoding (BISG)

Eric Moore and Kailas Venkitasubramanian

Lunch and Learn Series
UNC Charlotte Urban Institute
January 24, 2023

The Problem

- Questions of Interest
- Personal Identification Information (PII), and its limitations
- One possible remedy

Bayesian Improved Surname Geocoding

- What does the method do?
  - Describing the probability of an event occurring
  - Combines surname analysis with demographic counts
- How has it been used empirically?
- The purpose of this presentation

Using BISG: The Basic Code

The first thing to do is to acquire the census data:

```
ncCensusData <- get_census_data(
key = "020934bc9b3faa8c9f629492650c8eb446349f2d",
states = "NC", census.geo = "county")
```

Using BISG: The Basic Code

Now we can run the predict_race function:

```
predict_race(voter.file = ncvoter,
     census.geo = "county",
     census.data = ncCensusData)
```

Using BISG: Output

pred.whi	pred.bla	pred.his	pred.asi	pred.oth
0.849	0.026	0.105	0.002	0.017
0.519	0.437	0.005	0.002	0.036
0.887	0.079	0.01	0.003	0.021
0.937	0.037	0.005	0.004	0.017
0.059	0.772	0.026	0.101	0.042
0.953	0.037	0.001	0	0.009
0.072	0.89	0	0	0.038
0.228	0.736	0	0	0.036
0.683	0.001	0	0.001	0.315
0.585	0.399	0.001	0.001	0.014

Validating BISG Predictions

- Going back to the North Carolina voter file
- Filtering self-reported race and ethnicity
- Question of determining an individual's race/ethnicity

Using BISG: The Final Code

```
predict_race(voter.file = ncvoter,
             surname.year = 2020,
             census.geo = "county",
             census.data = ncCensusData,
             year = "2020",
             model = "BISG",
             names.to.use = "surname, first, middle")

mutate(predict_race = colnames(ncpredict %>%
       dplyr::select(pred.whi:pred.oth) )
       [max.col(ncpredict %>%
       dplyr::select(pred.whi:pred.oth),
       ties.method = "random")],
          prace_new = recode(predict_race,
                      "pred.whi" = "W",
                      "pred.bla" = "B",
                      "pred.his" = "HL",
                      "pred.asi" = "A",
                      "pred.oth" = "O"))
```

Validating BISG Predictions

	Surname (76%)		Sur+First (83%)		S+F+Mid (87%)	
	Acc.	Diff.	Acc.	Diff.	Acc.	Diff.
Asian	61.5	-14.4	74.4	-8.64	81.3	-5.73
Black	54.5	-21.4	65.1	-17.9	70.9	-16.1
Hispanic	72.9	-2.96	76	-6.98	78.1	-8.93
Other	15.8	-60.1	21.1	-61.9	28.5	-58.5
White	84.9	9.03	91.2	8.16	94	6.95

Final Thoughts

- BISG as a viable research tool, though the drawbacks must be acknowledged
- New extensions and modifications

Applications for UI

- Studies with unreliable or missing race/ethnicity data
- Augmenting Data Integration Process

Guidelines for Practice

- Privacy Policy of Client/Partner Organization
- Informed Consent
- Rare populations and Groups
- Choosing when to collect race data
- Reporting imputation and constraints
- Collecting name and address information
- Reidentification Risks

Next Steps

- Formalizing the practice
- Documentation and Resources
- Community and partner feedback