Modeling Traffic Stop Demographics with Population Estimates

Presented by: Chad Kite

December 7th, 2023

Problem

According to the Stanford Open Policing Project (SOPP), there are over 50,000 traffic stops per day in the U.S.

There is no federal requirement to publicly report or disclose data related to these stops, and most states do not have one either, so we do not even know the exact number of stops, much less what groups, if any, get stopped more or less often than others.

Given the lack of complete data, how can we estimate the demographics of the population of traffic stops?

Opportunity

The SOPP has collected over 200,000,000 traffic stop records.

Voluntary submissions means they are not standardized and not a random sample

How can we build a model that uses data available for the entire United States to predict the demographics of these traffic stop records?

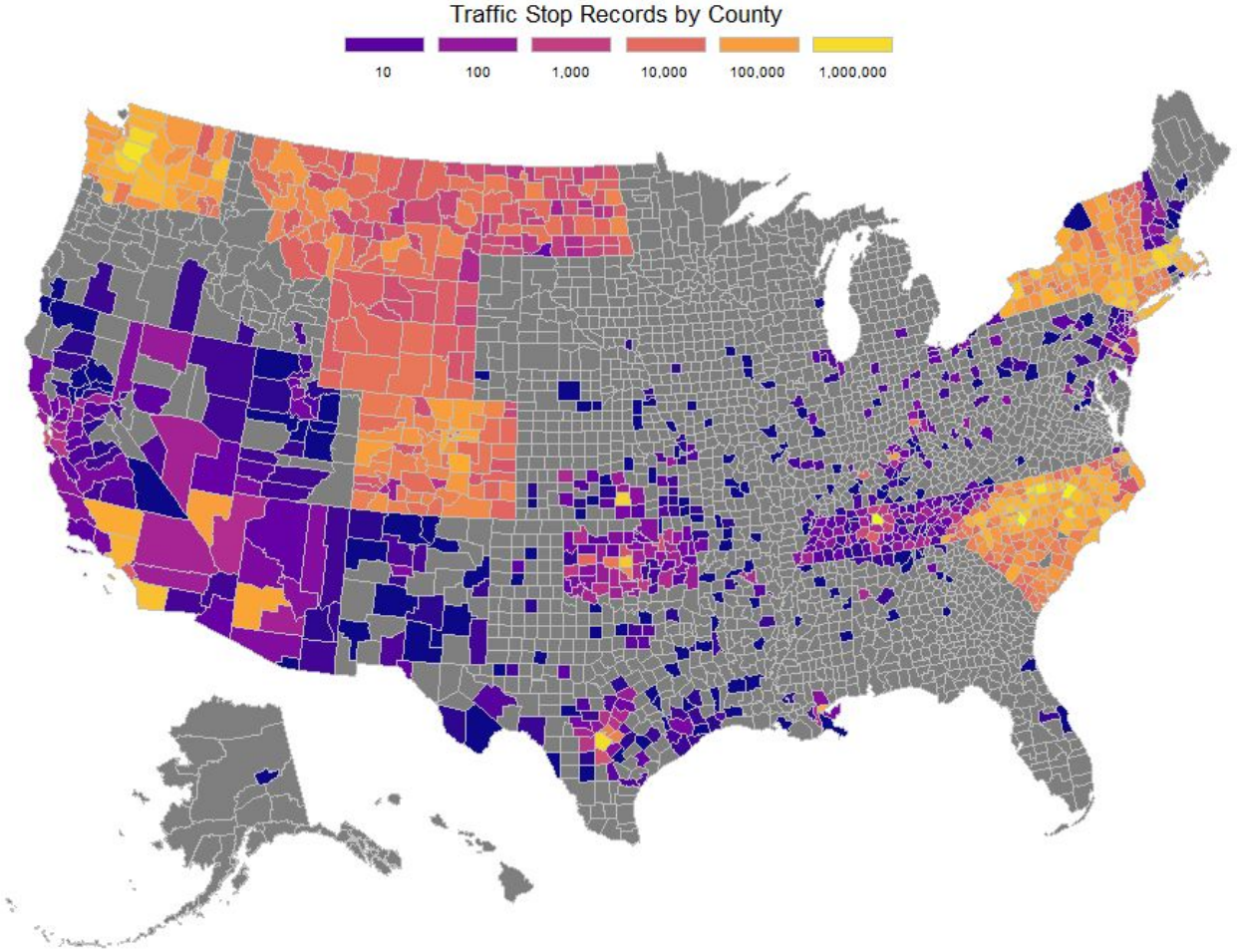
Proposal

Model traffic stop demographics using population demographics from the U.S. Census Bureau

Data

SOPP Traffic Stop Data[1]

- Records containing the date, usable location data, and the age, race and sex of the subject from 30 departments in 23 states
- These records represent 52,644,187 traffic stops
- Age is an integer
- Race: Asian or Pacific Islander, Black, Hispanic, Other, or White
- Sex: Female or Male


Traffic Stop Records by County
10
100
1,000
10,000
100,000
1,000,000

Data

U.S. Census Estimates[2]

- Available at the county level for every year from 1990 to 2020
- Contain population by age, race, sex and origin
- Age is binned: 0, 1-4, 5-9, 10-14, … , 75-79, 80-84, 85+
- Race: American Indian/Alaska Native, Asian or Pacific Islander, Black, White
- Sex: Female or Male
- Origin: Hispanic or Non-Hispanic

Data Processing

Traffic Stop Records

- Dropped records with NAs in required fields
- Dropped records representing pedestrian stops
- Rounded latitude and longitude to 4 digits
- Simplified date to year
- Binned subject age: 0-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81+
- Matched records by county name or latitude and longitude to Federal Information Processing System county code
- Aggregated traffic stops by year and county

Data Processing

Census Estimates

- Converted numeric representations of demographic data into factors with appropriate labels
- Converted records with Origin = Hispanic to Race = Hispanic to match traffic stop records
- Combined age bins to match traffic stop records: 0-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81+
- Created min-max and z-score normalized versions of data

Exploratory Data Analysis

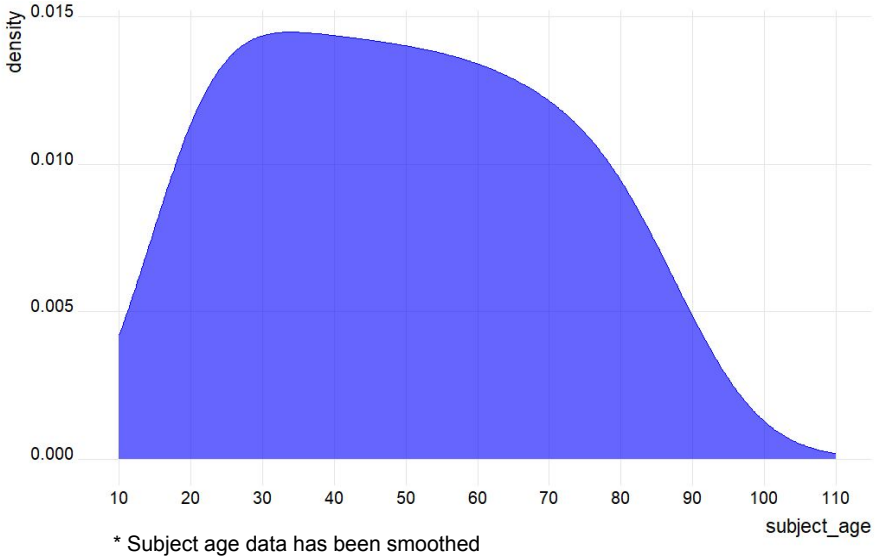
Traffic Stops by Year of Stop

~61% in 2010-2015

Year	# of Stops	% of Stops
2000	252,343	0.48%
2001	570,413	1.08%
2002	1,287,163	2.45%
2003	1,071,016	2.03%
2004	992,787	1.89%
2005	995,976	1.89%
2006	1,290,312	2.45%
2007	2,117,018	4.02%
2008	2,270,066	4.31%
2009	3,050,840	5.80%
2010	5,223,562	9.92%
2011	5,272,328	10.00%
2012	5,370,608	10.20%
2013	5,188,463	9.86%
2014	5,489,812	10.40%
2015	5,459,726	10.40%
2016	2,897,259	5.50%
2017	2,640,829	5.02%
2018	987,398	1.88%
2019	169,517	0.32%
2020	46,751	0.09%

Exploratory Data Analysis

Traffic Stops by Subject Age, Race, and Sex



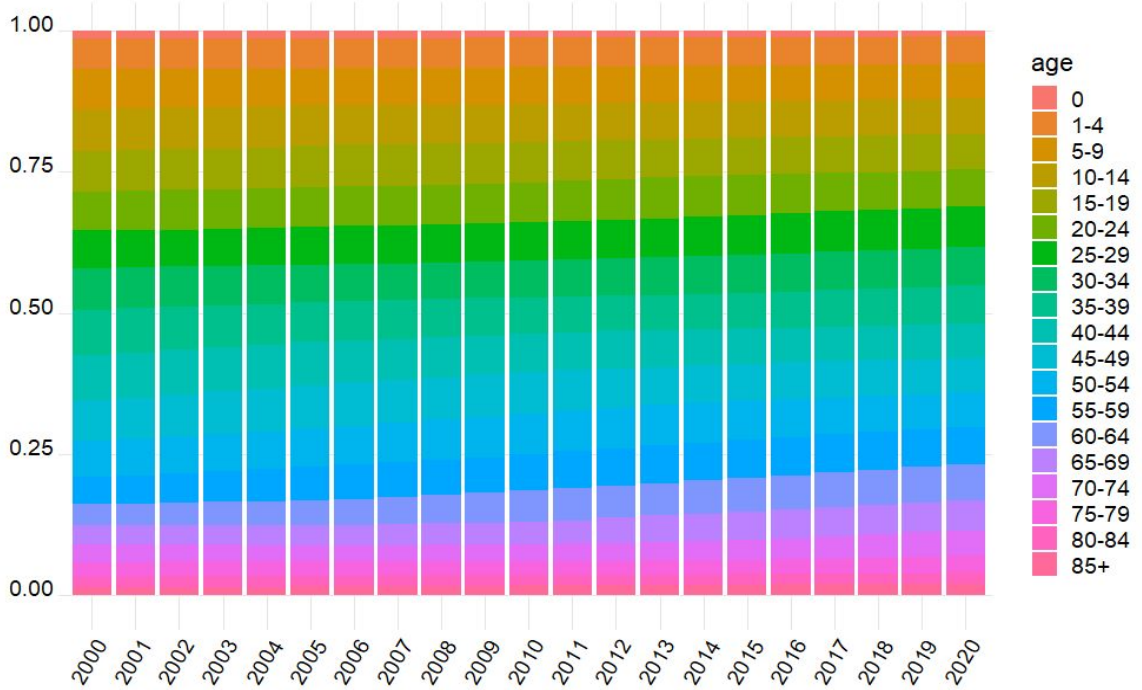
* Subject age data has been smoothed

Subject Race	# of Stops	% of Stops
Asian/Pacific Islander	1,500,866	2.85%
Black	11,183,670	21.24%
Hispanic	4,976,955	9.45%
White	34,084,360	64.74%
Other	898,336	1.70%

Subject Sex	# of Stops	% of Stops
Female	18,074,879	34.33%
Male	34,569,308	65.67%

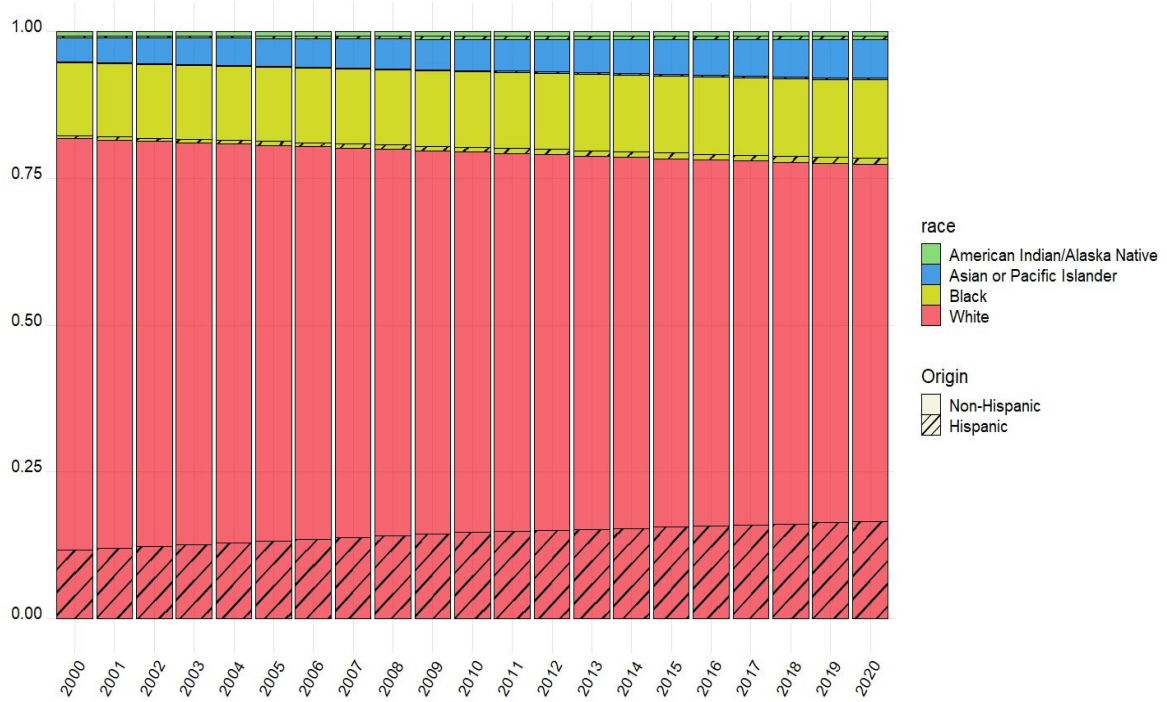
Exploratory Data Analysis

Census Estimates by Age and Year



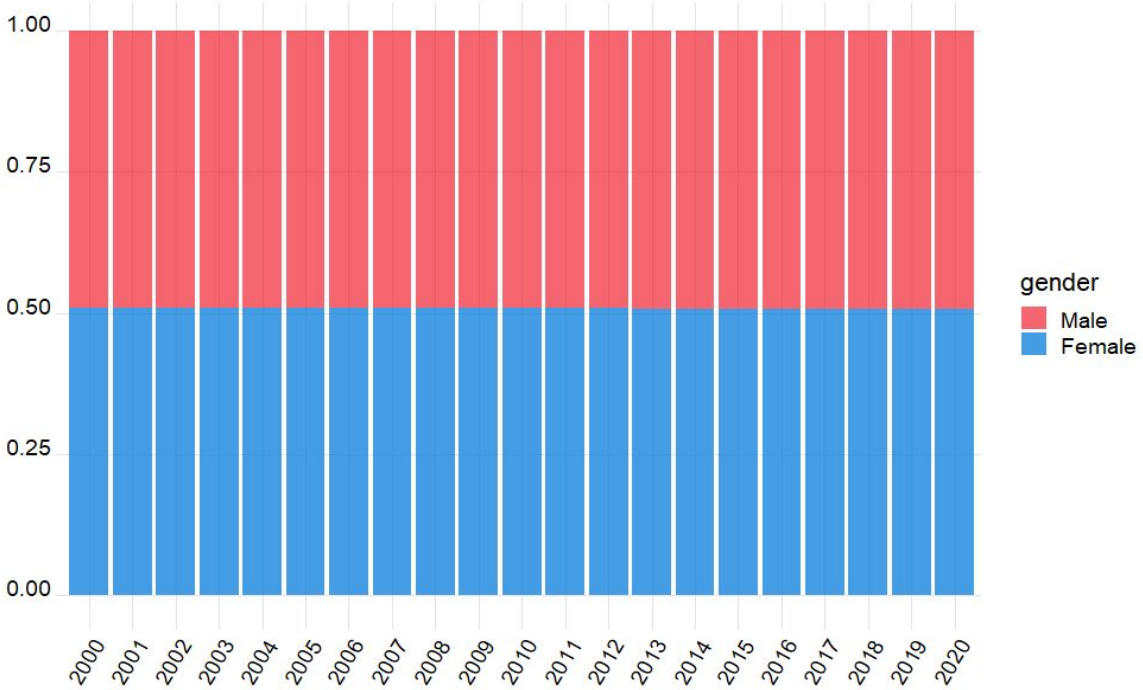
Exploratory Data Analysis

Census Estimates by Race and Year



Exploratory Data Analysis

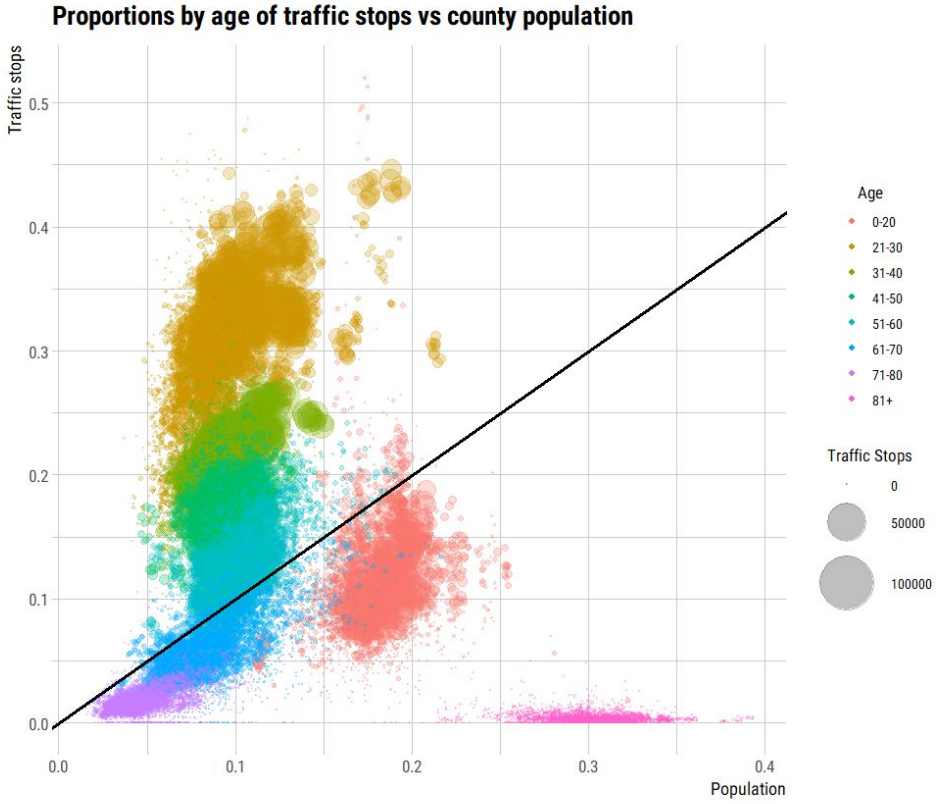
Census Estimates by Sex and Year



Demographic Comparison

Comparison of subject age to population demographics by county:

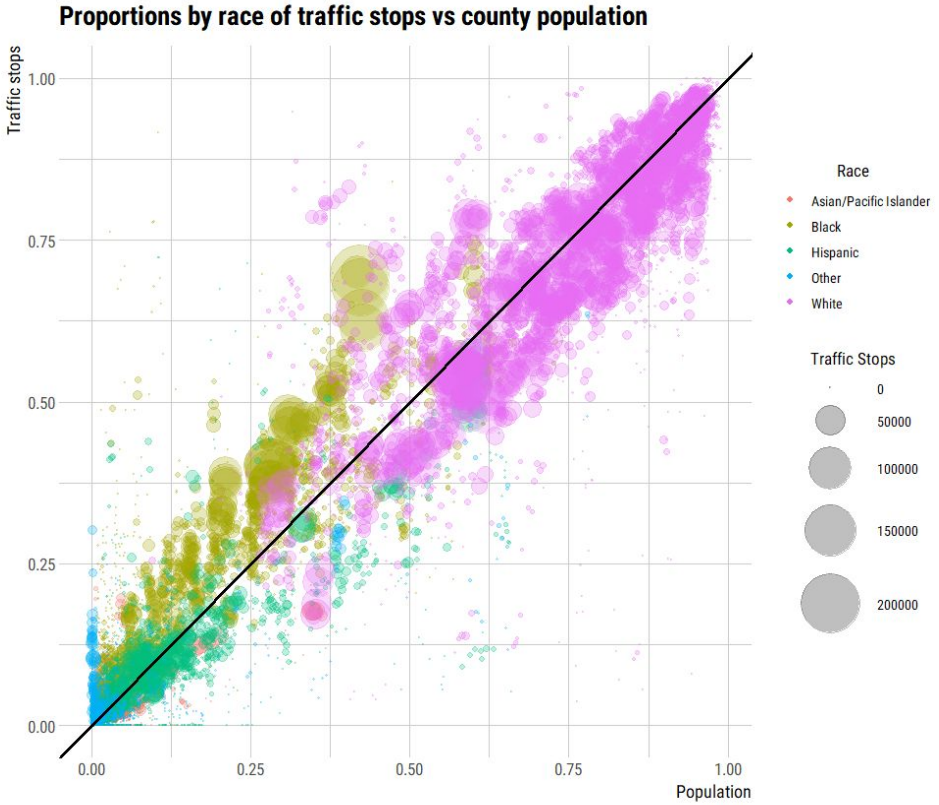
- Traffic stops are clearly delineated by age group.
- Higher age groups have fewer stops despite similar representation in county population demographics.
- Outliers are 0-20 and 81+ age groups, both of which include large portions of the population that do not drive.



Demographic Comparison

Comparison of subject race to population demographics by county:

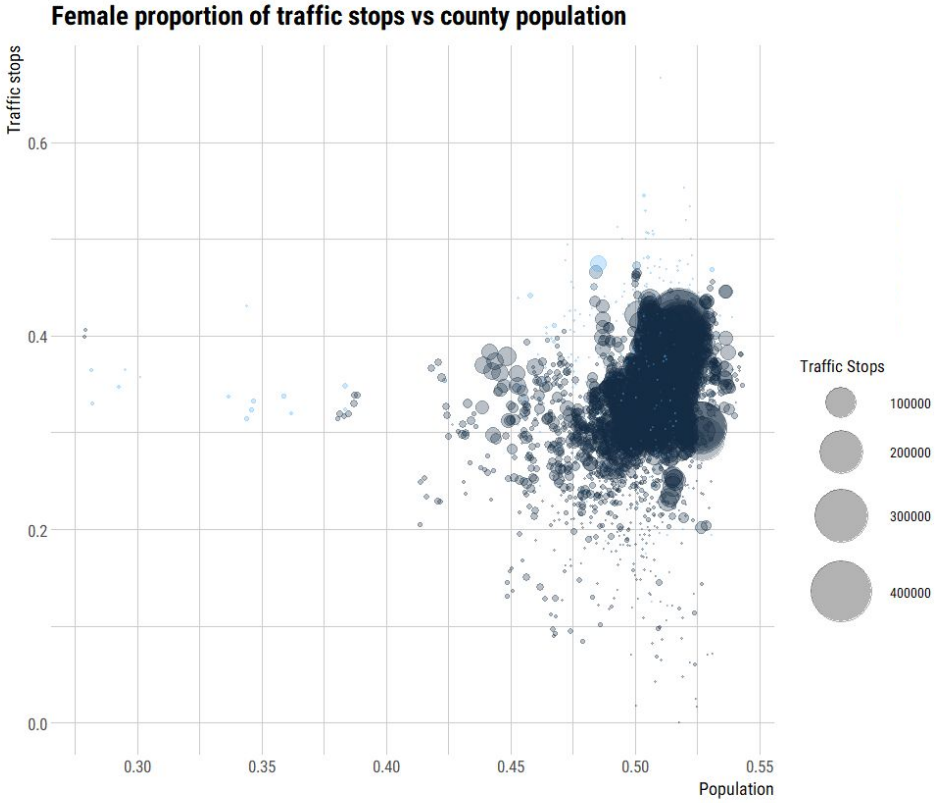
- Ratios of traffic stops by race to county population demographics are generally consistent.
- Subjects recorded as Black are clearly stopped at a higher rate than their representation in county population demographics.



Demographic Comparison

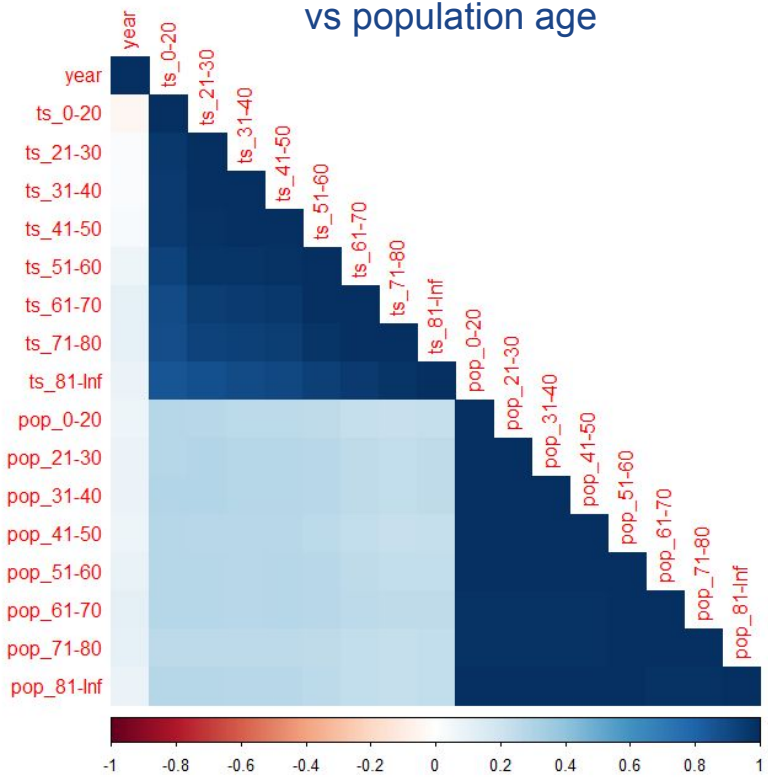
Comparison of subject sex to population demographics by county:

- Ratios of traffic stops by sex to county population demographics are very consistent.
- Subjects recorded as Female are stopped at a lower rate than their representation in county population demographics

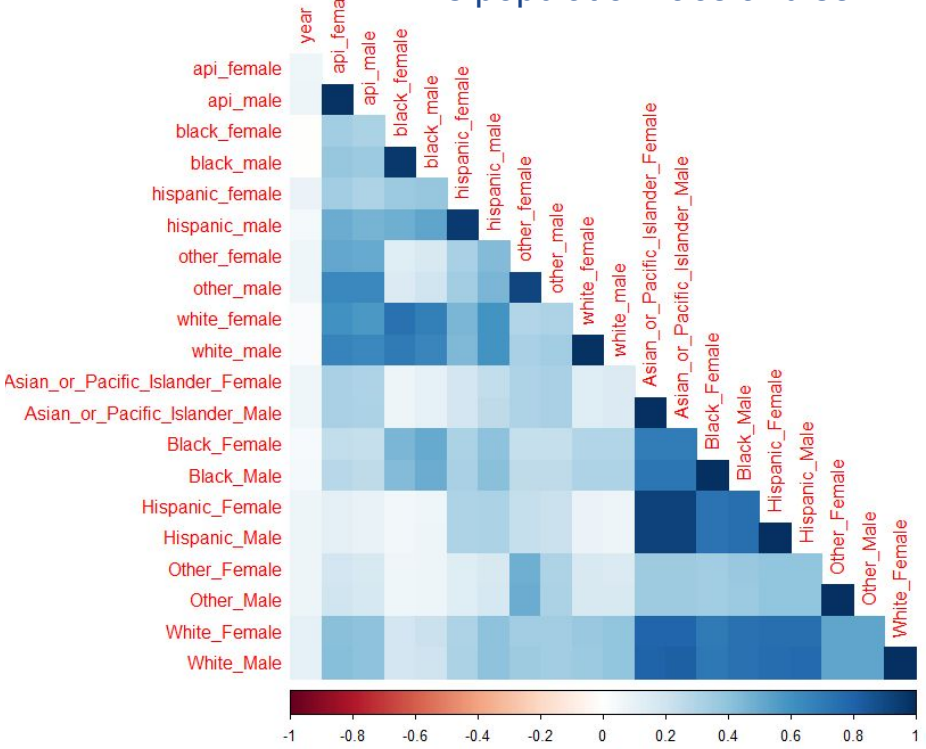


Correlation

Traffic stop subject age vs population age



Traffic stop subject race and sex vs population race and sex



Exploratory Data Analysis

- 20 years, 1083 counties, and 5436 county-years
- χ^2 test for goodness of fit:
    - None of the subject demographic distributions are the same as their respective population demographic distribution
    - The distribution of subject sex differs by subject race
    - The distribution of subject age is approximately the same for each combination of subject race and subject sex

Splitting and Merging the Data

Randomly selected 200 of the 1083 counties represented in the data for the test set and set aside all years for those counties

Aggregated traffic stops in training set and population demographics 3 different ways

- by year, county, age, race and sex
- by year, county, race and sex
- by year, county and age

Merged training and test sets with their corresponding normalized and unnormalized U.S. Census estimates

Assessing Model Performance

Predictors for each county-year are identical for every observation, so predicting the correct class for each observation is not relevant.

County-level Accuracy

1. Multiply predicted proportions by number of stops in a county-year
2. Sum correct predictions
3. Divide by number of stops in the county-year

Example 3 classes with frequencies 23, 24, and 53, for which the predicted probabilities are .2, .3, and .5.

Predicted probabilities	*	# of stops	=	Predicted stops		min(Pred. stops, stops)
.2		100		20		min(20, 23) = 20
.3		100		30		min(30, 24) = 24
.5		100		40		min(40, 53) = 50
						94

County-level accuracy: 94 / 100 = 94%

Modeling Subject Age

Multinomial Logistic Regression

- Base model not significantly better than null model
- Normalization ➡ small improvement

- Best model found by stepwise selection in both directions:
  response ~ year + age 61-70 using z-score normalized predictors

Normalization	AIC	χ^2 vs null model	p-value
Unnormalized	12274.03	12.85	1.0
Min-max	12237.26	49.61	0.8905
Z-score	12237.53	49.34	0.8957

Normalization	AIC	χ^2 vs null model	p-value
Min-max	12148.84	40.04 (14 df)	0.00025
Z-score	12144.75	44.13 (14 df)	0.00005

County-level accuracy: 93.90%

Modeling Subject Age

Multinomial Logistic Regression

- Base model not significantly better than null model
- Normalization ➡ small improvement

- Best model found by stepwise selection in both directions:
  response ~ year + age 61-70 using z-score normalized predictors

Response	(Intercept)	year	pop_61-70
stops_per_21-30	-52.1367	0.0264	0.0175
stops_per_31-40	-65.9730	0.0331	0.0074
stops_per_41-50	-93.5460	0.0467	-0.0080
stops_per_51-60	-157.3465	0.0783	-0.0337
stops_per_61-70	-184.4481	0.0914	-0.0722
stops_per_71-80	-187.3894	0.0923	-0.0994
stops_per_81-Inf	-62.3742	0.0293	-0.0308

County-level accuracy: 93.90%

Modeling Subject Age

Linear Discriminant Analysis

- Required the conversion of aggregated data back to individual observations

	year	`pop_0-20`	`pop_21-30`	`pop_31-40`	`pop_41-50`	`pop_51-60`	`pop_61-70`	`pop_71-80`	`pop_81-Inf`	name	count
1	2000	0.0120	0.0114	0.0126	0.0132	0.0114	0.00995	0.0142	0.0124	0-20	1180
2	2000	0.0120	0.0114	0.0126	0.0132	0.0114	0.00995	0.0142	0.0124	0-20	1180
3	2000	0.0120	0.0114	0.0126	0.0132	0.0114	0.00995	0.0142	0.0124	0-20	1180
4	2000	0.0120	0.0114	0.0126	0.0132	0.0114	0.00995	0.0142	0.0124	0-20	1180
5	2000	0.0120	0.0114	0.0126	0.0132	0.0114	0.00995	0.0142	0.0124	0-20	1180

- Normalization ➡ small improvement
- No difference between methods of normalization

Normalization	County-level accuracy
Unnormalized	94.14%
Min-max	95.09%
Z-score	95.09%

Modeling Subject Age

Linear Discriminant Analysis

Predictor	LD1	LD2	LD3	LD4	LD5	LD6	LD7
year	-0.2691	0.0470	0.0045	-0.0052	0.0471	0.0366	0.0168
`pop_0-20`	71.0608	469.9542	-163.8615	-239.0605	-294.3471	175.3673	317.8429
`pop_21-30`	9.5304	314.9625	-195.6983	-53.7806	-33.9351	111.6845	289.4849
`pop_31-40`	0.2971	-121.8927	93.4058	-127.0440	-147.8592	-31.1660	12.5604
`pop_41-50`	-17.9366	137.9225	9.6168	81.4872	316.1810	81.8148	-40.8422
`pop_51-60`	-20.2419	-75.5018	16.2295	-0.1043	-380.6436	-42.1984	172.5459
`pop_61-70`	24.8779	30.8674	3.0206	27.3932	227.6167	-52.7127	-213.2455
`pop_71-80`	-4.2586	63.4026	-38.3057	-50.0963	-107.8974	128.1945	139.5816
`pop_81-Inf`	-53.3224	-789.2829	263.4058	336.5066	406.9096	-339.5263	-664.0746

Proportion of Trace:	LD1	LD2	LD3	LD4	LD5	LD6	LD7
	0.7077	0.2020	0.0515	0.0349	0.0027	0.0010	0.0002

County-level accuracy: 95.09%

Modeling Subject Age

Extreme Gradient Boosting

- Unable to process entire training set as observations
  - Using downsampled data produced predictions too inaccurate to consider
  - Using random sample of 20% of observations to build model produced a viable model that could be processed quickly enough to tune hyperparameters
- Normalization ➡ no effect

Grid search hyperparameters

max.depth = {4, 6, 8}
eta = {.2, .3, .4}
nrounds = {5, 8, 10}

max.depth	eta	nrounds	County-level accuracy
4	.4	10	92.66%
6	.4	10	92.17%
4	.4	8	91.93%
8	.4	10	91.91%
4	.3	10	91.66%

County-level accuracy: 92.66%

K-fold Cross-Validation for Subject Age

Evaluated all 3 models with k=10

	MNLR	LDA	XGB
Mean county-level accuracy	94.21%	94.99%	94.94%

The LDA model produced the highest average county-level accuracy.

Used a random sample of 40% of the observations available for model building on each iteration to train the boosted model, which had a significant effect on model performance. The ability to use even more of the data to train the model could result in further improvement.

County-level accuracy: 94.99%

Modeling Subject Race and Sex

Multinomial Logistic Regression

- All 3 types of data produced results significantly better than the null model
- Model built on z-score normalized data slightly outperformed the others

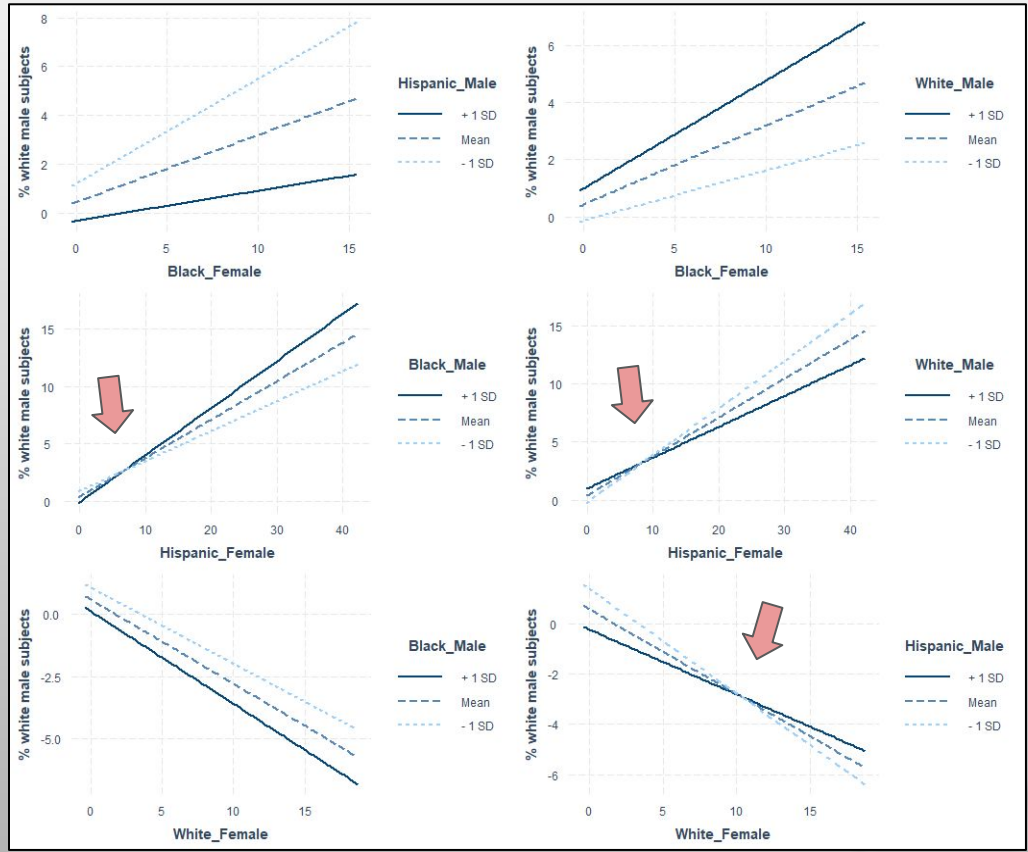
Normalization	AIC	χ^2 vs null model	p-value
Unnormalized	11398.01	346.1 (99 df)	0
Min-max	11401.76	342.35 (99 df)	0
Z-score	11397.89	346.2 (99 df)	0

Modeling Subject Race and Sex

Multinomial Logistic Regression

Checked for interactions by building linear models with two predictors

- Limited search space by testing across race and sex for the three most common subject races in the data: Black, Hispanic, and White.
- For the dependent variable in each round of inspection, selected response variable with highest variance that had not yet been used



Potential Interactions

Hispanic Female x Black Male

Hispanic Female x White Male

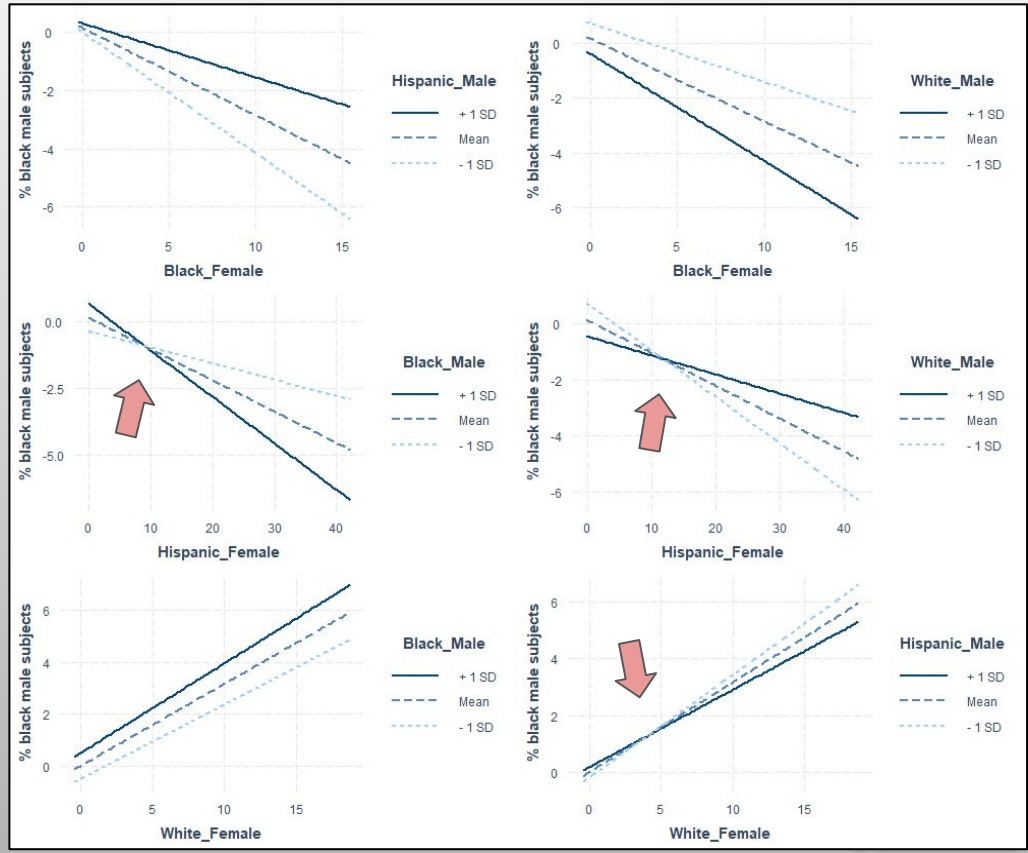
White Female x Hispanic Male

Modeling Subject Race and Sex

Multinomial Logistic Regression

Checked for interactions by building linear models with two predictors

- Limited search space by testing across race and sex for the three most common subject races in the data: Black, Hispanic, and White.
- For the dependent variable in each round of inspection, selected response variable with highest variance that had not yet been used



Potential Interactions
Hispanic Female x Black Male
Hispanic Female x White Male
White Female x Hispanic Male

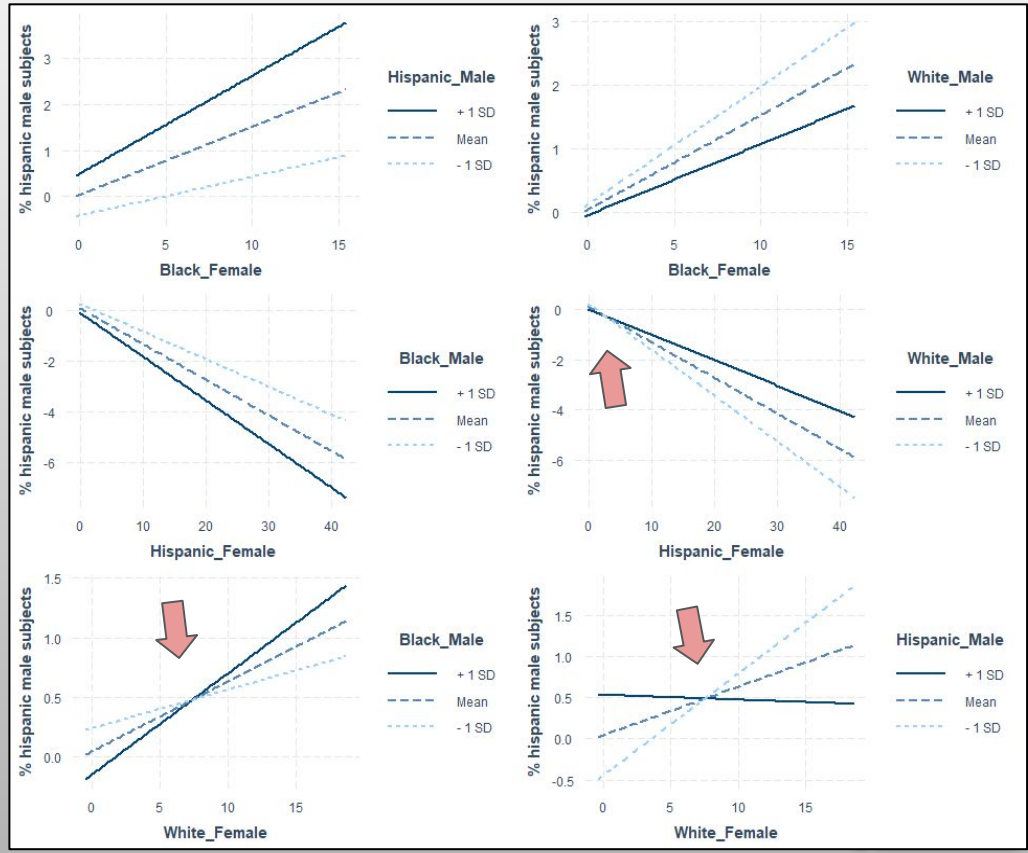
Modeling Subject Race and Sex

Multinomial Logistic Regression

Checked for interactions by building linear models with two predictors

- Limited search space by testing across race and sex for the three most common subject races in the data: Black, Hispanic, and White.
- For the dependent variable in each round of inspection, selected response variable with highest variance that had not yet been used

Potential Interactions
Hispanic Female x White Male
White Female x Black Male
White Female x Hispanic Male



Modeling Subject Race and Sex

Multinomial Logistic Regression

Each potential interaction term and its opposite pairing by subject sex was used as the predictor in a multinomial model with proportions of traffic stops by race and sex at the county level as the response and then compared with the null model.

Term 1	Term 2	χ^2 vs null model	p-value
Hispanic Female	Black Male	16.17	0.0635
Hispanic Male	Black Female	15.81	0.0710
Hispanic Female	White Male	19.37	0.0222*
Hispanic Male	White Female	19.48	0.0214*
White Female	Black Male	22.36	0.0078*
White Male	Black Female	21.67	0.0100*

* result is significant at the p=.05 level

Modeling Subject Race and Sex

Multinomial Logistic Regression

Included the four significant interaction terms with the base model as input to stepwise selection in both directions, which found the following model:

Response	(Intercept)	year	Black_Male	Hispanic_Male	White_Female	Other_Male	Black_Male: White_Female	Hispanic_Male:White_Female
stops_per_api_m	-71.8421	0.0362	-0.1501	-0.0093	0.0043	-0.0291	0.0258	-0.0066
stops_per_b_f	126.0057	-0.0613	1.1397	-0.7966	-0.2656	0.0570	-0.2591	0.1246
stops_per_b_m	150.8361	-0.0734	1.0439	-0.6810	-0.2406	0.0517	-0.2283	0.1082
stops_per_h_f	-108.7589	0.0548	0.1093	0.1885	-0.1608	-0.0066	-0.0176	-0.0002
stops_per_h_m	-7.9211	0.0053	0.2025	0.1000	-0.1840	0.0098	-0.0368	0.0101
stops_per_o_f	-46.4323	0.0233	-0.4871	0.2241	-0.5696	0.2837	0.0387	-0.0049
stops_per_o_m	-77.2702	0.0390	0.1134	0.0879	-0.4828	0.2397	-0.0276	0.0138
stops_per_w_f	31.8108	-0.0137	-0.0828	-0.1703	-0.2376	0.0622	0.0061	0.0116
stops_per_w_m	35.3436	-0.0151	-0.2660	-0.1114	-0.2871	0.0460	0.0639	-0.0088

AIC: 11291.29 χ^2 vs null model: 380.82 p-value: 0

County-level accuracy: 86.35%

Modeling Subject Race and Sex

Linear Discriminant Analysis

- Required the conversion of aggregated data back to individual observations

	year	Asian_or_Pacific_Islander_Female	Asian_or_Pacific_Islander_Male	Black_Female	Black_Male	Hispanic_Female	Hispanic_Male	Other_Female	Other_Male	White_Female	White_Male	name
1	2000	0.000799	0.000884	0.0176	0.0178	0.00153	0.00215	0.00487	0.00507	0.0322	0.0298	api_female
2	2000	0.000799	0.000884	0.0176	0.0178	0.00153	0.00215	0.00487	0.00507	0.0322	0.0298	api_female
3	2000	0.000799	0.000884	0.0176	0.0178	0.00153	0.00215	0.00487	0.00507	0.0322	0.0298	api_female
4	2000	0.000799	0.000884	0.0176	0.0178	0.00153	0.00215	0.00487	0.00507	0.0322	0.0298	api_female
5	2000	0.000799	0.000884	0.0176	0.0178	0.00153	0.00215	0.00487	0.00507	0.0322	0.0298	api_female

- Normalization ➡ small improvement
- No difference between methods of normalization

Normalization	County-level accuracy
Unnormalized	79.16%
Min-max	83.97%
Z-score	83.97%

Modeling Subject Race and Sex

Linear Discriminant Analysis

Using the same predictors as the final MNLR model produced the best results:

Predictor	LD1	LD2	LD3	LD4	LD5	LD6	LD7
year	0.0478	0.0020	-0.0382	-0.0099	-0.1383	0.1517	-0.1650
Black_Male	-28.8218	0.7935	-0.2386	7.7065	-27.6336	-11.4443	0.8167
Hispanic_Male	15.9097	-33.1562	11.0740	-1.0335	24.7219	15.6093	10.2733
White_Female	-0.5205	1.0062	-1.3461	-9.6367	-10.0834	-8.4135	1.7814
Other_Male	0.0745	-0.4344	-9.6929	8.3292	0.5684	2.2313	4.2278
Black_Male:White_Female	100.8618	-2.6899	-8.2938	-48.9489	197.1665	93.8036	-10.8167
Hispanic_Male:White_Female	-63.2992	32.3274	-4.8040	35.8047	-121.5915	-71.0841	-19.1821

Proportion of Trace:	LD1	LD2	LD3	LD4	LD5	LD6	LD7
	0.5587	0.3312	0.0576	0.0452	0.0048	0.0022	0.0003

County-level accuracy: 84.53%

Modeling Subject Race and Sex

Random Forest

- Unable to process entire training set as observations
  - Using downsampled data produced predictions too inaccurate to consider
  - Using random sample of 20% of observations to build model produced a viable model that could be processed quickly enough to tune hyperparameters
- Normalization ➡ no effect

Grid search hyperparameters
num.trees = {200, 500, 1000}
mtry = {2, 3, 4}
min.bucket = {400000, 600000, 800000}

num.trees	mtry	min.bucket	County-level accuracy
200	2	400000	83.67%
500	2	400000	83.62%
1000	2	400000	83.61%
500	3	400000	83.60%
1000	3	400000	83.60%

County-level accuracy: 83.67%

K-fold Cross-Validation for Subject Race and Sex

Evaluated all 3 models with k=10

	MNLR	LDA	RF
Mean county-level accuracy	86.32%	84.53%	85.03%

The MNLR model had the highest average county-level accuracy.

The Random Forest model was trained on only 40% of the available observations on each iteration. It is very possible that the ability to use even more of the data to train the model could result in better accuracy.

County-level accuracy: 86.32%

Combining Models

Distributed predicted proportions for subject age across predicted proportions for subject race and sex to create 80 predictions of proportions by subject age, race and sex for each county-year.

	ts_per_api_f_0.20	ts_per_api_f_21.30	ts_per_api_f_31.40	ts_per_api_41.50	ts_per_api_f_51.60	ts_per_api_f_61.70	ts_per_api_f_71.80	ts_per_api_f_81.Inf
1	0.0005294619	0.0009839117	0.0006628537	0.0004523444	0.0002212322	0.00007976571	0.00002445703	0.000005925729
2	0.0004842104	0.0009131087	0.0006116606	0.0004192012	0.0002050523	0.00007345667	0.00002197505	0.000005251605
3	0.0004692597	0.0008827732	0.0005878289	0.0004054295	0.0001984732	0.00007143160	0.00002135992	0.000005084309
4	0.0004724955	0.0008884560	0.0005928593	0.0004088047	0.0002003048	0.00007199331	0.00002158448	0.000005153882
5	0.0004741804	0.0008932590	0.0005951326	0.0004097728	0.0002009128	0.00007232162	0.00002167530	0.000005161974

Example □ Classes 1, 2, and 3 of model one with predicted probabilities of .2 and .8 combined with classes A and B of model two with predicted probabilities of .4 and .6.

Model one proportions	* Model two proportions	□ Combined model proportions	
.2	.4, .6	1A = .08	1B = .12
.8	.4, .6	2A = .32	2B = .48

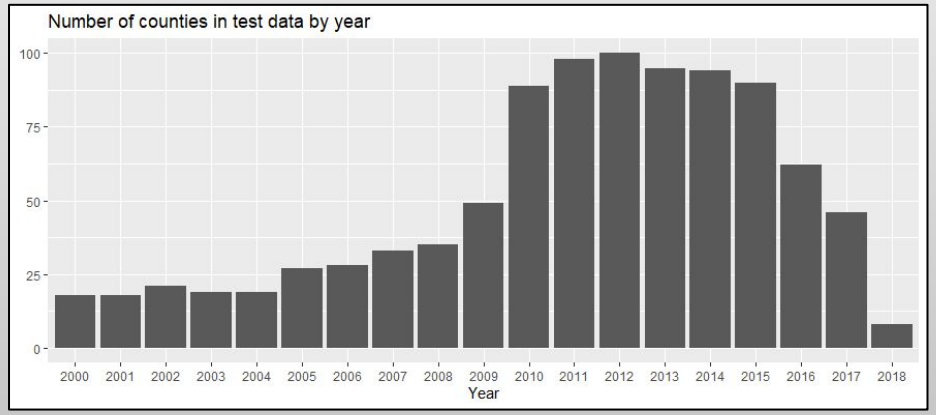
County-level accuracy: 83.58%

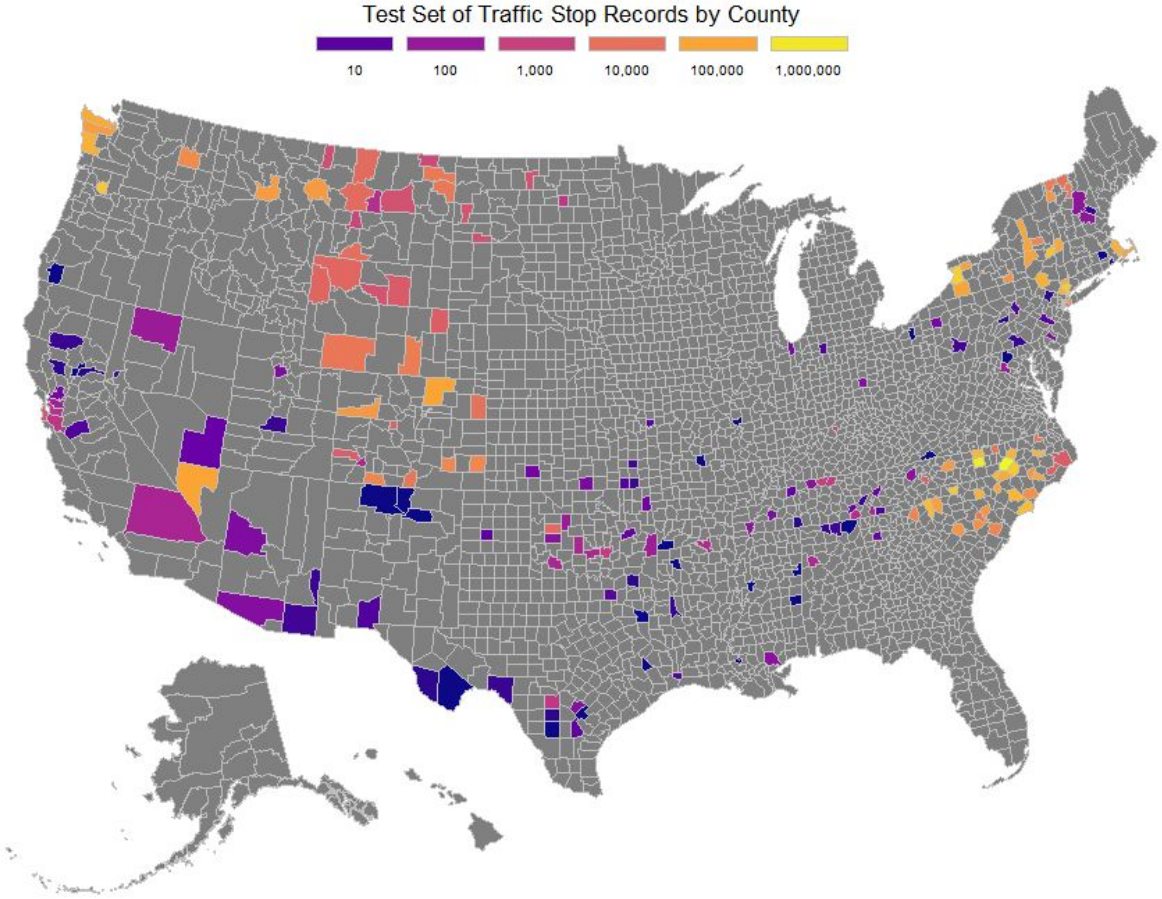
Overall accuracy: 94.32%

Testing

The 6 years from 2010 - 2015 each have at least 75 counties, which should serve as a large enough sample to evaluate the predictive power of the models.

We compared the accuracy of the model for subject age, the model for subject race and sex, And the combined model for all three by county and overall for each of these years and the entire 6 year period.




Test Set of Traffic Stop Records by County
10
100
1,000
10,000
100,000
1,000,000

Model Testing

Subject Age

Year	County-level accuracy	Total accuracy
2010	94.45%	98.40%
2011	94.58%	97.95%
2012	94.62%	98.20%
2013	94.76%	98.63%
2014	94.85%	98.90%
2015	94.59%	98.05%

Average county-level accuracy: 94.64%
Total overall accuracy: 98.83%

Subject Race and Sex

Year	County-level accuracy	Overall accuracy
2010	86.49%	97.23%
2011	86.36%	97.91%
2012	85.94%	97.95%
2013	85.80%	97.88%
2014	85.94%	96.74%
2015	84.75%	95.38%

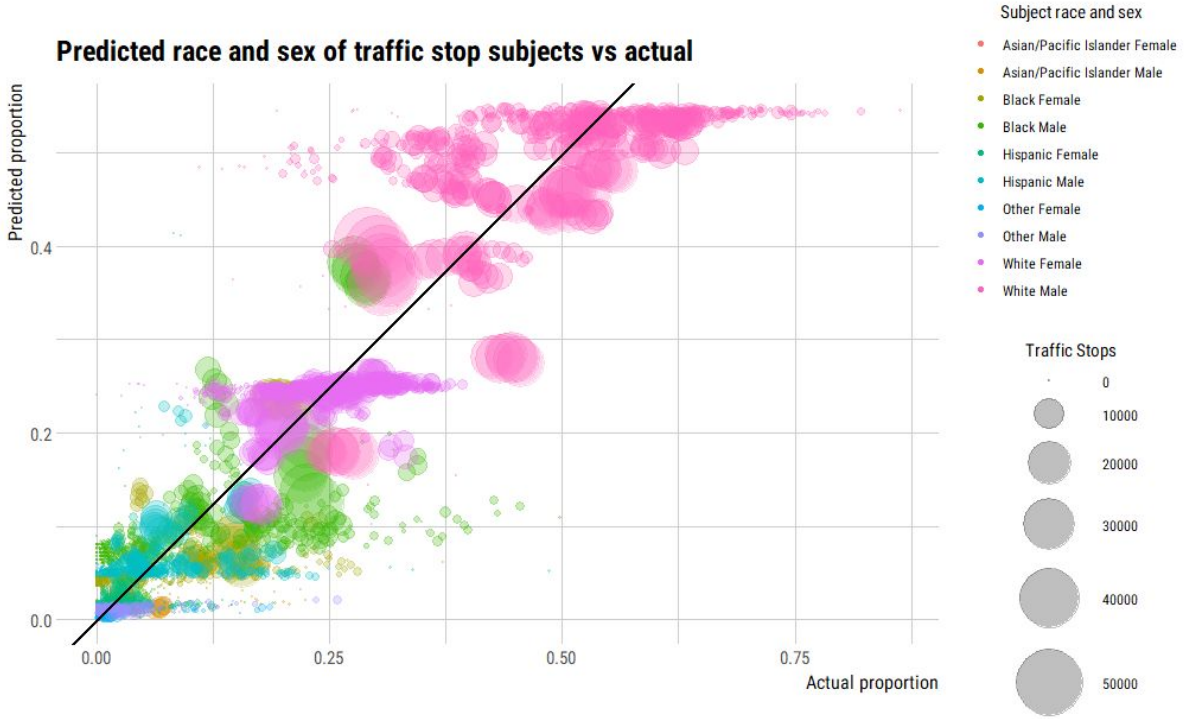
Average county-level accuracy: 85.91%
Total overall accuracy: 98.26%

Subject Age, Race and Sex

Year	County-level accuracy	Total accuracy
2010	84.02%	94.02%
2011	84.08%	94.28%
2012	83.73%	94.11%
2013	83.61%	93.86%
2014	83.82%	94.11%
2015	82.60%	93.36%

Average county-level accuracy: 83.67%
Total overall accuracy: 94.42%

Visualizing Predictions vs Reality



Discussion

- It is tempting to over-simplify results such as these into spurious conclusions, like “police stop black subjects at a higher rate than their representation in the county where the stop occurs, so police traffic stop decisions are biased against black subjects”.
- Though true that this data shows police stop black subjects at a higher rate than their representation in the county where the stops occurs, there are numerous other potential causes besides individual bias, e.g. access to vehicles, need to drive/ability to live close to work, police resourcing and allocation decisions, etc.
- A data source more closely related to who drives and how much, such as the population of drivers or car owners, would likely produce more accurate models and provide more interpretable results.

Further Research

- Add more features to the set of predictors, e.g. population density
- During this project, the SOPP made more data available which could potentially improve these models or serve as additional data for testing
- Model number of stops by county by year using population demographics with additional features, e.g. population density

Conclusion

- Population demographics are an excellent base to start with for modeling traffic stop demographics.
- These models perform much better when measured against a large number of county-years.
- There are significant differences in population demographics and traffic stop demographics which warrant far more research to define and explain.

Sources

1. The Stanford Open Policing Project. https://openpolicing.stanford.edu/
2. U.S. Census Estimates. https://seer.cancer.gov/popdata/popdic.html