The Hidden Biases of BISG-Based Disparity Estimates

Modern Fair Lending Analysis

Richard R. Pace

The BISG Proxy Model: Yesterday and Today

- The BISG Proxy Model was originally developed to analyze demographic disparities in **health care outcomes** leveraging U.S. Census-based geography and surname demographic distributions.
- Around 2013, certain federal financial regulators employed the BISG Proxy Model to test for potential disparate impact discrimination in non-HMDA consumer lending outcomes – primarily **automobile loan pricing**.
- The BISG Proxy Model and its corresponding fair lending disparities were the basis for numerous public enforcement agreements – and non-public MOUs and MRAs – alleging **fair lending violations** and requiring lenders to:
    - Pay millions of dollars in customer remediation payments.
    - Implement on-going fair lending monitoring using the BISG proxy-based testing approach and, where applicable, take corrective actions against third-party business partners and pay further customer remediation payments.
    - Implement changes to pricing models and compensation controls.
- More recently, consumer lenders have deployed the BISG proxy model to test for **potential algorithmic bias** in customer-impacting AI/ML models.

How Reliable are the BISG Proxy-Based Fair Lending Disparities?

Unfortunately, we do not really know.

- No publicly-released model validation testing results have been shared by the federal financial regulators – although such model testing, in general, is mandatory for federally-supervised banking institutions (i.e., SR 11-7).
- Certain ad-hoc accuracy studies have been performed using HMDA data with self-reported race / ethnicity data. Common findings include:
  - Identification of both aggregate-level and individual-level accuracy errors.
  - Statistically-derived price disparities based on the BISG proxies are different from those obtained using the self-reported race / ethnicity data.

Yet the model continues to be used due to regulatory requirements / expectations.

It's Long Past Time to Address This Important Question

Our study of the BISG Proxy Model focuses on the following **four questions** relative to its use for fair lending testing:

1) What is its **inherent accuracy** – in the aggregate (i.e., at the overall sample level) and at the individual level?
2) Do the BISG race / ethnicity proxies suffer from any type of **socio-economic biases**?
3) Do the BISG proxies impact the accurate measurement of **fair lending price disparities**?
4) If biases are present, what are the **implications and potential remedies**?

However, **unlike prior work**, our study is designed:

- Without the demographic and economic biases of HMDA data
- With known **"ground truth"** fair lending price disparities.
- Analyzing both **disparate treatment** and **disparate impact** price discrimination scenarios.
- Using both the **"BISG Continuous"** and the **"BISG Classification"** approaches when estimating fair lending price disparities.

What Data Do We Use For Our Testing?

We create a **large sample of synthetic U.S. adults** by leveraging the U.S. Census data distributions that underlie the BISG Proxy Model. More specifically,

- We simulate a sample of “actual” U.S. adults whose:
  - Frequencies of surnames and Census Block Groups (“CBGs”) are consistent with the 2010 U.S. Census population data.
  - “Actual” race / ethnicities are consistent with their corresponding BISG probabilities.
- To avoid biasing our results due to small sample sizes, we create a synthetic sample of **10 million U.S. adults**.
- We append **2010 median CBG household income** values to each sample member to explore potential socioeconomic biases.
- The synthetic sample allows us to design specific discrimination scenarios in which the “**ground truth” discrimination activity is known.**

Probabilistic Assignment of Synthetic Individual 1’s Race / Ethnicity

Surname	Block Group ID
Beech	250277309013
Race / Ethnicity	**BISG Probability**
White	95.5%
Black	2.5%
API	0.5%
Hispanic	0.5%
Other	1.0%

$+$ **0.48** **Random Number** $=$ **Simulated Race / Ethnicity**

White

Probabilistic Assignment of Synthetic Individual 2’s Race / Ethnicity

Surname	Block Group ID
Brancheau	120570118022
Race / Ethnicity	**BISG Probability**
White	68.9%
Black	0.2%
API	0.5%
Hispanic	30.2%
Other	0.2%

$+$ **0.75** **Random Number** $=$ **Simulated Race / Ethnicity**

Hispanic

Key Findings

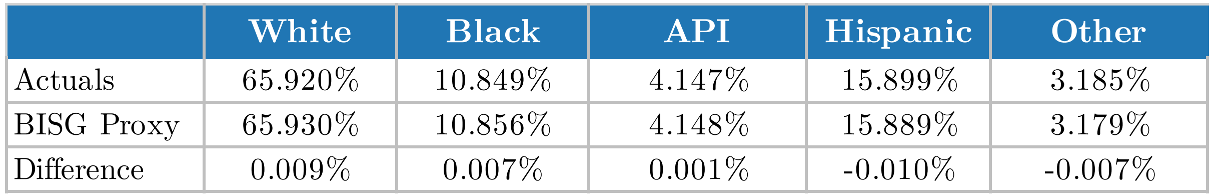
- Aggregate-Level Accuracy
- Individual-Level Accuracy
- Socioeconomic Biases
- Disparity Estimation Biases
- Potential Mitigants

Aggregate-Level Proxy Accuracy

Key Finding #1: The BISG proxy model is not inherently biased or error-prone in estimating aggregate race / ethnicity distributions ***for samples that are consistent with the model's underlying census data properties*****.**

- Previously reported measures of aggregate proxy error are really due to the authors' application of the BISG proxy model to data samples that are known to differ from the Census data properties underlying the model (e.g., HMDA data samples).
- To the extent that consumer lending samples are skewed toward particular socioeconomic profiles – such as higher income / assets or higher credit quality – then the BISG proxy probabilities can produce materially biased / inaccurate aggregate group membership counts and fair lending disparity estimates under the BISG Continuous approach (more on this later).

Figure 8: Actual vs. Expected Race / Ethnicity Distribution For 10 Million Synthetic Individuals



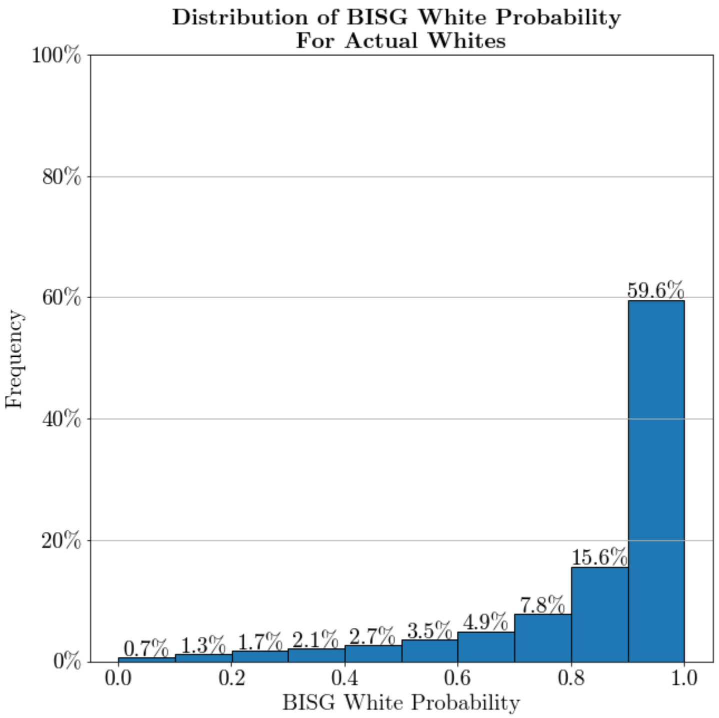
	White	Black	API	Hispanic	Other
Actuals	65.920%	10.849%	4.147%	15.899%	3.185%
BISG Proxy	65.930%	10.856%	4.148%	15.889%	3.179%
Difference	0.009%	0.007%	0.001%	-0.010%	-0.007%

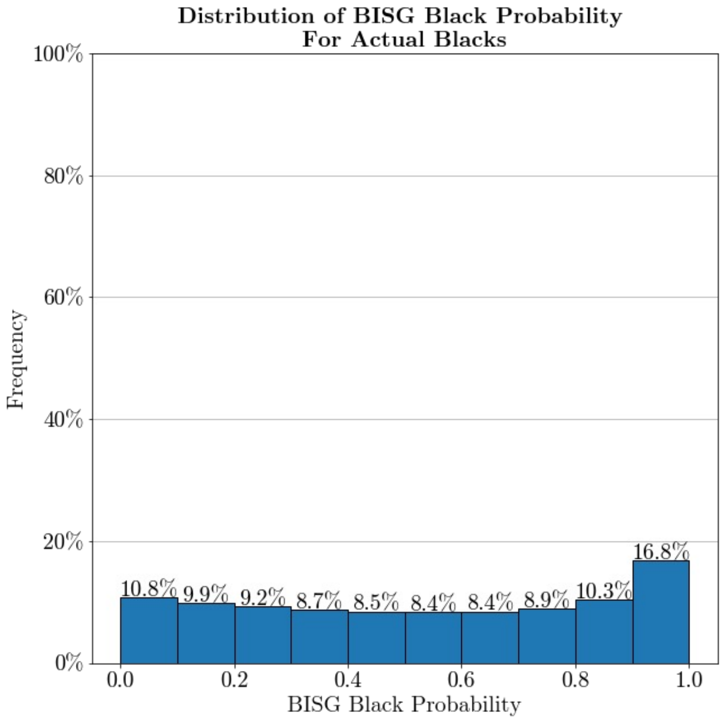
This finding is consistent with the well-known model risk management principle that a model should be applied to data samples that are materially aligned with the key properties of the original dataset used to estimate or train the model.

Low Proxy "Confidence" Levels For Blacks

Key Finding #2: The BISG Proxy Model produces relatively undifferentiated BISG Black probability values for Actual Black sample members.

- Rather than a model flaw, this property is simply a reflection of the Census data upon which the BISG proxy probabilities are based – thereby limiting its predictive power at the individual level.
- It is caused by a general lack of segregation – both geographic and surname – for the vast majority of Blacks.
- According to the 2010 U.S. Census data, Blacks in our broad-based national sample reside in micro-geographies that are, on average, only 39% Black while Whites in our overall sample reside in micro-geographies that are, on average, 76% White.
- Additionally, 59% of micro-geographies in which our sample Whites reside are at least 80% White, while only 18% of micro-geographies in which our sample Blacks reside are at least 80% Black.

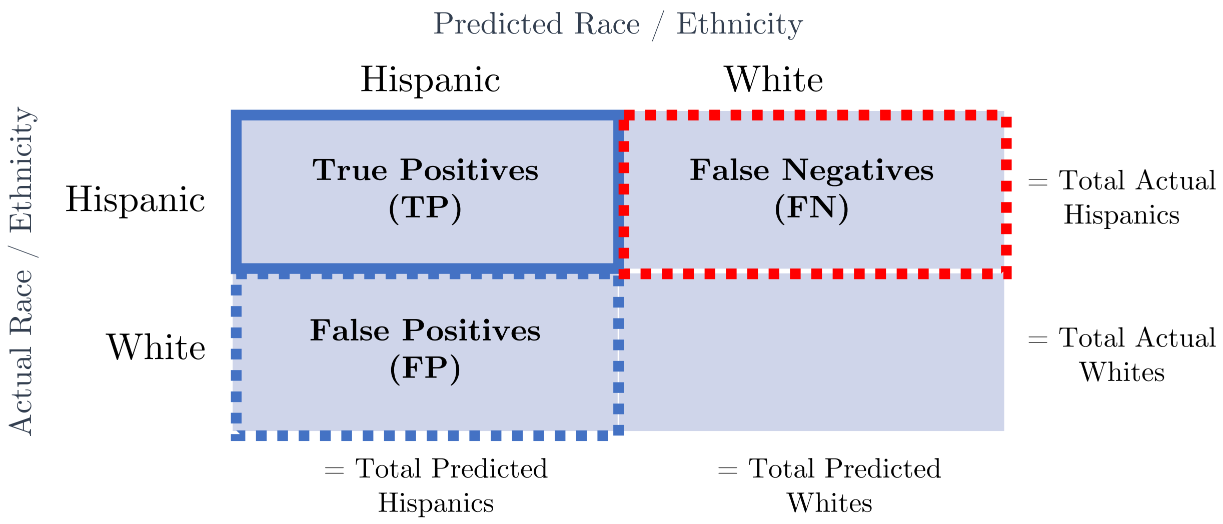




Evaluating Individual-Level Accuracy

To evaluate individual-level accuracy, we need two constructs:

- First, we need a **classification rule** that we apply to each individual's set of BISG probabilities to determine the race / ethnicity to which they will be assigned.
- Second, we need **accuracy metrics** that measure classification rule's accuracy in predicting the individual race / ethnicity of sample members. The diagram to the right illustrates this construct for classifying two races / ethnicities.
- After applying the classification rule, there are three outcomes of interest:
  - True Positives ("TPs") – these are the correctly predicted Actual Hispanics
  - False Negatives ("FNs") – these are the Actual Hispanics who were incorrectly predicted to be another race / ethnicity (here, White)
  - False Positives ("FPs") – these are non-Hispanics (here, White) who were incorrectly predicted to be Hispanics.



		Predicted Race / Ethnicity		
Actual Race / Ethnicity		Hispanic	White	
	Hispanic	**True Positives (TP)**	**False Negatives (FN)**	= Total Actual Hispanics
	White	**False Positives (FP)**		= Total Actual Whites
		= Total Predicted Hispanics	= Total Predicted Whites	

We use these three classification outcomes to derive the following accuracy metrics:

- Recall Accuracy = TP / (TP + FN) = % of Actuals that are correctly predicted
- Precision Accuracy = TP / (TP + FP) = % of Predicteds that are correct
- F1 Accuracy = an "average" of the Recall and Precision Accuracy metrics

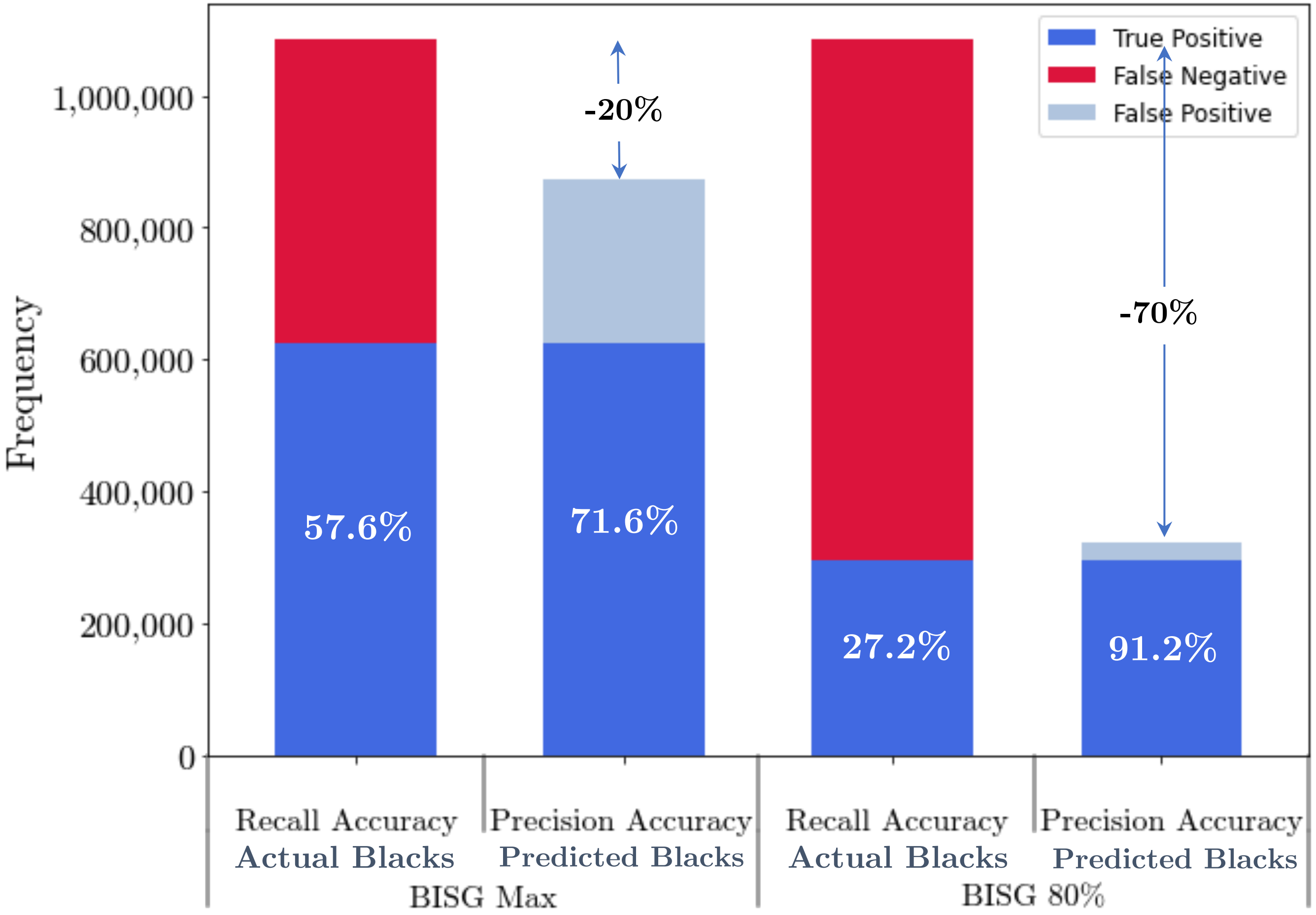
Individual-Level Proxy Accuracy: Blacks

Key Finding #3: Using common BISG Classification rules, we find that:

- **The BISG proxy model accurately identifies only 27% - 58% of Actual Blacks** due to the high rate of False Negatives caused by the model's low confidence level for Blacks.
- While the BISG proxy model's **Predicted Black groups are 72% - 91% accurate**, this is largely due to the relatively small size of the Predicted Black group. In fact,
  - Under the BISG Max classification rule, the Predicted Black group is only 80% the size of the Actual Black group.
  - Under the BISG 80% Threshold rule, the Predicted Black group is only 30% the size of the Actual Black group.

Additionally, as shown next, the Predicted Black groups also suffer from significant socioeconomic bias.

Recall and Precision Accuracy Components: Actual and Predicted Blacks



Individual-Level Race / Ethnicity Proxy Bias

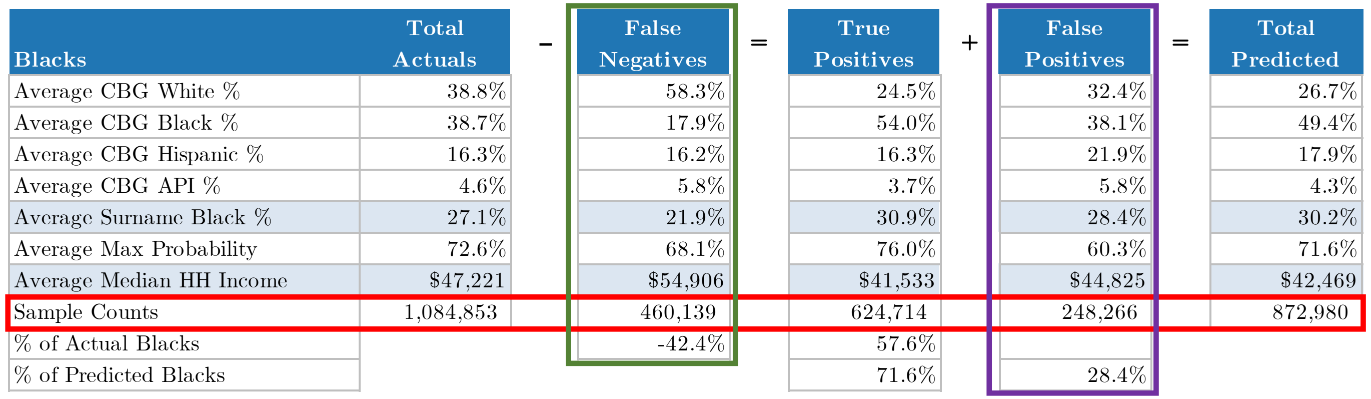
<u>Key Finding #4</u>: Overall, Predicted Blacks are a small biased subset of relatively low-income individuals living in highly segregated geographies – thereby providing a highly biased group of "Blacks" for fair lending testing.

For example, when used to predict individual Blacks using the BISG Max classification rule, we obtain:

- A 20% undercount of Actual Blacks
- The exclusion of 42% of Actual Blacks (False Negatives) who who have higher average CBG median incomes ($54,906 vs. $47,221) and live in racially-diverse areas (17.9% Black vs. 38.7% Black)
- The inclusion of 28.4% Non-Blacks (False Positives) – primarily White – who live in high minority areas (38.1% Black) and have below average CBG median incomes ($44,825 vs. $47,221).

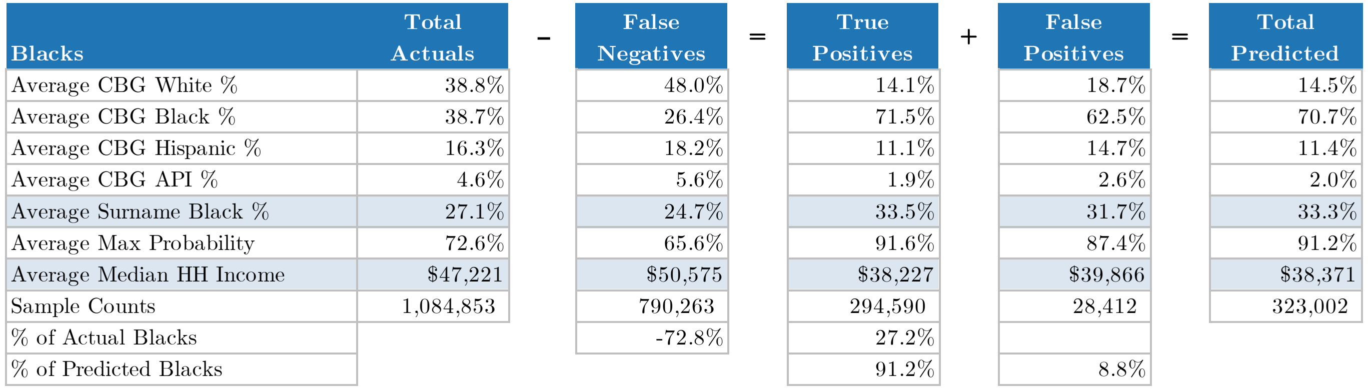
These group undercounts and socioeconomic biases are even more extreme when the BISG 80% Threshold classification rule is used.

Figure 19b: Comparative Characteristics of Actual vs. Predicted Blacks
BISG Max Classification Rule



Blacks	Total Actuals	–	False Negatives	=	True Positives	+	False Positives	=	Total Predicted
Average CBG White %	38.8%		58.3%		24.5%		32.4%		26.7%
Average CBG Black %	38.7%		17.9%		54.0%		38.1%		49.4%
Average CBG Hispanic %	16.3%		16.2%		16.3%		21.9%		17.9%
Average CBG API %	4.6%		5.8%		3.7%		5.8%		4.3%
Average Surname Black %	27.1%		21.9%		30.9%		28.4%		30.2%
Average Max Probability	72.6%		68.1%		76.0%		60.3%		71.6%
Average Median HH Income	$47,221		$54,906		$41,533		$44,825		$42,469
Sample Counts	1,084,853		460,139		624,714		248,266		872,980
% of Actual Blacks			-42.4%		57.6%				
% of Predicted Blacks					71.6%		28.4%		

Figure 25b: Comparative Characteristics of Actual vs. Predicted Blacks
BISG 80% Threshold Classification Rule



Blacks	Total Actuals	–	False Negatives	=	True Positives	+	False Positives	=	Total Predicted
Average CBG White %	38.8%		48.0%		14.1%		18.7%		14.5%
Average CBG Black %	38.7%		26.4%		71.5%		62.5%		70.7%
Average CBG Hispanic %	16.3%		18.2%		11.1%		14.7%		11.4%
Average CBG API %	4.6%		5.6%		1.9%		2.6%		2.0%
Average Surname Black %	27.1%		24.7%		33.5%		31.7%		33.3%
Average Max Probability	72.6%		65.6%		91.6%		87.4%		91.2%
Average Median HH Income	$47,221		$50,575		$38,227		$39,866		$38,371
Sample Counts	1,084,853		790,263		294,590		28,412		323,002
% of Actual Blacks			-72.8%		27.2%				
% of Predicted Blacks					91.2%		8.8%		

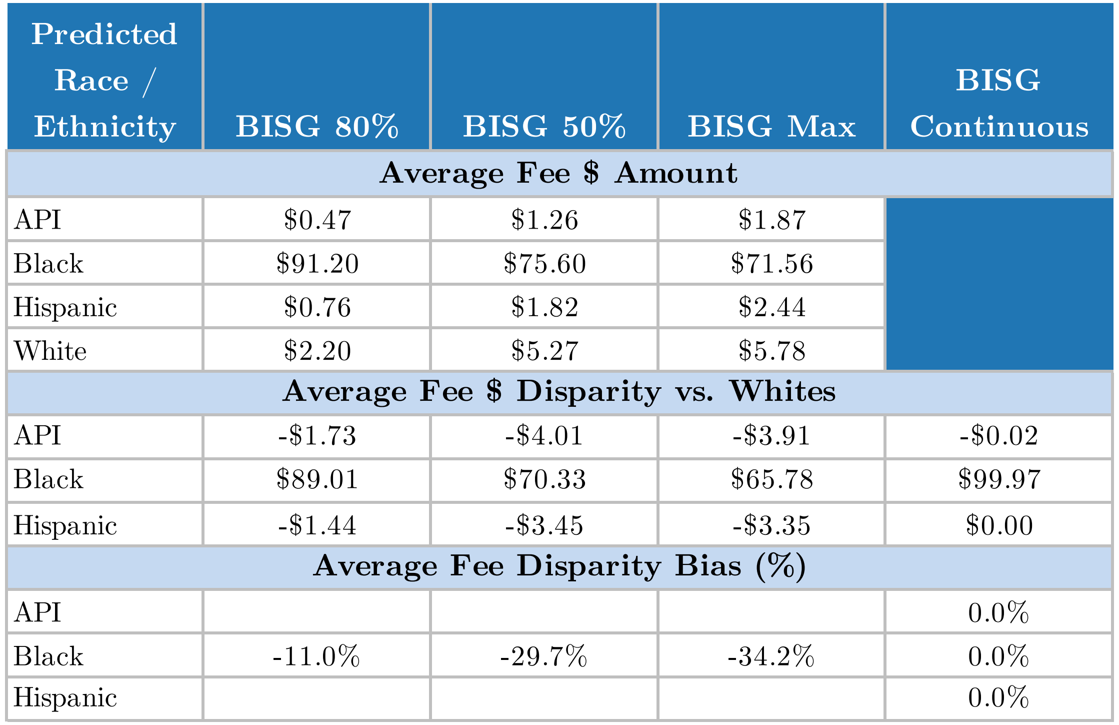
Disparate Treatment Disparity Estimate Bias

Key Finding #5: Under a disparate treatment pricing scenario in which only Blacks are assessed a $100 discretionary fee:

- All Black fee disparity estimates under the BISG Classification approaches **are biased downward by 11% - 34%.**
- The downward bias is primarily driven by the **cross-contamination** of False Negatives / False Positives.
- BISG Max has the most cross-contamination and BISG 80% the least – which explains the difference in their downward biases.
- The Black fee disparity estimate under the BISG Continuous approach is **unbiased**. However,
  - This result **only holds** if the underlying socio-demographic distribution of the transaction sample is aligned to the U.S. Census data on which the BISG proxy model is based.
  - To the extent that the transaction sample is biased – such as toward higher income or wealthier individuals – then the BISG Continuous approach will also bias the estimated fee disparities downward.

Figure 27: Disparate Treatment Scenario Test Results

Scenario: Blacks = $100, All Others = $0



Predicted Race / Ethnicity	BISG 80%	BISG 50%	BISG Max	BISG Continuous
Average Fee $ Amount				
API	$0.47	$1.26	$1.87	
Black	$91.20	$75.60	$71.56	
Hispanic	$0.76	$1.82	$2.44	
White	$2.20	$5.27	$5.78	
Average Fee $ Disparity vs. Whites				
API	-$1.73	-$4.01	-$3.91	-$0.02
Black	$89.01	$70.33	$65.78	$99.97
Hispanic	-$1.44	-$3.45	-$3.35	$0.00
Average Fee Disparity Bias (%)				
API				0.0%
Black	-11.0%	-29.7%	-34.2%	0.0%
Hispanic				0.0%

Disparate Impact Disparity Estimate Bias

For our disparate impact scenario, we assume that discretionary fees are charged based on the average CBG median income level where the individual resides – with higher income areas receiving lower fees and vice versa.

Average CBG Median Income Decile	Discretionary Fee Amount
1	$100
2	$90
3	$80
4	$70
5	$60
6	$50
7	$40
8	$30
9	$20
10	$10

Race / Ethnicity	Actuals	Actual Disparity Amounts
Average Fee $ Amount		
API	$48.65	-$3.51
Black	$67.25	$15.09
Hispanic	$59.10	$6.94
White	$52.17	

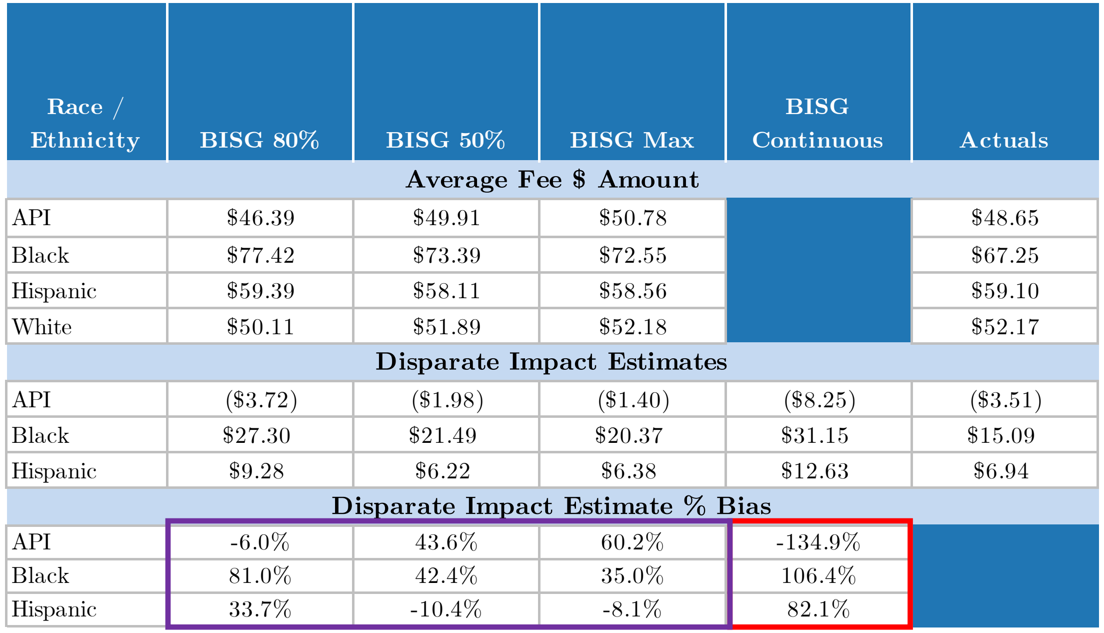
Based on the socioeconomics of the CBGs where each race / ethnicity group reside, we see that **this fee assessment "policy" leads to higher average fee amounts for Blacks and Hispanics – and lower average fee amounts for APIs (relative to Whites)**. These are the "ground truth" disparities we seek to estimate using the BISG Continuous and Classification approaches.

Disparate Impact Disparity Estimate Bias

Key Finding #6: Under a disparate impact pricing scenario in which discretionary fees are correlated with average CBG median income levels:

- All proxy-based estimates are biased – none come close to the “ground truth” values.
- **The BISG Continuous approach produces disparity estimates with the greatest bias**. In fact, the disparity amounts for all three minority groups are severely amplified in amount – **with Blacks and Hispanics experiencing 82% - 106% overstatement, and APIs experiencing 135% understatement.**
- The biases for the BISG Continuous approach occur **even if the underlying socio-demographic distribution of the transaction sample is aligned to the U.S. Census data on which the BISG proxy model is based.**
- **The BISG Classification approach also produces significantly biased results** – particularly for Blacks where the disparities are 35% - 81% overstated.

Figure 34: Disparate Impact Scenario Results: Discretionary Fee Schedule Based on Income



Race / Ethnicity	BISG 80%	BISG 50%	BISG Max	BISG Continuous	Actuals
Average Fee $ Amount					
API	$46.39	$49.91	$50.78		$48.65
Black	$77.42	$73.39	$72.55		$67.25
Hispanic	$59.39	$58.11	$58.56		$59.10
White	$50.11	$51.89	$52.18		$52.17
Disparate Impact Estimates					
API	($3.72)	($1.98)	($1.40)	($8.25)	($3.51)
Black	$27.30	$21.49	$20.37	$31.15	$15.09
Hispanic	$9.28	$6.22	$6.38	$12.63	$6.94
Disparate Impact Estimate % Bias					
API	-6.0%	43.6%	60.2%	-134.9%	
Black	81.0%	42.4%	35.0%	106.4%	
Hispanic	33.7%	-10.4%	-8.1%	82.1%	

Can The Disparate Impact Bias Be Mitigated?

Key Finding #7: We propose two alternative BISG Continuous estimation approaches to address the root cause of the disparity estimation biases. Both approaches move the uncertainty of race / ethnicity membership out of the regressors – thereby eliminating the cause of the bias.

1) Bootstrap Regression Approach

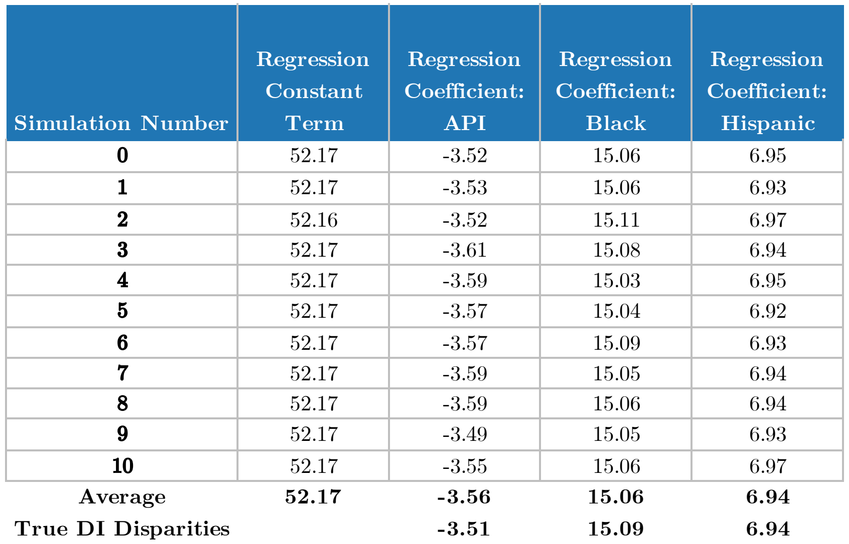
2) Proportional Regression Approach

Both approaches produce the same results, and we demonstrate that **they produce disparate impact disparity estimates that match the "ground truth" disparities**.

However,

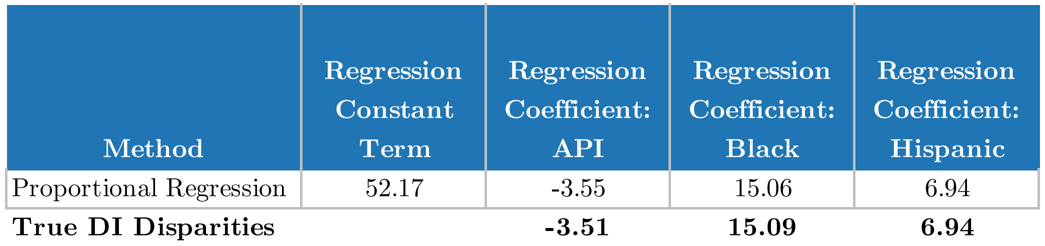
- They **only apply** to disparate impact disparity estimates.
- They only address the root cause of the estimation bias - **they do not produce "ground truth" estimates if the transaction sample is also biased**.
- **Caution must be exercised** in measuring the statistical significance of the disparity estimates.

Figure 38: Bootstrap Regression Coefficient Estimates



Simulation Number	Regression Constant Term	Regression Coefficient: API	Regression Coefficient: Black	Regression Coefficient: Hispanic
0	52.17	-3.52	15.06	6.95
1	52.17	-3.53	15.06	6.93
2	52.16	-3.52	15.11	6.97
3	52.17	-3.61	15.08	6.94
4	52.17	-3.59	15.03	6.95
5	52.17	-3.57	15.04	6.92
6	52.17	-3.57	15.09	6.93
7	52.17	-3.59	15.05	6.94
8	52.17	-3.59	15.06	6.94
9	52.17	-3.49	15.05	6.93
10	52.17	-3.55	15.06	6.97
Average	**52.17**	**-3.56**	**15.06**	**6.94**
True DI Disparities		**-3.51**	**15.09**	**6.94**

Figure 41: Proportional Regression Coefficient Estimates



Method	Regression Constant Term	Regression Coefficient: API	Regression Coefficient: Black	Regression Coefficient: Hispanic
Proportional Regression	52.17	-3.55	15.06	6.94
True DI Disparities		**-3.51**	**15.09**	**6.94**

Final Thoughts

- These results indicate that disparate impact pricing disparities are likely consistently biased upward for Blacks and Hispanics. Further research would be fruitful to explore a broader range of potential disparate impact scenarios as well as other types of lending outcomes (e.g., credit decisions).
- The degree of bias present for each lender will depend on the specific geo-surname distribution of their book-of-business. We would expect that these findings translate qualitatively to all lenders; however, the magnitudes of the disparities will likely vary with the overall segregation profile of the lender’s customers.
- These results reinforce the importance of model validation testing prior to the implementation of a new model for a high-stakes use.
- There is no easy solution to these findings.
- Currently, the best solution may be to advocate for expanded GMI data collection so that non-HMDA testing is based on actual demographic data.