Supplementary Materials for Correcting Sociodemographic Selection Biases for Population Prediction from Social Media

Salvatore Giorgi,[1] Veronica E. Lynn,[2] Keshav Gupta,[2] Farhan Ahmed,[2]
Sandra Matz,[3] Lyle H. Ungar,[1] H. Andrew Schwartz[2]
[1] University of Pennsylvania
[2] Stony Brook University
[3] Columbia University
sgiorgi@sas.upenn.edu, has@cs.stonybrook.edu

Cross Correlations

The correlations between our four health outcomes and four socio-demographic variables are presented in Table 1. When correcting for selection biases it is not always clear which biases exist in your sample. While this paper was limited to age, gender, income, and education, there exist many other variables one could correct for. To this end, we present the correlations between all variables used in this paper, in order to interpret our results — given these relationships, would we expect correcting for certain socio-demographic variables to increase predictive performance on a specific outcome? The correlations in Table 1 show that both income and education are highly associated with all of our outcome variables. On the other hand, age and gender are not, with the exception of suicide. Given the size of the income correlations we might expect correcting for income to give us the biggest benefit.

	Suicide	Life Satisfaction	Fair/Poor Health	Income	Education	Percent Female	Median Age
Heart Disease	.18	-.34	.60	-.58	-.59	.09	.03
Suicide	-	-.04	.20	-.31	-.34	-.15	.30
Life Satisfaction		-	-.35	.37	.39	-.06	-.05
Fair/Poor Health			-	-.65	-.62	.06	.02

Table 1: Correlations between county level health and socio-demographic variables

Method Examples

Here we give brief examples of our more complex methods to aid in understanding.

Estimator Redistribution Here we redistribute our estimated socio-demographics at the national level (i.e., across all counties) to match the national distribution reported by PEW. (See **Data** for a description of the PEW data.) The national percentage of people on Twitter between $\min_h^{(t)} = 18$ and $\max_h^{(t)} = 29$ is 51.1% (as reported by PEW's Social Media update (Duggan et al. 2015; Greenwood, Perrin, and Duggan 2016); averaged across 2013-2016). We then start with our minimum predicted age in our sample $\min_h^{(s)} = 13$ and then find the age $\max_h^{(s)}$ such that the percentage of Twitter users in our sample between 13 and $\max_h^{(s)}$ equals 51.1%. We then adjust all age estimates with that bin using Equation 6. Next, we set $\min_h^{(t)} = 30$ and $\max_h^{(t)} = 49$ and note that the PEW reported national percentage of people on Twitter in this age bin is 31.8%. We then set $\min_h^{(s)}$ equal to $\max^{(s)}$ from the previous iteration. Finally, we find the age $\max_h^{(s)}$ such that the percentage of Twitter users in our sample between $\min_h^{(s)}$ and $\max_h^{(s)}$ equals 31.8%. This process is repeated for all bins.

Adaptive Binning In adaptive binning we set a minimum number of observations per bin and collapse all bins with the smallest adjacent bin if they do not meet this threshold. This is repeated until all bins meet the threshold or we have a single bin. For example, in a given county we have at most 11 age bins. We start with the bin with the smallest number of Twitter users and see if this number meets our minimum threshold. In this example lets define our minimum bin threshold as 50 and assume the age bin with the smallest number of Twitter users mapped to our county is 45-49 years old. We check if there are at least 50 users who are between 45 and 49 years old. If not, we combine this bin with the smallest, adjacent bin (either 40-44 or 50-54). Assume that the bin 40-44 has less users than 50-54. We combine then combine the bins 40-44 and 45-49, resulting in the bin 40-49, and discard the bins 40-44 and 45-49. We then start the process over: identify the smallest bin and repeat the above steps until all bins meet our threshold or we have a single bin.

Additional Experiments

Add One Smoothing We examine the effects of "add one" smoothing in Table 2. Here we add 1 to each socio-demographic bin. We see that the prediction accuracies are comparable to binning, with a minimum bin size of 1, and smoothing with $k = 1$, though this method fails to increase performance over baseline.

References

Duggan, M.; Ellison, N. B.; Lampe, C.; Lenhart, A.; and Madden, M. 2015. Social media update 2014. *Pew research center*, 19.

	Post-stratification	Naive Post-Statification	Raking
Age	.587	-	-
Gender	.641	-	-
Income	.617	-	-
Education	.644	-	-
Age + Gender	-	.590	.607
Income + Education	-	.629	.630
Age + Gen. + Inc. + Edu.	-	.640	.628

Table 2: Evaluation of “add one” Smoothing. Results are comparable to adaptive binning with bin size = 1 and informed smoothing with k = 1, but no increase above baseline.

Greenwood, S.; Perrin, A.; and Duggan, M. 2016. Social media update 2016: Facebook usage and engagement is on the rise, while adoption of other platforms holds steady. *Pew Research Center*.