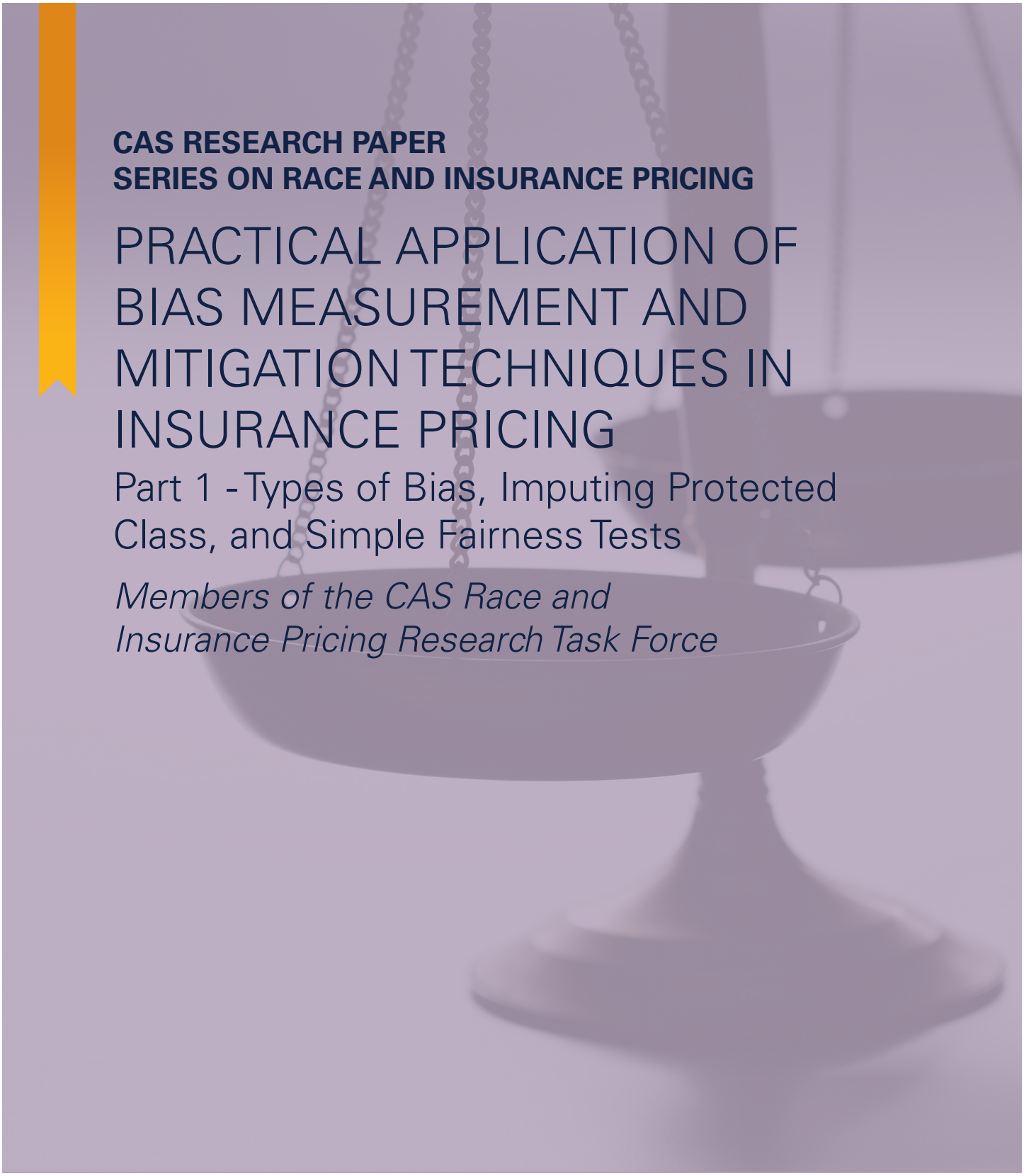


CAS RESEARCH PAPER SERIES ON RACE AND INSURANCE PRICING

PRACTICAL APPLICATION OF BIAS MEASUREMENT AND MITIGATION TECHNIQUES IN INSURANCE PRICING

Part 1 - Types of Bias, Imputing Protected Class, and Simple Fairness Tests

Members of the CAS Race and Insurance Pricing Research Task Force

CASUALTY ACTUARIAL SOCIETY

The CAS Research Paper Series on Race & Insurance Pricing was created to guide the insurance industry toward proactive, quantitative solutions that address potential racial bias in insurance pricing. These reports explore different aspects of unintentional potential bias in insurance pricing, address historical foundations and offer forward-looking solutions to quantify and handle possible bias. Through these reports, the CAS aims to encourage actuaries to discuss this topic with their stakeholders across all areas of insurance pricing and operations. For more information on the series, visit casact.org/raceandinsuranceresearch.

The Casualty Actuarial Society (CAS) is a leading international organization for credentialing, professional education and research. Founded in 1914, the CAS is the world's only actuarial organization focused exclusively on property-casualty risks and serves over 10,000 members worldwide. CAS members are sought after globally for their insights and ability to apply analytics to solve insurance and risk management problems. As the world's premier P&C actuarial research organization, the CAS reaches practicing actuaries across the globe with thought-leading concepts and solutions. The CAS has been conducting research since its inception. Today, the CAS provides thousands of open-source research papers, including its prestigious publication, Variance — all of which advance actuarial science and enhance the P&C insurance industry. Learn more at casact.org.

CAS RESEARCH PAPER
SERIES ON RACE AND INSURANCE PRICING

CAS RACE AND INSURANCE PRICING RESEARCH TASK FORCE

Mallika Bender, FCAS; Margaret (Peggy) Brinkmann, FCAS, CSPA; Eric Krafcheck, FCAS, CSPA; Craig Sloss, PhD, FCAS, FCIA; Gary Wang FCAS, CSPA; Mike Woods, FCAS, CSPA

Casualty Actuarial Society
4350 North Fairfax Drive, Suite 250
Arlington, VA 22203
www.casact.org
(703) 276-3100

Contents

Practical Application of Bias Measurement and Mitigation Techniques in Insurance Pricing: Part 1 – Types of Bias, Imputing Protected Class, and Simple Fairness Tests

By Members of the Casualty Actuarial Society Race and Insurance Pricing Research Task Force

Introduction to Part 1

Industry views on fairness in insurance pricing are evolving to include both the traditional understanding that insurance rates should not be "unfairly discriminatory" – that is, they should reflect differentials in risk among policyholders – and the potential that insurance rating may result in "discriminatory effects" where certain legally protected groups are subject to disproportionately higher or lower insurance rates than others. In the United States, many jurisdictions are taking regulatory and/or legislative action to encourage or require insurers to evaluate their own data and models for both of these types of fairness.[1]

This paper is intended as a practitioners' guide for actuaries and insurance professionals responsible for building, maintaining, or updating insurance pricing models that satisfy multiple views of fairness.

The paper is presented in two parts, **of which this is Part 1**. **Part 1** consists of three sections:

Section 1 of the paper introduces categories of "bias" as defined by the National Institute of Standards and Technology: systemic, statistical and computational, and human. This section reviews different types of bias and explain how they can affect data, model design, implementation, use, and monitoring. An understanding of these issues can help the analyst evaluate the results of fairness tests, identify the source of unfair outcomes, and/or determine how to address unfair outcomes.

[1] For more detail on recent regulatory and legislative actions, as of May 2024, refer to the following three papers in the CAS Research Paper Series on Race and Insurance Pricing: (1) "Regulatory Perspectives on Algorithmic Bias and Unfair Discrimination," (2) "A Practical Guide to Navigating Fairness in Insurance," and (3) "Comparison of Regulatory Framework for Non-Discriminatory AI Usage in Insurance."

Section 2 covers a data preparation step that is essential for evaluating fairness – appending protected class information to insurance data when it is not already available or practical to collect. The U.S. insurance industry is currently focused on measuring potential discriminatory effects between policyholders of different racial and ethnic groups. This paper explores the Bayesian Improved First Name Surname Geocoding (BIFSG) approach to impute race or ethnicity using information that is generally accessible to insurers – policyholder name and location. It begins with the theory and data supporting the BIFSG method, then covers challenges arising from the two potential first name data files that have so far been developed for use in the United States, with regard to data cleansing and performance of the imputation method. This section also outlines different options for using the BIFSG output – a set of probabilities representing the likelihood that a policyholder belongs to one of six race/ethnicity groups – including classification of records based on maximum probability, direct use of probabilities, and simulated imputations.

Section 3 extends the three families of fairness testing methods introduced in a previous paper in this series[2] – independence, separation, and sufficiency – from their original applications on binary classification models to equivalent application to continuous models such as those used for insurance pricing. This section discusses how these fairness tests may be interpreted through an insurance pricing lens as premium parity and loss ratio parity tests. Illustration of these tests suggests that satisfying multiple types of fairness tests at the same time may be impossible. Thus, there is still a need for a robust discussion within the insurance industry on what types of parity insurers should aim to achieve.

Part 2 of the paper, which can be found on casact.org/raceandinsuranceresearch, is also presented in three sections.

- Part 2, Section 4 delves into more complex fairness analyses that take into consideration multiple rating factors and distributional differences between protected classes across the levels of certain rating factors, conditional demographic parity, the proxy ("control variable") test, and nonparametric matching.
- Part 2, Section 5 reviews several technical bias mitigation methods that can be applied to insurance pricing data, models, or model outputs.
- Part 2, Section 6 discusses several important non-modeling considerations that can contribute to fairness concerns, such as targeted marketing practices, regulatory restrictions, and discounts.

While this paper may be read from start to finish, readers are invited to navigate directly to the part and/or section of the paper that is most relevant to their current responsibilities.

[2] These methods are introduced in the CAS Research Paper Series on Race and Insurance Pricing report, "Methods for Quantifying Discriminatory Effects on Protected Classes in Insurance".

Section 1. Types of Bias to Identify and Mitigate

In recent years, fairness in insurance practices, particularly with respect to protected attributes such as race and ethnicity, has become a significant and sustained focus of insurers and regulators. While all stages of the insurance life cycle, including marketing, underwriting, pricing, claims handling, and fraud investigation, are under scrutiny, much of the attention so far has been on fairness in insurance underwriting and pricing. Actuaries play a significant role in developing, implementing, and evaluating insurance underwriting and pricing models. New legislative and regulatory activity has aimed to better understand and address fairness in insurance. The 2022 paper *Defining Discrimination in Insurance* examines several terms used to describe fairness (or its absence) as they relate to insurance, such as *unfair discrimination*, *proxy discrimination*, and *disproportionate impact* and *disparate impact* (Chibanda 2022). *Bias* is often considered an underlying driver of these unfair outcomes, but the term itself is an umbrella that encompasses a variety of phenomena that can occur in a variety of contexts.

Understanding how biases can enter the modeling process and contribute to potential unfair discrimination or unfair outcomes can help inform choices in diagnostics and mitigation approaches, where appropriate. This section defines several types of bias and how each affects the stages of model development.

1.1. Types of Bias

There are dozens of types of bias. The National Institute of Standards and Technology groups them into three categories:

- *Systemic.* Systemic biases result from societal and institutional procedures and practices that result in certain social groups being advantaged and others being disadvantaged. These biases can result from conscious social prejudices but can also arise from the majority following existing rules or norms. They can be occurring in the present or can present themselves as residual effects of historical procedures and practices.
- *Statistical and computational.* Statistical and computational biases stem from errors that result when the sample is not representative of the population and from model design/validation choices. These biases can occur without discriminatory intent.
- *Human.* Human biases reflect systematic errors in human thought. Human biases come in a wide variety, some of which are a fundamental part of the human mind. They may also be based on limited information or influenced by individual experiences (Korteling and Toet 2022).

Such biases can enter predictive models at multiple points, including but not limited to data collection and selection, model design, and model implementation/monitoring/use (American Academy of Actuaries 2023a). The following sections discuss potential impacts of bias at various phases of modeling.

1.2. Impacts on Data

Systematic, statistical, or human bias can affect data in multiple ways.

Data selection. Human availability biases can lead modelers to use datasets and/or variables that are more available or accessible instead of a dataset that might be more suitable (NIST 2022). Such datasets may not appropriately reflect the population to which the model is relevant (Jager et al. 2020). Confirmation bias can result in choosing datasets and/or predictor variables that fit or confirm the modeler's existing beliefs.

Data sampling. Some groups may be underrepresented in datasets due to systemic bias or other reasons, which can result in statistical bias if models do not perform well for underrepresented groups. For example, when Amazon trained a model to screen applicants in its hiring process on a dataset that was disproportionately male, the model performed better when evaluating men versus women (Datta 2021).

In insurance pricing, sampling bias may be present if the historical batch of insurance policies used to develop a model contains a different mix of risk characteristics than the future policies to which the model-informed decisions are applied. Underwriting risk selection guidelines, marketing strategies, and other generally acceptable business practices can result in subgroups being over- or underrepresented in the historical data used to build a model and model outputs that are imprecise and/or biased away from the true mean.

Data values. Even when datasets are representative, the data values contained within may still exhibit entrenched systemic and human bias in the target variables, predictor variables, or both.

Target variables. Bias in a target variable is also known as negative legacy (Datta 2021) or label bias (Verma 2021). It occurs when the target variable is affected by systemic and/or human bias. This could be human judgment in manually labeling outcomes or historical/institutional practices that have generated differential outcomes.

In insurance pricing, target bias could result from biased claim processes, such as higher rates of denied claims or lower claim payouts for certain groups. Claims fraud identification is also at risk for label bias, as cognitive biases and social prejudices may influence which claimants are more or less likely to be flagged for investigation. Fraud models built on such data are likely to produce results that reflect those same biases and prejudices. Overreliance on automation can unintentionally embed past human biases into future automated decisions (automation bias).

Predictor variables. Distributions of predictor variables can vary by subgroup due to systemic and human biases, posing a risk of creating disproportionate impacts and/or unfair discrimination in rates or other insurance outcomes. This type of bias can be the result of human judgments influenced by cognitive biases or social prejudices or simply mechanical errors. For example, some studies suggest that policing practices may be influenced by social prejudices and result in disproportionate numbers of traffic stops or citations between

drivers of different races (2021 CAS Race and Insurance Research Task Force 2022). Including unwarranted traffic citations as an indicator for risky driving in an auto insurance rating model may result in incorrect predictions of higher or lower loss costs for certain protected groups.

It is possible for predictor variables to both be directly predictive of insurance claims and act as proxies for discriminatory characteristics. For example, using the presence of diabetes as a rating factor will be directly predictive of health insurance costs, but because certain racial or ethnic groups may be predisposed to develop diabetes, including diabetes as a rating factor may lead to disproportionate impacts by race (Lindholm et al. 2022). In such situations, it is helpful to consider other attributes of the variable that influence society's assessment of whether it is fair for insurance purposes, such as controllability and causality (see discussion of considerations in Frees and Huang 2023). Excluding or significantly tempering the effects of such variables in insurance pricing models can lead to moral hazards and escalating loss costs, which can have an adverse impact on insurance affordability and/or availability.

1.3. Impact on Model Design

Typically, insurance pricing models are developed with significant human involvement in the modeling process, which opens the possibility of human bias and behaviors affecting the model design and selection. That said, it is important to keep in mind that models built with less human involvement are not necessarily less susceptible to bias. Some examples of statistical and human biases that can affect model design are discussed below.

Aggregation bias, also known as ecological fallacy or confounding bias, refers to error introduced when trends in a predictor do not apply to all subgroups (American Academy of Actuaries 2023a). It can occur when another predictor is omitted or an interaction effect is overlooked. In the most extreme cases, aggregation bias can reverse the coefficient of the predictor, a phenomenon known as Simpson's paradox.

Confirmation bias can harm predictive modeling analyses. For example, an actuary may have an existing bias going into the modeling exercise, expecting to see a particular result. The actuary does not see the results they expect, so they continue to tweak the model parameters until they eventually arrive at a scenario that confirms their existing bias. By becoming aware of confirmation bias, actuaries can be more open to accepting results that do not confirm their original hypotheses (American Academy of Actuaries 2023a).

Omitted variable bias can have a significant impact on risk classification models. Modeling algorithms can use only the predictors included in the training dataset; in some cases, additional data may be required to produce accurate results. Omitted variable bias can be seen when the output of an algorithm is based on certain learned correlations while a different, and potentially more accurate, output may have been produced had the algorithm considered different or additional information (Serwin and Perkins, n.d.).

By leaving out important variables, correlations can arise between other variables to try to account for this lost signal, or the signal can be lost altogether. This will lead to a less

explanatory model overall and potentially skew the coefficients of the other included variables (American Academy of Actuaries 2023a). For example, some studies have shown that differing levels of infrastructure investment may create or promote existing racial inequality (Norwood 2021). If that affects road quality, congestion, or other driving-related factors, it could result in an apparent variation in auto insurance risk by subgroup, especially if infrastructure information is not included in the rating model. This may present as a less desirable outcome for some policyholders, which may or may not be deemed unfair discrimination.

Other model design choices that bias can affect include these:

- **Choice of target variable.** For example, a model to predict employee "productivity" could be trained on hours worked, which could disadvantage women with higher childcare burdens, as opposed to other measures.
- **Choice of model.** Data dredging bias is the misuse of statistical inference by probing the data in unplanned ways to find "attractive" results (Catalog of Bias Collaboration 2020). Examples include testing of large numbers of hypotheses to produce statistically significant results even when the results are statistically nonsignificant (NIST 2022), selectively reporting the "best" model, or generating a hypothesis to explain results that have already been obtained but presenting it as if it were a hypothesis one had prior to collecting the data (Catalog of Bias Collaboration 2020).

1.4. Impacts on Model Implementation, Use, and Monitoring

Deployment bias happens when a model is used in ways the developers did not intend. For example, developers of an algorithm used by major U.S. cities to assist in coordinating housing to homeless people began phasing it out after several cities inappropriately used the algorithm as an assessment tool rather than as the prescreening tool as it was designed.[3] In insurance pricing, an example of potential deployment bias might be the use of a submodel that was not developed for insurance risk assessment, such as a submodel that predicted credit default risk, as an input to a loss cost model.

Selective adherence is a human bias where decision makers selectively adopt algorithmic advice, such as, when it matches their preexisting beliefs and stereotypes (NIST 2022). Underwriters or claims adjusters, for example, may selectively vary from the recommendations of a model in their workflow.

Emergent bias and "concept drift" can occur when a model is used outside of the domain on which it was trained or in an unanticipated/new context (NIST 2022). In insurance, this could result from using a model trained on historical experience in new, different

[3] The Vulnerability Index – Service Prioritization Decision Assistance Tool (VI-SPDAT) was meant to help local social service providers assess what type of housing assistance might best suit a homeless person's needs. Instead, resource-strapped cities relied on VI-SPDAT to make a binary choice: who gets housing and who doesn't (Thompson 2021).

markets/segments or continuing to use models that have not been trained on examples of new and emerging types of claims.

Automation bias is a tendency to favor results generated by automated systems over those generated by nonautomated systems, irrespective of their relative error rates (Alon-Barkat and Busuioc 2021). For example, insurers may implement models for making underwriting decisions or claims adjusting without testing whether the models are more accurate than a human.

Bias is not new to predictive modeling, and it is not possible to eliminate the risk of bias in model data, design, and implementation. However, the risk can be managed through governance and practice improvements for identifying, understanding, measuring, managing, and reducing bias.

Section 2. Appending Protected Attributes to a Dataset

To evaluate data for model fairness or discriminatory effects, insurers must first be able to label the data and categorize the information as belonging to various classes of interest. In principle, that would mean accessing and handling sensitive data. However, property and casualty insurers do not typically collect protected class attributes other than gender and age.

Given that limitation, researchers have relied on imputation methodologies. A popular choice today is the Bayesian Improved First Name Surname Geocoding (BIFSG) approach.

2.1. Data Sources Underlying BIFSG

The U.S. Census Bureau publishes demographic summary information from its data efforts. In particular, the Census Bureau publishes demographic breakdowns for race and ethnicity at various geographic levels of detail, such as by ZIP Code Tabulation Areas (ZCTAs), by census tract, and by census block group. Historically, this is the starting point for the imputation. The geocode is captured or derived from the data, and the racial and ethnic proportions are then appended at that geographic level of detail.

The Census Bureau also separately provides summary racial and ethnic distributions by surname. Elliott et al. (2009) introduced a manner to blend these two sets of information utilizing the Bayesian formula and some independence assumptions, which they coined the Bayesian Improved Surname Geocoding (BISG) approach. Voicu (2018) expanded the methodology to incorporate a third dataset containing summary racial and ethnic distributions by first name from the mortgage data–gathering efforts of Tzioumis (2018) commonly known as the Bayesian Improved First Name Surname Geocoding (BIFSG) approach. Rosenman et al. (2022) introduced a variation of BIFSG that uses first name and race and ethnicity breakdowns based on voter registration data from six southeastern U.S. states. Two tools were developed for the execution of the aforementioned approaches. The Python package *surgeo* uses the first name summary information based on mortgage data, while the R package *wru* uses the information based on voter registration.

2.2. Derivation of BIFSG

The mathematics behind BISG and BIFSG can be found in various sources, such as the documentation for the *surgeo* package or as part of the Elliott et al. (2009) paper. For the readers' convenience, a simple presentation of the BIFSG formulation is offered here:

Given r, g, s, and f, where

r = race and ethnicity,

$$r \in \left\{ \begin{array}{l} White, Black, Hispanic, Asian\ and\ Pacific\ Islander\ (API), \\ American\ Indian\ and\ Alaska\ Native\ (AIAN), Multiracial \end{array} \right\};$$

g = geocode – either ZCTA, census tract, or census block group;
s = surname; and
f = first name;

$$P(r|g,s,f) = \frac{P(r,g,s,f)}{\sum_i P(r_i,g,s,f)}.$$

Using the probability chain rule,

$$P(r,g,s,f) = P(g) * P(r|g) * P(s|r,g) * P(f|r,g,s).$$

Assuming

$$P(s|r,g) = P(s|r), and$$

$$P(f|r,g,s) = P(f|r),$$

the desired conditional probability formula becomes

$$P(r|g,s,f) = \frac{p(g) * P(r|g) * P(s|r) * P(f|r)}{\sum_i p(g) * P(r_i|g) * P(s|r_i) * P(f|r_i)}$$

$$= \frac{P(r|g) * P(s|r) * P(f|r)}{\sum_i P(r_i|g) * P(s|r_i) * P(f|r_i)}.$$

The above simplifying assumptions for conditional probabilities on surname given race and on first name given race are major simplification steps, essentially trading local conditional probabilities for global conditional probabilities in the equation. This allows for the most direct blending of information from the three datasets.

The evaluation that follows in 2.3 includes comparisons against the BISG approach, which does not utilize the first name and the associated race and ethnicity summary information. The formulation for BISG follows the above derivation for BIFSG in principle, and is not separately demonstrated here:

$$P(r|g,s) = \frac{P(r|g) * P(s|r)}{\sum_i P(r_i|g) * P(s|r_i)}.$$

2.3. Evaluating the Relative Performance of Imputation Methods

The discussion that follows uses the following data sources to perform imputations:

- Census 2010 demographics data summary by ZCTA for the geocode information,
- Census 2010 demographics data summary by surname for the surname information, and
- mortgage information–based demographics data summary by first name for the first name information.

The practitioner should consider whether the data sources are appropriate given their intended purpose. The evaluation here reflects the use of common sources of information for the BIFSG approach and is not itself an endorsement or recommendation of any one particular set of data. For those using alternative data for imputing race and ethnicity based on the BIFSG approach, due diligence is needed to understand the selected imputation process and understand and document key assumptions and considerations when evaluating the results.

To evaluate the various implementations of BISG and BIFSG, this analysis relies on North Carolina voter registration data.[4] The data contain zip code, surname, first name, and race and ethnicity information. For the purposes of this paper, only those records where BIFSG can be applied were kept in the dataset. Records where the race/ethnicity information is either unknown or entered as "OTHER" were removed. In total, out of 8.4 million records (8,424,012), 5.7 million records (5,659,018) were retained for evaluation purposes.

This section compares the various imputation implementations with a consistent suite of summary statistics. The analysis begins with the proportion of each race/ethnicity from the imputation, as a check on whether the imputation distribution is statistically biased relative to the listed race/ethnicity distribution. Then, for each race/ethnicity group, the imputation is evaluated using the following metrics:

- Precision: True Positive/(True Positive + False Positive)
  - reflects the correctness of the imputation in designating records as being of the specific race or ethnicity

[4] It is important to note that this test set may not be representative of the United States as a whole or other parts of the United States, so the results should not be extended to other datasets/populations.

- Recall: True Positive/(True Positive + False Negative)
  - reflects the proportion of a specific race or ethnicity the imputation correctly identifies
- F1 Score: 2 ∗ Precision ∗ Recall/(Precision + Recall)
  - reflects the harmonic mean of Precision and Recall and is a single measure that reflects both the false positives and the false negatives
- Accuracy: True Positive/Total Records Imputed
  - calculated across all race and ethnicity groups

There is often a tradeoff between precision and recall. If the imputation process is careful and only indicates the most probable records to be of the designated race or ethnicity, it will achieve a higher precision for being wrong less often in the records the method tags, at the cost of tagging fewer of the records belonging to that race or ethnicity. That is, the method gains precision at the cost of lower recall.

A common method for using the imputation results is to assign to each record the race/ethnicity with the highest imputed probability. The classification statistics in Tables 2.1 and 2.2 are the results from using this imputation approach and selecting the race/ethnicity based on the maximum probability of the six imputed categories.

As can be seen in Table 2.1, introducing surname and first name information to the imputation process using the BIFSG approach leads to more accurate imputations overall. Furthermore, as Table 2.2 shows, using additional information improves the precision and recall of the imputation for most groups. More specifically, the improvement is seen for the four race and ethnicity groups white, Black, Hispanic, and Asian and Pacific Islander (API), which represent

Table 2.1. Comparison of Accuracy for Imputation Methods

Percentage of Total Records				
Race/Ethnicity	Actual	Geocode	BISG	BIFSG
White	76%	88%	78%	78%
Black	17%	11%	18%	17%
Hispanic	4%	0%	3%	3%
Asian and Pacific Islander (API)	1%	0%	1%	1%
American Indian and Alaskan Native (AIAN)	1%	1%	1%	0%
Multiracial	0%	0%	0%	0%
Total	100%	100%	100%	100%
Overall accuracy		**78%**	**81%**	**84%**

Table 2.2. Comparison of Precision, Recall, and F1 Score for Imputation Methods

Race/ Ethnicity	Percentage of Total Records	Precision			Recall			F1 Score		
		Geocode	BISG	BIFSG	Geocode	BISG	BIFSG	Geocode	BISG	BIFSG
White	76%	81%	88%	89%	94%	89%	91%	87%	88%	90%
Black	17%	52%	56%	62%	33%	57%	62%	40%	57%	62%
Hispanic	4%	13%	68%	74%	0%	53%	66%	0%	60%	70%
API	1%	7%	76%	81%	0%	58%	63%	0%	66%	71%
AIAN	1%	60%	66%	65%	44%	51%	40%	51%	58%	49%
Multiracial	0%	0.0%	3.5%	2.8%	0.0%	0.1%	0.4%		0.1%	0.7%

nearly all the data under evaluation. The results are mixed for American Indian and Alaska Native (AIAN) or those identified as multiracial. The improvement from incorporating additional information can also be seen from a review of the overall imputation accuracy.

Because BIFSG is a formula-driven approach to impute the race and ethnicity probabilities, the decisions made along the way of computation can lead to different probability imputations. The paragraphs below highlight a few such junctures to illustrate the variations.

First, note that the three datasets are not consistent representations of the U.S. population. In particular, the racial breakdowns vary across the three datasets (the first name file in the example that follows is based on the mortgage data aggregation, as used in *surgeo*).

As seen in Table 2.3, while the ZCTA table and the surname table are comparable, the results are not exact. The mortgage data supplied a robust 2.66 million observations and 2.45 million Specific First Name levels for the process, after the sufficiently unique first name records were collated into a single All Other First Names level. However, the biased sample representation of the mortgage data compared with the geocode and the surname table is very pronounced. This is further exacerbated by the more unique first names generally representing those in various minority groups. Typical implementations for the first name file use only the specific first names data, within which 85% of the names are categorized as white. When the first name from the record does not match a specific entry in the first name file, the typical treatment is to use an alternative approach, such as the BISG approach, rather than using the All Other First Names proportions from the first name file and continuing with the BIFSG algorithm.

In the BIFSG formulation, there is an order to look up the geocode, surname, and first name probabilities, The first probability takes a different form than the subsequent two probabilities. This gives rise to three possible variations to BIFSG. The formulation above reflects starting with the geocode probability, and we will distinguish this variation with a subscript $BIFSG_g$.

Table 2.3. Comparison of Race/Ethnicity Tables Underlying Imputation Methods

Source	Data Segments	Total Records	White	Black	Hispanic	API	AIAN	Multiracial
ZCTA*	**TOTAL**	**311,857,728**	**63%**	**12%**	**17%**	**5%**	**1%**	**2%**
Surname file	Specific Surnames	*265,660,058*	64%	12%	17%	5%	1%	2%
	All Other Surnames	*29,312,001*	67%	9%	14%	8%	1%	2%
	Blank	*7,170*	94%	2%	2%	1%	0%	1%
	TOTAL	**294,979,229**	**64%**	**12%**	**16%**	**5%**	**1%**	**2%**
First name file	Specific First Names	*2,449,240*	85%	4%	7%	4%	0%	0%
	All Other First Names	*214,124*	51%	12%	8%	28%	0%	0%
	TOTAL	**2,663,364**	**82%**	**4%**	**7%**	**6%**	**0%**	**0%**

*ZCTA = ZIP Code Tabulation Area.

The other two alternative formulations, leading with surname and with first name, are referred to as BIFSG_s and BIFSG_f, respectively:

$$\text{BIFSG}_g: P(r|g,s,f) = \frac{P(r|g) * P(s|r) * p(f|r)}{\sum_i p(r_i|g) * p(s|r_i) * p(f|r_i)}.$$

$$\text{BIFSG}_s: P(r|g,s,f) = \frac{P(r|s) * P(g|r) * p(f|r)}{\sum_i p(r_i|s) * p(g|r_i) * p(f|r_i)}.$$

$$\text{BIFSG}_f: P(r|g,s,f) = \frac{P(r|f) * P(g|r) * p(s|r)}{\sum_i p(r_i|f) * p(g|r_i) * p(s|r_i)}.$$

A summary of the classification statistics for these three variations is shown in Tables 2.4 and 2.5.

The choice of order in the probability chain rule between leading with geocode – $P(r|g)$ – and leading with surname – $P(r|s)$ – results in comparable performance. However, whereas leading with first name – $P(r|f)$ – yields comparable overall accuracy, as Table 2.4 shows, this choice trades away recall performance in return for greater precision, as Table 2.5 shows. The choice of leading with first name – $P(r|f)$ – also produces a more biased imputation distribution of race/ethnicity compared with the other two approaches. In practice, most practitioners

Table 2.4. Comparison of Accuracy for BIFSG Variations

Percentage of Total Records				
Race/Ethnicity	Actual	$BIFSG_g$	$BIFSG_s$	$BIFSG_f$
White	76%	78%	78%	88%
Black	17%	17%	17%	8%
Hispanic	4%	3%	3%	3%
API	1%	1%	1%	1%
AIAN	1%	0%	0%	0%
Multiracial	0%	0%	0%	0%
Total	100%	100%	100%	100%
Overall accuracy		**83.8%**	**83.9%**	**83.9%**

Table 2.5. Comparison of Precision, Recall, and F1 Score for BIFSG Variations

Race/ Ethnicity	Percentage of Total Records	Precision			Recall			F1 Score		
		Geocode	BISG	BIFSG	Geocode	BISG	BIFSG	Geocode	BISG	BIFSG
White	76%	89%	89%	85%	91%	91%	97%	90%	90%	91%
Black	17%	62%	62%	78%	62%	61%	37%	62%	62%	50%
Hispanic	4%	74%	75%	81%	66%	65%	56%	70%	70%	66%
API	1%	81%	81%	77%	63%	63%	64%	71%	71%	70%
AIAN	1%	65%	64%	80%	40%	40%	30%	49%	49%	43%
Multiracial	0%	3%	3%	3%	0%	0%	0%	1%	1%	0%

choose either to lead with the geocode information as Elliott and colleagues (2009) did in their BISG paper or with the surname information as implemented in the *surgeo* package.

A Caution When Utilizing the Python surgeo Package

Further study into the algorithm under the *surgeo* package reveals an additional variation. The summary data with the first name and race/ethnicity information are provided with counts N_i and $P(r|f_i)$, for each first name f_i. The Bayesian formula is then typically applied to derive, for a given race r and each first name f_i, $P(f_i|r)$ via

$$P\left(f_i \middle| r\right) = \frac{N_i P\left(r \middle| f_i\right)}{\sum_j N_j P\left(r \middle| f_j\right)}.$$

The probabilities used in the *surgeo* package can be replicated by surrendering knowledge of the counts associated with each first name. Assume equal likelihood of each first name in general (i.e., ignore observed counts and assume instead $N_i = N_j$, for all i, j). The above formula would then reduce to

$$P(f_i|r) = \frac{P(r|f_i)}{\sum_j P(r|f_j)}.$$

Tables 2.6 and 2.7 show a comparison of $BIFSG_s$ and $BIFSG_{surgeo}$.

Table 2.6. Comparison of Accuracy for Standard versus *surgeo* BIFSG

Percentage of Total Records			
Race/Ethnicity	Actual	$BIFSG_s$	$BIFSG_{surgeo}$
White	76%	78%	83%
Black	17%	17%	13%
Hispanic	4%	3%	3%
API	1%	1%	1%
AIAN	1%	0%	0%
Multiracial	0%	0%	0%
Total	100%	100%	100%
Overall accuracy		**83.9%**	**84.5%**

Table 2.7. Comparison of Precision, Recall, and F1 Score for Standard versus *surgeo* BIFSG

Race/ Ethnicity	Percentage of Total Records	Precision		Recall		F1 Score	
		$BIFSG_s$	$BIFSG_{surgeo}$	$BIFSG_s$	$BIFSG_{surgeo}$	$BIFSG_s$	$BIFSG_{surgeo}$
White	76%	89%	87%	91%	94%	90%	91%
Black	17%	62%	69%	61%	52%	62%	59%
Hispanic	4%	75%	79%	65%	60%	70%	68%
API	1%	81%	87%	63%	57%	71%	69%
AIAN	1%	64%	69%	40%	37%	49%	49%
Multiracial	0%	2.9%	3.4%	0.4%	0.4%	0.7%	0.8%

Table 2.8. Summary of Predicted Race/Ethnicity from Standard versus *surgeo* BIFSG

Predicted Records by Race/Ethnicity		$BIFSG_{surgeo}$					
		White	Black	Hispanic	API	AIAN	Multiracial
$BIFSG_s$	**White**	4,420,821	0	0	0	0	0
	Black	226,094	742,634	0	0	0	40
	Hispanic	23,095	503	169,613	0	12	28
	API	6,578	428	274	40,038	12	621
	AIAN	2,847	503	0	0	22,042	4
	Multiracial	852	0	0	0	0	1,979

While the reasonableness of surrendering the counts information available in the first name and race summary data and making a uniform occurrence assumption can be debated, Table 2.6 shows that the *surgeo* approach appears to result in improved overall accuracy. In particular, per the recall metric in Table 2.7, the *surgeo* approach improves tagging more of the records listed as white, at the cost of tagging less of the non-white groups. The improved accuracy of the *surgeo* approach comes with a cost in bias in the imputed distribution, as can be seen in Table 2.8. The effect observed is specific to evaluating the imputations on the North Carolina voter registration data and highlights the need for practitioners to consider the choices made in their specific adaptations of the BIFSG algorithm. The classification matching and difference between the $BIFSG_s$ and $BIFSG_{surgeo}$ is summarized in Table 2.8.

The overall match rate between the two imputations is 95.4%. *Surgeo*'s variation in approach leads to greater conservativeness in imputing a record as being of a minority class than if one uses the frequency of first names information. This results in a biased imputation, in return for an improvement in imputation accuracy. This is consistent with what the observations for the other summary statistics shared earlier.

2.4. Data Cleansing – Handling Surname Conventions

Of some interest for the analyst are some of the common conventions in the surname data. This section discusses a few common conventions and explores their treatment, both to illustrate some practical explorations and to illustrate the difficulties in cleansing the names data. This treatment is by no means complete in breadth or in depth, and the analyst is encouraged both to develop techniques for standardizing names data and recognize the cost–benefit tradeoffs associated with such efforts.

Generational Suffixes

The census surname demographics summary dataset does not take into consideration generational suffixes (e.g., JR, SR, III, IV). Thus, it is assumed that generational suffixes are removed from the surname column before going through the imputation process.

Spaces, Hyphens, and Apostrophes

Of the 8.4 million records in the North Carolina voter registration data, approximately 111,000 records contain spaces in the surname, 103,000 records contain hyphens, and 9,000 records contain apostrophes (with a select few situations where multiple phenomena are present). Together these records represent 2.6% of the total data. Thus, a reasonable resolution may be to develop holistically some basic processing rules and move forward.

In dealing with spaces, a difficulty arises in that there is not a dominant common prefix causing the issue. For example, summarizing by the leading three characters of surnames with spaces shows no single three-leading-character combination that exceeds 6,000 occurrences. However, a survey of the data shows that multiple surnames (a variation of hyphenated last names) account for a large number of occurrences. In these situations, reasonable choices include matching to one of the names or matching to the compound name. Where the race or ethnicity is comparable, the choices lead to comparable components.

As Table 2.9 shows, where the two names are of similar percentages (e.g., both LOPEZ and HERNANDEZ lead to greater-than-90% Hispanic probability and small percentages for the other race and ethnicity groups), the choice of method should generally lead to comparable final imputed results. However, where the two names have different probability distributions, the choice of methods can lead to different imputation outcomes. These variations should be considered in the context that hyphenation and spaces represent in total a small percentage of all the records, and the compound name surname situations make up a subset of this segment of the data.

There are a few other situations worth mentioning. Prefixes such as "VAN," "ST," or "DE" are another common reason for spaces in surnames. In these situations, a survey of the surname file shows that it is reasonable to make an exact match with spaces removed.

Table 2.9. Race/Ethnicity Probabilities for Surname Adjustments to "Hernandez-Lopez"

Method	Name	Count	White	Black	Hispanic	API	AIAN	Multiracial
Exact lookup	HERNANDEZLOPEZ	1,508	1.1%	0.1%	98.6%	0.1%	0.1%	0.0%
First surname	HERNANDEZ	1,043,281	3.8%	0.4%	94.9%	0.6%	0.2%	0.2%
Second surname	LOPEZ	874,523	4.9%	0.6%	92.9%	1.0%	0.4%	0.3%
Reverse surname	LOPEZHERNANDEZ	1,893	1.6%	0.1%	97.8%	0.1%	0.1%	0.1%
Straight avg*	–	–	4.3%	0.5%	93.9%	0.8%	0.3%	0.2%
Weighted avg**	–	–	4.3%	0.5%	93.7%	0.9%	0.3%	0.2%

* Straight avg: $P(r = \text{Hispanic}|s) = (94.9\% + 92.9\%)/2 = 93.9\%$.
** Weighted avg: $P(r = \text{Hispanic}|s) = [1.04/(1.04 + 0.87)] * 94.9\% + [0.87/(1.04 + 0.87)] * 92.9\% = 93.7\%$.

Alternatively, stripping the prefix and matching the main body of the surname may be serviceable, particularly if the exact match is not available.

Two common forms of apostrophized prefixes are with O' and with D.' The exact lookup (without the apostrophe) can be found in common cases. There are a few interesting observations to be made here. First, generally, the prefixes D' and O' reflect surnames that are predominantly listed as white in the census data. The proportion listed as white drops notably when the prefix is dropped in the lookup. Second, this observation is not universal. The common O' prefix name O'NEAL shows meaningfully different race and ethnicity distributions than O'BRIEN and O'CONNOR (and D'ANGELO and D'AMICO). Names containing apostrophes represent a small segment of the North Carolina voter registration data. Thus, efforts to enhance the imputation process to better predict this segment of the data will likely have limited overall benefit in the analysis.

Additional Options for Missing Surname Joins

As a default for missing surname joins, one can use the All Other Names entry from the census data for purposes of calculating the imputation (Table 2.10). Although crude relative to the above-mentioned treatments, the simplicity in implementation is a benefit with this option.

Alternatively, where the surname match fails, one can shift to a coarser imputation approach. As the surname probabilities would be needed for BIFSG and BISG, one can use the geocode-only probabilities for imputation purposes. That is, one can use the census race/ethnicity percentages by geocode directly, without the surname and first name adjustments.

This idea can be similarly leveraged where the first name join fails against the mortgage summary data. The researcher can either use the All Other First Names entry in the mortgage summary file or shift to imputing using the BISG process, which does not require matching the first name.

These represent but a few of the surname complications that may arise. While this section has presented a variety of basic approaches to handle these complications, researchers are encouraged to explore their own treatment of data cleansing, while balancing the benefits they expect to gain relative to the efforts put forth.

The foregoing discussion illustrates the probabilities associated with $P(r|s)$, which would be employed as part of $BIFSG_s$. Similar explorations can be done with $P(s|r)$, the component used for $BIFSG_g$ and $BIFSG_f$.

Table 2.10. Race/Ethnicity Probabilities for "All Other Names" Census Surname Category

Category	Count	White	Black	Hispanic	API	AIAN	Multiracial
All Other Names	29,312,001	66.7%	8.5%	13.7%	8.0%	0.9%	2.3%

2.5. Utilizing Imputed Probabilities

The geocode-only, BISG, and BIFSG imputation approaches result in probabilities for each of the six race and ethnicity categories as categorized in the geocode, surname, and first name files: white, Black, Hispanic, API, AIAN, and multiracial. The following are three common ways to use the probabilities from any of these three approaches:

1. Classification for each record, based on maximum probability
2. Use of each probability, as imputed
3. Simulated classification – random assignment of race and ethnicity, with likelihoods based on the imputed probabilities

The following example illustrates the three approaches. Appendix A provides the R code that was used to create this illustration. Consider a simulated dataset, with the following information:

- The number of records is $N = 1{,}000$.
- Each has a likelihood of being class A based on a uniform draw on probabilities between 0 and 1.
- A random draw is made for each record to determine class, based on the probabilities determined above.
- For class A, the probability that outcome = 1 is 0.6. For class B it is 0.4.
- A random draw is made based on the above-defined probabilities to determine outcome.

Table 2.11 shows the first 10 records from this simulation.

Given the probabilities, an imputation can be made for each record by choosing the class with the maximum probability. The results can be found in the "Class Based on Max Imputed Percent" column (column e). The imputed class A feature can be developed from this column:

$$X_A: x = 1 \textit{ if } \textit{Imputed Prob A} (\textit{col b}) > \textit{Imputed Prob B} (\textit{col c}), \textit{and } 0 \textit{ otherwise}.$$

A second manner is to use the probabilities directly ("Imputed Prob A," column b) to develop the class A probability feature:

$$X_A: x = \textit{Imputed Prob A} (\textit{col b}).$$

This creates a probability feature instead of a binary feature and is usable for modeling and statistical summary computations in many instances. These probabilities can serve as more meaningful weights than the imputed classifications/counts for purposes of analysis that is based on summarizing data, such as premium parity or loss ratio parity. Furthermore, the probabilities can be useful for developing and evaluating visual tests. In particular, the

Table 2.11. Approaches to Utilize Imputed Probabilities – Simulated Example

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
ID	Class	Imputed Prob A	Imputed Prob B	Outcome	Class Based on Max Imputed Percent	Sim 1	Sim 2	Sim 3	Sim 4
1	A	72%	28%	0	A	A	A	A	A
2	A	88%	12%	0	A	A	A	A	A
3	A	76%	24%	1	A	A	B	A	A
4	A	89%	11%	0	A	A	A	A	A
5	A	46%	54%	1	B	B	A	B	A
6	B	17%	83%	1	B	B	B	B	B
7	B	33%	68%	0	B	B	B	A	B
8	B	51%	49%	0	A	A	B	B	B
9	B	73%	27%	0	A	A	B	A	B
10	A	99%	1%	0	A	A	A	A	A

relevant target or fairness metric can be plotted against the probabilities or probability bands rather than in a binary nature and reviewed for patterns or lack of patterns.

Finally, instead of taking the class with maximum probability, a class feature vector can be developed through simulation. Randomly draw between A and B using the probabilities, rather than deterministically assigning the class value with the highest probability. Four simulated class columns are shown in Table 2.11 in the columns "Sim 1" through "Sim 4" (columns f–i). The analysis can be run repeatedly on simulated class features:

$$X_A: x = 1 \textit{ if simulated class is } A, \textit{ and } 0 \textit{ otherwise}.$$

This last approach has the benefit of providing a distribution of model outputs or statistical measures, at the cost of increased processing time.

For those interested in carrying out the described computations and evaluating proportions of outcome = 1 for each class in Table 2.11, Table 2.12 shows a summary for comparison. In every case, the proportion for class A can be computed as

$$Proportion = \frac{\sum x_A * Outcome}{\sum x_A}.$$

With 1,000 records and 100 simulations, Table 2.12 shows the results of each classification approach.

Table 2.12. Summary of Simulated Outcomes for Approaches to Utilizing Imputed Probabilities

	Outcome			Simulation Results (100 simulations)			
Class	Actual	Classify Based on Max Pct	Direct Use of Pct	mean	min	median	max
A	60%	54%	53%	53%	49%	53%	56%
B	39%	45%	46%	46%	43%	46%	50%

Please note that the purpose of the foregoing example is to illustrate some of the uses of the imputation outcomes only. In practice, the relationships between features and outcomes are more complex than the example shown here. Practitioners should carefully consider the possible sources of errors as they evaluate the results that utilize the feature imputations. To further complicate matters, in practice the actual distribution is unknown. While the population distribution can be used as a reference, it is possible the underlying distribution of the data under evaluation may not reflect that of the population.

Section 3. Fairness Criteria to Identify and Measure Potential Disparities

3.1. Relationship of Fairness Criteria to Pricing Models

In an earlier paper in this series, Mosley and Wenman (2022) explained how three types of fairness criteria apply to binary classification models, i.e., models that produce an output of either 0 or 1. Binary classification models are commonly used in the insurance industry for use cases such as fraud detection and claims segmentation, but these definitions need to be modified to apply to rating models built to predict a continuous response variable. This paper continues this discussion by providing additional detail on how these concepts apply to regression models, such as loss cost models used for pricing, and tying the general-purpose fairness criteria to traditional insurance metrics such as premium and loss ratio.

The three fairness criteria are

- **independence –** the model's predictions and the protected attribute are statistically independent;
- **separation –** conditional on the *observed response variable*, the model's predictions and the protected attribute are statistically independent; and
- **sufficiency –** conditional on the *model's predicted values*, actual outcomes and the protected attribute are statistically independent.

Table 3.1. Hypothetical Example Where Effect of Rating Factor Differs by Protected Group

Full Information (Protected Group Known)					
Level of Rating Factor (R)	Protected Group (k)	Exposure ($X_{R,k}$)	Losses ($L_{R,k}$)	Loss Cost ($\bar{L}_{R,k} = L_{R,k}/X_{R,k}$)	Loss Cost Relativity ($\bar{L}_{R,k}/\bar{L}_{N,a}$)
N	a	90	$90,000	$1,000	1.00
N	b	10	$8,000	$800	0.80
Y	a	90	$135,000	$1,500	1.50
Y	b	10	$10,000	$1,000	1.00

These are broad criteria that can be satisfied by more than one metric, and this paper examines their application to metrics that may be appropriate for regression models. This discussion will use the following notation:

- Y denotes a response variable and $\hat{Y}$ is an estimator for Y.
- A represents the protected attribute, with levels a and b. For k equal to either a or b,
  - P_k is the aggregate premium charged to members of group k;
  - L_k is the aggregate losses among members of group k; and
  - X_k is the aggregate exposure among members of group k.

The concepts in this section will be demonstrated using the following scenario, which illustrates how a loss ratio disparity can arise when the effect of a rating factor is different for members of different groups.[5] Assume the loss experience shown in Table 3.1 exists for groups a and b within a protected class, segmented by a binary rating factor, R, having levels "N" and "Y."

The "full information" loss cost relativity is what could be calculated using both the rating factor and the protected class. In practice,[6] it may not be possible or permissible to review loss experience by protected group. From the perspective of an analyst who does not have information on the protected attribute, the data would look like that in Table 3.2.

Using these data would lead to a one-way relativity of 1.48 for level Y of the rating factor, using level N as the base level. Comparing this with the full information relativities illustrates a potential source of unfairness under the sufficiency fairness criterion; although the effect

[5] As discussed in Section 1, multiple potential sources of bias could affect a model. Appendix B contains a variation on this example in which the source of bias is use of a proxy variable. The reader may find it instructive to compare and contrast these two examples.

[6] Use of protected class information in insurance rating varies across jurisdictions. In some jurisdictions, characteristics that are protected for general usage – such as gender, age, and marital status – are permitted for insurance rating purposes.

Table 3.2. Calculating Relativities without Knowing Protected Group

Protected Group Unknown				
Level of Rating Factor (R)	Exposure (X_R)	Losses (L_R)	Loss Cost ($\bar{L}_R = L_R/X_R$)	Loss Cost Relativity ($\bar{L}_R/\bar{L}_N$)
N	100	$98,000	$980	1.00
Y	100	$145,000	$1,450	1.48

of the rating factor is lower for the minority group *b,* the estimate of the one-way relativity is driven by the majority group *a* due to higher data volume. The result is a relativity that is too high for the minority group *b* and slightly too low for the majority group *a.*

3.2. Testing Independence Using Premium Parity

For binary classification models, the independence criterion is satisfied when the following condition, known as *demographic parity,* is met:

$$P\left(\hat{Y} = 1 \mid A = a\right) = P\left(\hat{Y} = 1 \mid A = b\right).$$

When Y is binary, this is equivalent to

$$E\left[\hat{Y} \mid A = a\right] = E\left[\hat{Y} \mid A = b\right].$$

In other words, the average model prediction is equal for each level of the protected attribute. The latter formulation can be applied to models with a continuous response variable because it does not rely on the probability that a binary variable is equal to 1. This criterion can be referred to as premium parity when the model involved is a model for charged premium. In this context, a demographic parity test for an insurance pricing model corresponds to checking that the average premium is the same, within a materiality threshold, for each level of the protected variable. The test can also be used to compare two models by identifying which one has a lower premium disparity among groups.

Because the focus of a premium parity test is on a specific pricing algorithm, rather than the company's historical premium over multiple rate changes, it is assumed throughout this section that all premiums are at the level of the algorithm that is being evaluated for fairness. If a proposed rating model is being assessed, this means that the historical dataset used to perform the test needs to be re-rated using the proposed rating model. If we are monitoring the fairness of a pricing algorithm that is currently in effect, re-rating would not be needed, but it is important to use only premium data for policies rated using the current model.

Table 3.3 illustrates the concept using the scenario described earlier, where a rating factor of 1.48 is used for level *Y* of the rating factor. The first step is to calculate the premium

Table 3.3. Rating Structure Based on One-Way Relativities

Level of Rating Factor (R)	Relativity	Protected Group (k)	Exposure ($X_{R,k}$)	Individual Customer Premium	Total Premium ($P_{R,k}$)
N	1.00	*a*	90	$1,508	$135,720
N	1.00	*b*	10	$1,508	$15,080
Y	1.48	*a*	90	$2,232	$200,880
Y	1.48	*b*	10	$2,232	$22,320

Table 3.4. Summarizing Proposed Premium by Protected Group (One-Way Relativities)

Protected Group (k)	Exposure (X_k)	Premium (P_k)	Average Premium (P_k/X_k)
a	180	$336,600	$1,870
b	20	$37,400	$1,870

under the proposed rating structure. To fully define the rating structure, assume a base rate of $1,508.[7]

Next, as shown in Table 3.4, the data can be summarized by protected group to calculate the average premium.

In this scenario, the premium parity criterion is satisfied because the average premium is the same for groups *a* and *b,* which effectively demonstrates that the proposed rating structure is independent of the protected attribute (i.e., uncorrelated). This implies that in order to satisfy the independence criterion, the distribution of the protected attribute must be consistent across all levels of the proposed rating structure, or that when distributional differences exist in the rating structure, the differences have offsetting effects on the average premium.

Now suppose an alternate rating structure (Table 3.5) is used based on the full information two-way relativities from Table 3.1, using a base rate of $1,538.[8] Under this rating structure, the premiums would be as shown in the table.

Table 3.6 shows the results when summarized by protected group.

Under this alternate rating structure, premium parity is not achieved due to the difference in average premium for groups *a* and *b,* which is caused by differences in the actual loss

[7] In this example, the overall loss cost is $1,215, and assuming a target loss ratio of 65% requires an average premium of $1,869. This is what results in a base rate of $1,508.
[8] This was selected to produce the same overall rate level as the rating algorithm in Table 3.3 to ensure that differences between the two algorithms do not result from changes in the overall rate level.

Table 3.5. Rating Structure Based on Full Information Relativities

Level of Rating Factor (R)	Protected Group (k)	Exposure ($X_{R,k}$)	Full Information Individual Premium	Full Information Total Premium ($P_{R,k}$)
N	*a*	90	$1,538	$138,420
N	*b*	10	$1,230	$12,300
Y	*a*	90	$2,307	$207,630
Y	*b*	10	$1,538	$15,380

Table 3.6. Summarizing Proposed Premium by Protected Group (Full Information Relativities)

Protected Group (k)	Exposure (X_k)	Full Information Alternate Premium (P_k)	Average Premium (P_k/X_k)
a	180	$346,050	$1,923
b	20	$27,680	$1,384

experience by protected group and use of a rating model that picks up on those differences. In this example, premium parity can be achieved only at the expense of other notions of fairness: specifically, with Table 3.5 providing individual premiums that more closely correspond to loss experience, a comparison to Table 3.3 shows that customers in group *b* are overcharged, while customers in group *a* are undercharged. In later sections, this paper explores this from a loss ratio perspective, and then as a general phenomenon, i.e., that a difference in actual loss experience leads to a need to make choices between different fairness criteria.

In practice, significant work is involved in creating the theoretical models underlying the premium calculation for a book of business, and the model output is often adjusted, such as through selections and expense or profit loading, to produce the pricing algorithm. Fairness tests should be performed on this final pricing algorithm to assess the actual impact on policyholders most accurately. However, given the significant amount of time involved in developing a rating model, it is also prudent to perform preliminary fairness tests to provide an early indication of potential fairness concerns. For example, tests of premium parity can be performed on the theoretical models used to develop the pricing algorithm, such as checking that the average prediction of a loss cost model is similar for each group. Assuming that adjustments do not materially change the relative ordering of premiums – as is typically the case when loading for expenses and profit – a disparity in a theoretical model would lead to a disparity in the charged premium, so performing a demographic parity test on a theoretical model provides an early indication as to whether there might be a disparity in the charged premium. Performing the test before and after selections would provide insight as to whether business adjustments to the model may have introduced a disparity.

An additional practical consideration when performing a premium parity test is determining which data will be used to calculate the metrics and whether there are any associated limitations. Such considerations also affect what notions of "premium" will be used. Fairness should be considered throughout the model life cycle, with premium parity being relevant during the development and monitoring stages of the life cycle. The approach will vary depending on the point in the model life cycle at which the calculations are being performed.

During the development stage of the model, some examples of approaches to evaluating fairness of rates prior to implementation include the following:

1. Testing the predictions from a theoretical rating model, based on training data. This allows for early indications of potential fairness concerns.[9]
    a. **Pro:** Using modeling data for tests at this stage has the significant advantage that the data are readily available.
    b. **Con:** The output of a theoretical model is generally not the premium that is charged to the customer, so this needs to be supplemented with later tests on the final pricing model.
2. Testing the proposed rating algorithm on renewal policies:
    a. **Pro:** This allows for an assessment of fairness concerns for policies currently written by the company. Assessing impacts that a proposed rating algorithm has on current customers is already part of standard pricing processes, minimizing the amount of incremental effort.
    b. **Con:** This does not assess fairness for potential new customers, which is a concern when the distribution of the current book of business differs materially from the market of potential customers.
3. Testing the proposed rating algorithm on historical quote data:
    a. **Pro:** This may provide a better indication of fairness across the market of potential customers than inforce data would.
    b. **Con:** This adds complexity to the fairness testing process, as it requires re-rating old quotes using the model that is being developed. Quotes for customers who did not purchase a policy will not be available in the model training data.

During the monitoring stage of the model, approaches to assessing the fairness of rates post-implementation include the following:

1. Testing the current written premium for policies starting after the effective date of the current rating plan:
    a. **Pro:** Data are straightforward to obtain, as policies have already been rated and premiums recorded in the policy database.

[9] While not strictly a premium parity approach, comparing average losses by protected group within the training data can give an even earlier indication of a potential premium disparity, since a disparity in the model response variable could lead to a disparity in model predictions.

   b. **Con:** This will only capture customers who have purchased a policy, and it may fail to detect situations in which an unfairness in the rates has dissuaded customers from purchasing a policy.

2. Testing premium quotes produced after the effective date of the current rating plan:
   a. **Pro:** This can mitigate the impacts that consumer purchasing decisions can have on the results.
   b. **Con:** This could produce a biased sample of the marketplace, with customers who tend to shop around being overrepresented.

A decision that needs to be made across all approaches is choosing the number of years of data to use in the parity calculations. The broad objective is to base the analysis on data that have both a credible volume and are representative of the market of potential customers. Some considerations that can inform this decision include the following:

- If more than one year of data is used, an individual customer may appear in the data multiple times, which could distort the results. For example, if historical policy data are used, long-standing customers (who are presumably satisfied with their policy) will be overrepresented in the analysis. If quotes are used, customers who tend to shop around frequently will be overrepresented.
- Older data may not be representative of the market of potential future customers if material changes have occurred in the book of business resulting from deliberate changes in the company's risk appetite.
- Older data could be representative of the market of future potential customers if shifts in the book of business are a result of customer attrition, and if a new pricing algorithm is expected to attract previous customers.
- During the monitoring phase, limited data could be supplemented with re-rated historical data, but this adds additional complexity to the monitoring process.

Affordability Metrics

Variations on the premium parity methodology can incorporate other concerns related to fairness in insurance pricing, such as affordability of insurance. An example of an affordability metric would be the ratio of premium to total income, effectively performing a premium parity test after first normalizing based on income. Parity of the premium-to-income ratio would correspond to the notion that insurance should be *equally affordable* regardless of membership in a protected group. This approach has the advantage of being simple and easy to explain, but one of its limitations is that it does not account for differences in coverage selected, and individuals in low-income areas may select lower coverage to help with affordability concerns. The method could be refined by adjusting the premium data for differences in coverage prior to calculating a premium-to-income ratio.

An alternate approach could be to define a binary "affordable" indicator, based on the premium being below a certain income-sensitive threshold, and examine parity of this

indicator. This approach can avoid potential distortions resulting from high-income individuals having very low premium-to-income ratios. The corresponding notion of fairness would reflect whether the premium is below the affordability threshold, rather than *how much* it is below the threshold.

In practice, the income of an individual insured is not typically known, nor is there an algorithm analogous to BIFSG for inferring income. In the United States, because of the lack of precision at the individual level, affordability is often measured at a geographic level using census data related to income. For example, the Federal Insurance Office (2017), in its *Study on the Affordability of Personal Automobile Insurance*, calculates an affordability metric as average written voluntary market premium within a zip code divided by median household income within that zip code; it considers insurance to be affordable if this metric is below the national average of 2%.

Alternate approaches could use different measures of income in the census or measure affordability on a comparative rather than binary basis (e.g., an index of 1.8% is more affordable than 1.9%). Affordability parity for a protected class can be assessed by using census data related to that protected class. In other words, one could examine the extent to which insurance is less affordable in zip codes that have a higher percentage of a protected group.

3.3. Testing Separation Using Loss Ratio Parity

For classification models, the separation criterion corresponds to checking whether the error rates (false negative[10] and/or false positive rates) are the same for each protected group of interest. This involves comparing the model's predictions to actual outcomes. An example of a separation condition for a classification model is *equalized odds*, which is that

$$P\left(\hat{Y}=1\middle|Y=y,A=a\right)=P\left(\hat{Y}=1\middle|Y=y,A=b\right),$$

for $y=0$ and $y=1$. (A weaker separation condition, *equal opportunity,* requires only that this condition hold when $y=1$, i.e., that true positive rates are the same for both protected groups.)

For pricing models, there is no direct analogue to the formulas used for classification models because separation conditions such as equalized odds are conditional on *actual* outcomes, whereas the purpose of a pricing model is to predict the *expected* outcome. The vast majority of individuals will have no claims, which makes it much less meaningful to condition on actual claims outcomes. Dolman and Semenovich (2018) propose that for insurance premiums this condition be adjusted to

$$E\left[\hat{Y}\middle|\mu,A=a\right]=E\left[\hat{Y}\middle|\mu,A=b\right].$$

[10] False negative rate parity is mathematically equivalent to true positive rate parity, which is a term commonly used in fairness literature. Similarly, false positive rate parity is mathematically equivalent to true negative rate parity.

With this approach, μ is the "true risk type" of the individual, which in an insurance context would be the theoretical pure premium. They refer to this condition as *actuarial group fairness and* observe that the condition will be satisfied any time the premium is a function of μ alone, such as by pricing to a constant loss ratio. In practice, expense and profit loads may vary by client, in which case this condition corresponds to assessing whether there are material differences in expense and profit loading by protected group. With "true" risk being unobservable, they propose using a model for μ; however, a drawback of this approach is that the model for μ could itself be biased.

This section explores an approach for calculating loss ratios based on actual historical claims for each protected group, which avoids the need to use a model. When modeling a continuous variable, the error is also continuous, so the natural analogue of looking at error rate metrics is to compare the charged premium to actual losses. Calculating the loss ratio for each group is a natural way to test this condition, because it is based on a traditional actuarial metric. The condition that $L_a/P_a = L_b/P_b$ is referred to as *loss ratio parity.* In practice, achieving exact equality is not feasible due to volatility of claims data, so loss ratio parity tests generally assess whether the difference in loss ratio among groups is within a specified tolerance level or, when comparing models, identify which model has a smaller disparity in loss ratios among groups.

Revisiting the scenario described in Table 3.1 and applying the proposed rating structure from Table 3.4 produces the results displayed in Table 3.7. Here, the rating structure does not achieve loss ratio parity due to the difference in overall loss ratio between groups *a* and *b;* this is in contrast with Table 3.4, where premium parity *is* achieved.

However, applying the proposed rating structure from Table 3.6 to the scenario described in Table 3.1 produces the results displayed in Table 3.8. The full information two-way rating structure satisfies loss ratio parity in this scenario. Table 3.8 is in contrast with Table 3.6, in which this rating structure *did not* achieve premium parity.

Table 3.7. Summarizing Loss Ratio by Protected Group (One-Way Relativities)

Protected Group (k)	Losses (L_k)	Premium (P_k)	Loss Ratio (L_k/P_k)
a	$225,000	$336,600	66.8%
b	$18,000	$37,400	48.1%

Table 3.8. Summarizing Loss Ratio by Protected Group (Full Information Relativities)

Protected Group (k)	Losses (L_k)	Full Information Alternate Premium (P_k)	Loss Ratio (L_k/P_k)
a	$225,000	$346,050	65.0%
b	$18,000	$27,680	65.0%

Because loss ratio parity tests involve claims data, the analyst needs to consider two competing criteria when selecting data:

1. Because the loss ratio test involves comparing predictions of claims to actual claims, as with any predictive model, the testing data should be independent of the training data to assess the model's performance more realistically on "unseen data."
2. Because claims data are more volatile than data on rated premium, data volume needs to be high enough to provide a credible result. Data volume concerns can be amplified when performing fairness tests, since such tests involve segmenting the data based on protected group, some of which may have low data volume and higher expected variance in the actual outcomes. Comparing results on a confidence interval basis can help the analyst understand whether there is enough data to draw conclusions based on the test.

The ideal approach would be to perform the test using holdout data from the modeling process, provided data volume is sufficient. If holdout data volume is low, this test could be supplemented by tests on training data or combined training and holdout data. However, if training data are used in a loss ratio test, the results should be interpreted with caution; tests that use training data can overstate model performance, and as a result, disparities in loss ratios among groups could be understated because the loss ratios for each group are more likely to be closer to a target loss ratio than they would be on holdout data. (This concern did not arise in the discussion of premium parity because the premium parity approach is not conditioned on actual claim outcomes. Rather, premium parity tests only the mechanics of the rating algorithm.) When a generalized linear model is used, after holdout testing is complete, it is standard practice to refit the model on combined training and holdout data to increase credibility of the final model. In this case, the loss ratio test should be redone using the final model, using combined training and holdout data, to confirm that results are consistent with what was observed during holdout testing.

Like any other analysis involving premium and losses, adjustments for trend, development, and change in benefit level should be considered. (Recall that the discussion on premium parity relied on the assumption that premium data are at the level of the rating algorithm under evaluation, so additional premium adjustments are unlikely to be needed.) However, when performing a disparity test, consider that multiplicative adjustments will not have a material impact on the results if the magnitude of the adjustment is similar for each protected group; in this situation a disparity would exist before the adjustment if and only if it exists after the adjustment. Situations in which the differences may be material include longer-tailed lines, books of business where the distribution of protected groups has been shifting over time, or situations where protected groups experience different trend rates or development patterns.

Much like premium parity, loss ratio parity can be assessed throughout the model-building process by assessing the components that contribute to the rating model. Some examples include the following:

1. Initial tests can be performed on a theoretical loss cost model by assessing the parity of the ratio of aggregate losses to aggregate model predictions (weighted by exposure) for each group.

2. To assess the fairness of a proposed pricing algorithm, once the algorithm is finalized, the test from example 1 can be repeated using the actual charged premium, weighted based on the earned exposure underlying the claims data in place of the theoretical loss cost. The differences between this approach and the test described in example 1 would be material only if differences exist between the indicated and proposed rating factors. Comparing the results from this test and the test in example 1 can assess whether disparities in the proposed pricing algorithm can be attributed to disparities in the indicated plan or whether they can be attributed to adjustments made to the indicated plan, i.e., whether the differences between the indicated and proposed plans are correlated with protected class.
3. To monitor rates that are currently in effect, loss ratio parity can be monitored based on the ratio of actual losses to earned premium for each group, using data after the effective date of the rating plan.

A limitation across all three of these approaches is that the data are based only on policies that are already part of the company's book of business, so if the distribution of the company's current book of business differs from that of the potential customer base, results could be distorted. For example, if a rating model is consistently overestimating the premium for a group, then the corresponding loss ratio disparity could go undetected because members of that group may not purchase policies in the first place. This limitation is generally unavoidable given that loss ratio tests rely in a fundamental way on claims that are actually on a company's book, but the impact of this limitation could be assessed by monitoring the distribution of the book of business compared with the general public.

When a loss ratio parity test is performed on a one-way basis, it provides an assessment of bias and potential for unfair discrimination in the overall rating structure. The American Academy of Actuaries (2023b) recommends performing a loss ratio parity test on a two-way basis, considering protected class and a rating factor of interest. Consistency of loss ratios across both dimensions indicates that the rating factor is equally predictive across the protected groups. This approach can be effective in detecting situations in which the rating factor is a source of bias on the grounds that the effect is not consistent from one group to another (i.e., aggregation bias), or in other words, when there is an *interaction effect* between protected class and the rating factor. For this reason, this test is sometimes also referred to as an *interaction test.* Interactions such as this can arise through bias in the data values (in situations in which there is inconsistency involved in the collection of data on the rating factor) or through omitted variable bias (in situations where there is a true interaction present that is not being recognized in the model). When a loss ratio disparity is detected in the overall rating structure, performing this test for each rating factor can help identify any rating factors that could be contributing to the overall disparity.

3.4. Testing Sufficiency Using Lift Charts and Loss Ratio Parity

The previous section argued that while loss ratio parity doesn't exactly match the definition of separation (because separation conditions on actual rather than expected outcomes), it is a reasonable analogue of separation metrics because both look at parity of model error across groups. This section presents an argument that loss ratio parity, while not

precisely meeting the definition of sufficiency, is also a reasonable analogue of the sufficiency condition for a model with a continuous response. Conceptually, separation corresponds to all groups having the same premium after conditioning on losses – which means having the same loss ratio at each loss level. The sufficiency criterion is conditioned on premium to check whether each premium level has the same amount of losses – which would mean having the same loss ratio at each premium level. The lack of a clear one-to-one correspondence between parity metrics for models with a continuous response and models with a categorical response is consistent with the fact that classification models have two discrete types of errors (false positives and false negatives) and multiple types of error rates, while regression models with a continuous response do not directly fit into this framework.

For classification models, the sufficiency condition corresponds to parity of metrics such as the false discovery rate (probability of a false positive, given that a model has predicted positive) and false omission rate (probability of a false negative, given that a model has predicted negative) among protected groups. There are also more stringent conditions such as *calibration* and *well-calibration* that are described in Mosley and Wenman (2022). Like separation, these tests are based on a comparison of model predictions to actual outcomes, and for a premium model, a sufficiency test would similarly involve a comparison between premium and actual losses.

More formally, for a classification model, let R be a model's estimate of the probability that $Y = 1$. The calibration condition is defined as satisfying

$$P\left(Y = 1 \middle| R = p, A = a\right) = P\left(Y = 1 \middle| R = p, A = b\right)$$

for all probabilities p between 0 and 1.

For models with a continuous response (in the pricing use case, with premium P), the calibration condition generalizes to

$$E\left[Y \middle| P = p, A = a\right] = E\left[Y \middle| P = p, A = b\right],$$

where P is the model's prediction – in this case, the premium. Given that the number of unique values for premium is typically large, this may not be a meaningful calculation unless the premium is divided into buckets prior to calculating average losses within each bucket. Essentially, this corresponds to producing a quantile plot for each protected group, much like the traditional quantile plots that are used for evaluating the accuracy of rating models. (See, for example, Section 7.2.1 in Goldburd et al. [2020].) Lift charts can provide a visual confirmation of whether the calibration condition is satisfied. This would involve verifying that there is a lack of a consistent pattern of differences in losses across premium levels. When performing this test, the quantiles need to be defined across the whole dataset, rather than separately for each protected group, to ensure that a consistent set of premium quantiles is used for each protected group. This introduces a risk that the amount of data in each bucket could be different by protected group, so results need to be interpreted

carefully in lower-volume buckets to ensure that a lack of pattern isn't driven by usual volatility. The choice of the number of quantiles selected can be informed by the need to have sufficient data volume for minority groups in each bucket. An alternate approach to ensuring adequate data volume in each bucket could be to generate quantiles based on the premium for members of the minority group (to ensure even data volume among buckets among the group for whom data are sparsest) and use the same quantiles for the majority group.

This approach is closely related to loss ratio parity, because if both groups have the same losses for a given premium level, then they will have the same loss ratio at that premium level. The differences are the following:

- Loss ratio parity, as described in the previous section, considers parity only at a "global" level rather than ensuring that parity is consistent at different premium levels. (If it is reasonable for us to assume that the loss ratio is the same across all premium levels, then a global loss ratio parity does satisfy the calibration condition.)
- The calibration condition does not require that the loss ratio be the same across premium levels. Therefore, if there is a nonuniform distribution of protected groups across premium levels, it could result in a situation where loss ratio parity is achieved at each premium level, but not globally.

The examples in Figures 3.1 and 3.2 illustrate some possible outcomes of a visual sufficiency test and how to interpret those results. These are artificial examples created for explanatory purposes, and as such they will look "cleaner" than typical results.

Figure 3.1. Example of Visual Test in Which Sufficiency Is Satisfied

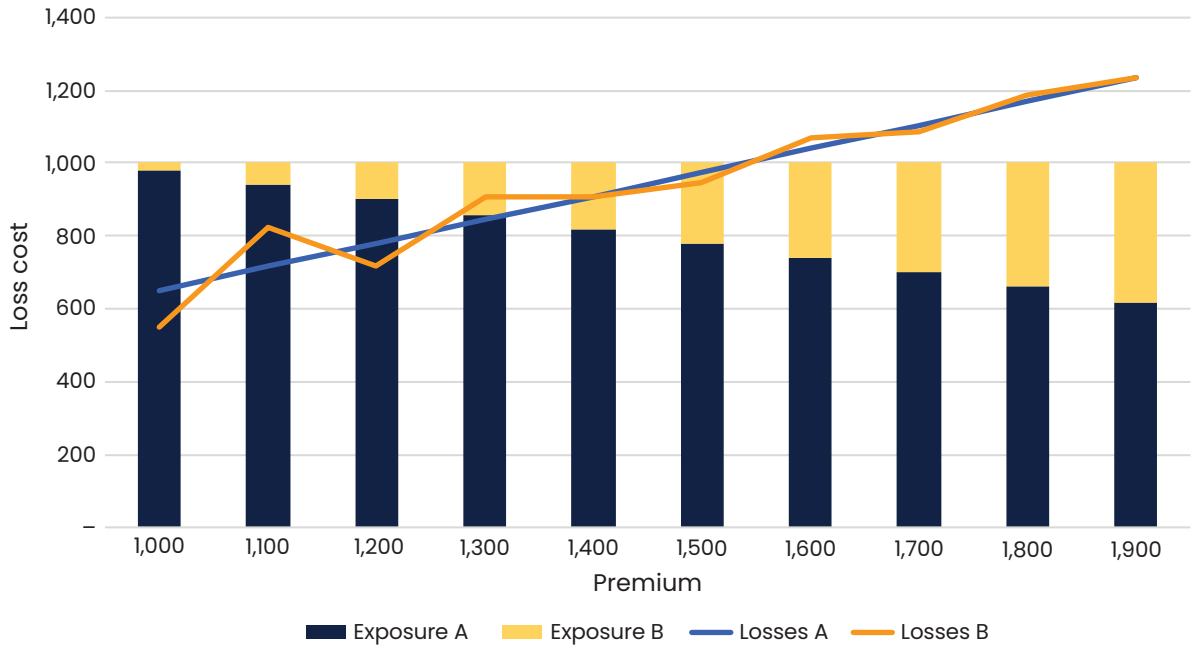
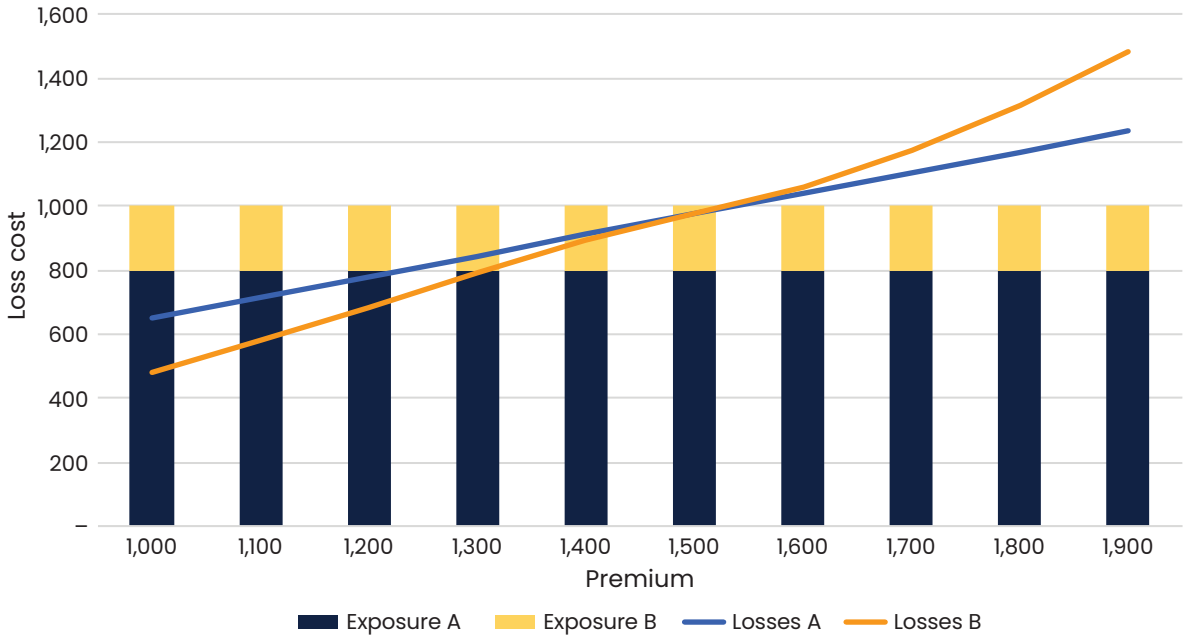


Figure 3.2. Example of Visual Test in Which Sufficiency Is Not Satisfied



The model passes the test when losses are similar across each bucket. In the example shown in Figure 3.1, even though there is volatility in the losses for group B on the left side of the chart, that can be explained by the lower exposure for group B. On the right side the losses for group B roughly track those of group A, indicating that this model passes a sufficiency test. In contrast, consider the example in Figure 3.2.

In this example, the model does not pass a sufficiency test, because losses for the two groups are different at each premium level. The consistency of the pattern – group B having lower losses than group A when premium is low, but the reverse being true when premium is high – suggests that this is *not* being driven by volatility of the data, as was the case in the previous example. Notably, in this example, the two groups in this example achieve aggregate loss ratio parity – each group's loss ratio is roughly 65% – but it does not satisfy the calibration test because loss parity is not consistent across premium levels.

3.5. Comparing Premium Parity and Loss Ratio Parity

A key advantage shared by both premium parity and loss ratio parity is their ease of calculation and explainability. They are based on familiar actuarial metrics – and checking these metrics for parity involves the same calculations that would be performed to produce the traditional breakdowns of premium and loss ratio by the levels of a rating factor. They should be easy to apply consistently by a wide range of insurers of various levels of analytical sophistication. Transparency is a key advantage when assessing models for bias, because a bias assessment will be less convincing to a wide range of stakeholders if it involves a method that is complex or opaque. Of course, simplicity comes at a cost – these metrics cannot capture nuances such as situations in which a disparity can be attributed to "acceptable" distributional differences of protected groups across levels of

a rating factor. Methods that can account for "fair" distributional differences are discussed later in this paper.

Premium parity corresponds to a notion of fairness based on the idea of equal impacts on different groups: that average premiums do not vary by protected group. In contrast, loss ratio parity corresponds more closely to traditional actuarial notions of fair rates – that rates should differentiate between insureds with differences in expected costs. Especially in situations where loadings for expenses and target underwriting profit are constant multiplicative or additive factors, disparities in loss ratio should correspond to disparities in underwriting profit, so loss ratio parity corresponds to a notion of fairness in which all protected groups are equally profitable on average. In other words, loss ratio parity means that no protected group gets overcharged, relative to any other protected group.

An additional advantage of premium parity is that it is simpler to calculate than loss ratio parity, as it does not require claims data, and generally a lower data volume is needed for credibility given that premiums are less volatile than claims. A key advantage of loss ratio parity is that it requires a consistently accurate assessment of risk across protected groups. However, in cases where individuals may be exposed to risks outside their control in part due to their protected group membership, a purely risk-proportional approach could be viewed as unfair on the grounds that achieving loss ratio parity could lead to charging people for risks outside their control.

With premium parity and loss ratio parity each corresponding to a different notion of fairness, a natural question that arises is "Why not just use a model that satisfies both?" The next section demonstrates that when average losses differ between protected groups, these two parity conditions cannot be achieved simultaneously.

Cannot Simultaneously Achieve Premium Parity and Loss Ratio Parity

One of the challenges with assessing model fairness is that there are multiple parity metrics available, and in many cases, it is mathematically impossible to achieve all forms of parity at once. For classification models, various authors[11] have shown that, provided the response variable is not statistically independent of the protected attribute, it's not possible to construct a model that simultaneously achieves demographic parity, false positive rate parity, and false negative rate parity. A short argument demonstrates that the analogous statement is true for loss cost models:

> *If groups* a *and* b *have different lost costs, then a pricing model cannot simultaneously achieve premium parity and loss ratio parity between these two groups.*

This statement can be verified by proving the logically equivalent statement that a pricing model that has achieved *both* premium parity and loss ratio parity can occur only when groups *a* and *b* have identical loss costs.

[11] A more general proof appears in Barocas et al. (2023).

Recall that for group k the losses, premium, and exposure are denoted by L_k, P_k, and X_k, respectively. Achieving loss ratio parity means that both protected groups have the same loss ratio, or

$$\frac{L_a}{P_a} = \frac{L_b}{P_b}.$$

Achieving premium parity means that both protected groups have the same average premium, or

$$\frac{P_a}{X_a} = \frac{P_b}{X_b}.$$

The premium parity condition can be re-expressed as

$$P_b = \frac{P_a X_b}{X_a}.$$

Combining this with the loss ratio parity condition gives

$$\frac{L_a}{P_a} = \frac{L_b X_a}{P_a X_b}.$$

Canceling out the P_a and rearranging gives

$$\frac{L_a}{X_a} = \frac{L_b}{X_b}.$$

which shows that the two groups must have the same loss cost for both premium parity and loss ratio parity to hold. The same argument would apply if P_a and P_b are outputs of a theoretical loss cost model rather than the actual premium, with the interpretation being that parity of average model predictions cannot be achieved simultaneously with parity of the ratio of aggregate losses to aggregate model predictions.

This means that when protected groups have differences in average losses, if a pricing algorithm achieves premium parity, then a loss ratio disparity will exist. This can lead to cross-subsidization between protected groups, and potentially to adverse selection if premium parity is not enforced consistently across insurers. Premium parity could be viewed as unfair; it could cause affordability concerns for the groups that are over-charged relative to risk and raise concerns about availability of insurance for the groups that are undercharged relative to risk. Loss ratio parity is an important metric in situations where insurance availability is a concern; absent a robustly enforced take-all-comers rule, a loss ratio disparity could discourage insurers from writing business to some protected groups.

A consequence of this fact is that it's not possible to address fairness through an exclusively technical analysis – since it's not possible to achieve both premium parity and loss ratio parity simultaneously, exogenous information is needed to identify which parity metric more closely corresponds to a socially accepted notion of fairness for a given use case. This need is amplified by the observation that there are reasons that each metric could be seen as fair, and reasons that it could be seen as unfair. Depending on the circumstances, the relevant notion of fairness could come from a variety of sources, such as regulatory definitions of fairness or a company's business strategy. Alternatively, it could be the case that neither premium parity nor loss ratio parity is considered an appropriate notion of fairness for a given context, and more nuanced metrics, such as those described in the second part of this paper, may be preferred.

Part 2 of the paper can be found on casact.org/raceandinsuranceresearch.

- Part 2, Section 4 delves into more complex fairness analyses that take into consideration multiple rating factors and distributional differences between protected classes across the levels of certain rating factors, conditional demographic parity, the proxy ("control variable") test, and nonparametric matching.
- Part 2, Section 5 reviews several technical bias mitigation methods that can be applied to insurance pricing data, models, or model outputs.
- Part 2, Section 6 discusses several important non-modeling considerations that can contribute to fairness concerns, such as targeted marketing practices, regulatory restrictions, and discounts.

Appendix A. Simulation for Utilizing Imputed Probabilities

The following R code was used to create a simulated dataset with 1,000 records, each classified as either class A or class B.

Assumptions:

- The number of records is $N = 1{,}000$.
- Each has random likelihood of being class A based on a uniform draw on probabilities between 0 and 1.
- A random draw is made for each record to determine class, based on the probabilities determined above.
- For class A, the probability that outcome = 1 is 0.6. For class B it is 0.4.
- A random draw is made based on the above-defined probabilities to determine outcome.

R Code:

```
set.seed(12345)
N <- 1000
prob.class.A <- round(runif(N),3)
prob.class.B <- 1 - prob.class.A
var.class.A <- rbinom(N, 1, prob.class.A)
var.class <- ifelse(var.class.A == 1, "A", "B")
prob.outcome <- ifelse(var.class.A == 1, 0.6, 0.4)
var.outcome <- rbinom(N, 1, prob.outcome)
var.class.impute <- ifelse(prob.class.A < prob.class.B, "B", "A")
df <- data.frame(var.class, prob.class.A, prob.class.B, var.outcome,
var.class.impute)
for (i in 1:100) {
  prob.name <- paste("prob.class.impute", i, sep="")
  prob.name <- runif(N)
  var.name <- paste("var.class.impute", i, sep="")
  df[,var.name] <- ifelse(prob.name < prob.class.A, "A", "B")
}
write.csv(df, "tmp_outfile.csv")
```

Appendix B. Illustration of Premium Parity and Loss Ratio Parity in a Proxy Variable Situation

The hypothetical example in Section 3.1 illustrated the concepts of premium parity and loss ratio parity in a situation where the source of a loss ratio disparity was an uncaptured interaction between protected class and a rating factor. This appendix revisits that example, using instead sample data that correspond to a correlated (or "proxy") variable as a potential source of bias. Assume that the loss experience in Table B.1 exists for groups *a* and *b* within a protected class, segmented by a binary rating factor.

In this example, because the effect of the rating factor does not vary by protected group, the "full information" relativities are the same as the relativities derived if data on protected class membership were unavailable, shown in Table B.2.

Assume a base rate of $1,538 is used, with a relativity of 1.5 for level *Y* of the rating factor. This would result in the aggregate premiums shown in Table B.3.

Next, summarize the data by level of the protected attribute and calculate the average premium, as shown in Table B.4.

In contrast with the example in Section 3.1, in this example we see that premium parity is not achieved, but loss ratio parity is. Notice that the premium disparity results from the fact that the protected class has a different distribution for each level of the rating factor, which was not the case in the example in Section 3.1.

Table B.1. Hypothetical Example Where a Rating Factor Is a Proxy for Protected Class

Full Information (Protected Group Known)					
Level of Rating Factor (R)	Protected Group (k)	Exposure ($X_{R,k}$)	Losses ($L_{R,k}$)	Loss Cost ($\bar{L}_{R,k} = L_{R,k}/X_{R,k}$)	Loss Cost Relativity ($\bar{L}_{R,k}/\bar{L}_{N,a}$)
N	*a*	70	$70,000	$1,000	1.00
N	*b*	10	$10,000	$1,000	1.00
Y	*a*	10	$15,000	$1,500	1.50
Y	*b*	10	$15,000	$1,500	1.50

Table B.2. Calculating Relativities without Knowing Protected Group

Protected Group Unknown				
Level of Rating Factor (R)	Exposure (X_R)	Losses (L_R)	Loss Cost ($\bar{L}_R = L_R/X_R$)	Loss Cost Relativity ($\bar{L}_R/\bar{L}_N$)
N	80	$80,000	$1,000	1.00
Y	20	$30,000	$1,500	1.50

Table B.3. Rating Structure Based on One-Way Relativities

Level of Rating Factor (R)	Protected Group (k)	Exposure ($X_{R,k}$)	Individual Customer Premium	Total Premium ($P_{R,k}$)
N	*a*	70	$1,538	$107,660
N	*b*	10	$1,538	$15,380
Y	*a*	10	$2,307	$23,070
Y	*b*	10	$2,307	$23,070

Table B.4. Summarizing Proposed Premium by Protected Group (One-Way Relativities)

Protected Class Group (k)	Exposure (X_k)	Premium (P_k)	Average Premium (P_k/X_k)	Losses (L_k)	Loss Ratio (L_k/P_k)
a	80	$130,730	$1,634	$85,000	65.0%
b	20	$38,450	$1,923	$25,000	65.0%

References

Alon-Barkat, S., and M. Busuioc. 2021. "Decision-Makers Processing of AI Algorithmic Advice: Automation Bias versus Selective Adherence." Preprint, arXiv, https://doi.org/10.48550/arXiv.2103.02381.

American Academy of Actuaries. 2023a. *An Actuarial View of Data Bias: Definitions, Impacts and Considerations.* Issue brief. Washington, DC. https://www.actuary.org/sites/default/files/2023-07/risk_brief_data_bias.pdf.

American Academy of Actuaries. 2023b. *Approaches to Identify and/or Mitigate Bias in Property and Casualty Insurance.* Issue brief. Washington, DC. https://www.actuary.org/sites/default/files/2023-02/CPCdataBiasIB.2.23_0.pdf.

Barocas, S., M. Hardt, and A. Narayanan. 2023. "Classification." Ch. 3 in *Fairness and Machine Learning: Limitations and Opportunities.* MIT Press. https://fairmlbook.org/.

Catalogue of Bias Collaboration (A. Erasmus, B. Holman, and J. P. A. Ioannidis). 2020. "Data-Dredging Bias." In Catalogue of Bias. https://catalogofbias.org/biases/data-dredging-bias/.

Cavanaugh, L., S. Merkord, T. Davis, and D. Heppen. 2024. *Regulatory Perspectives on Algorithmic Bias and Unfair Discrimination.* CAS Research Paper Series on Race and Insurance Pricing. Arlington, VA: CAS. https://www.casact.org/sites/default/files/2024-08/Regulatory_Perspectives_on_Algorithmic_Bias_and_Unfair_Discrimination.pdf.

Chibanda, K. F. 2022. *Defining Discrimination in Insurance.* CAS Research Paper Series on Race and Insurance Pricing. Arlington, VA: CAS. https://www.casact.org/sites/default/files/2022-03/Research-Paper_Defining_Discrimination_In_Insurance.pdf

Datta, A. 2021. "3 Kinds of Bias in AI Models – and How We Can Address Them." InfoWorld, February 24. https://www.infoworld.com/article/2262600/3-kinds-of-bias-in-ai-models-and-how-we-can-address-them.html.

Dolman, C., and D. Semenovich. 2018. "Algorithmic Fairness: Contemporary Ideas in the Insurance Context." Paper presented at the Institute and Faculty of Actuaries GIRO Conference 2018, Birmingham, England, October 24. https://www.actuaries.org.uk/system/files/field/document/B9_Chris%20Dolman%20%28paper%29.pdf and https://www.actuaries.org.uk/system/files/field/document/B9_Chris%20Dolman.pdf.

Elliott, M. N., P. A. Morrison, A. Fremont, D. F. McCaffrey, P. Pantoja, and N. Lurie. 2009. "Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities." *Health Services and Outcomes Research Methodology* 9 (2): 69–83.

Federal Insurance Office. 2017. *Study on the Affordability of Personal Automobile Insurance.* https://home.treasury.gov/system/files/311/FINAL%20Auto%20Affordability%20Study_web.pdf.

Frees, E. W., and F. Huang. 2023. "The Discriminating (Pricing) Actuary." *North American Actuarial Journal* 27 (1): 2–24. https://doi.org/10.1080/10920277.2021.1951296.

Goldburd, M., A. Khare, D. Tevet, and D. Guller. 2020. Generalized Linear Models for Insurance Rating, 2nd ed. CAS Monograph Series, no. 5. Arlington, VA: CAS. https://www.casact.org/sites/default/files/2021-01/05-Goldburd-Khare-Tevet.pdf.

Jager, K. J., G. Tripepi, N. C. Chesnaye, F. W. Dekker, C. Zoccali, and V. S. Stel. 2020. "Where to Look for the Most Frequent Biases?" *Nephrology (Carlton)* 25 (6): 435–41. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7318122/.

Korteling, J. E., and A. Toet. 2022. "Cognitive Biases." In *Encyclopedia of Behavioral Neuroscience,* 2nd ed., pp 610–19. Accessed March 6, 2024. https://www.sciencedirect.com/science/article/abs/pii/B9780128093245241059.

Leong, J., R. Moncher, and K. Jordan. 2024. *A Practical Guide to Navigating Fairness in Insurance Pricing.* CAS Research Paper Series on Race and Insurance Pricing. Arlington, VA: CAS. https://www.casact.org/sites/default/files/2024-08/A_Practical_Guide_to_Navigating_Fairness_in_Insurance_Pricing.pdf.

Lindholm, M., R. Richman, A. Tsanakas, and M. V. Wüthrich. 2022. "Discrimination-Free Insurance Pricing." *ASTIN Bulletin* 52 (1): 55–89. doi:10.1017/asb.2021.23.

Mosley, R., and R. Wenman. 2022. *Methods for Quantifying Discriminatory Effects on Protected Classes in Insurance.* CAS Research Paper Series on Race and Insurance Pricing. Arlington, VA: CAS. https://www.casact.org/sites/default/files/2022-03/Research-Paper_Methods-for-Quantifying-Discriminatory-Effects.pdf.

NIST (National Institute of Standards and Technology). 2022. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.* Special Publication 1270. *https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf.*

Norwood, C. 2021. "How Infrastructure Has Historically Promoted Inequality." PBS News, April 23. Accessed February 23, 2021. https://www.pbs.org/newshour/politics/how-infrastructure-has-historically-promoted-inequality.

Rosenman, E., S. Olivella, and K. Imai. 2022. "Race and Ethnicity Data for First, Middle, and Last Names." Harvard Dataverse, V9. https://doi.org/10.7910/DVN/SGKW0K.

Schraub, D., J. Lang, Z. Zhang, and M. A. Sayre. 2024. *Comparison of Regulatory Framework for Non-Discriminatory AI Usage in Insurance.* Society of Actuaries Research Institute. https://www.casact.org/sites/default/files/2024-08/Comparison-of-Regulatory-Framework-for-Non-Discriminatory-AI-usage-in-Insurance.pdf.

Serwin, K., and A. H. Perkins Jr. n.d. "Algorithmic Bias: A New Legal Frontier." https://www.iadclaw.org/assets/1/7/18.1_-_REVIEWED-_Serwin-_Algorithmic_Bias.pdf.

Thompson, C. 2021. "Who's Homeless Enough for Housing? In San Francisco, an Algorithm Decides." Coda, September 21. https://www.codastory.com/authoritarian-tech/san-francisco-homeless-algorithm/.

2021 Casualty Actuarial Society Race and Insurance Research Task Force. 2022. *Understanding Potential Influences of Racial Bias on P&C Insurance: Four Rating Factors Explored.* Casualty Actuarial Society Research Paper Series on Race and Insurance Pricing. Arlington, VA: CAS. https://www.casact.org/sites/default/files/2022-03/Research-Paper_Understanding_Potential_Influences.pdf?utm_source=Landing+Page&utm_medium=Website&utm_campaign=RIP+Series.

Tzioumis, K. 2018. "Data for: Demographic Aspects of First Names." Harvard Dataverse, V1. https://doi.org/10.7910/DVN/TYJKEZ.

Verma, S., Ernst, M., and R. Just. 2021. "Removing Biased Data to Improve Fairness and Accuracy." https://doi.org/10.48550/arXiv.2102.03054.

Voicu, I. 2018. "Using First Name Information to Improve Race and Ethnicity Classification." *Statistics and Public Policy* 5 (1): 1–13. doi:10.1080/2330443X.2018.1427012.