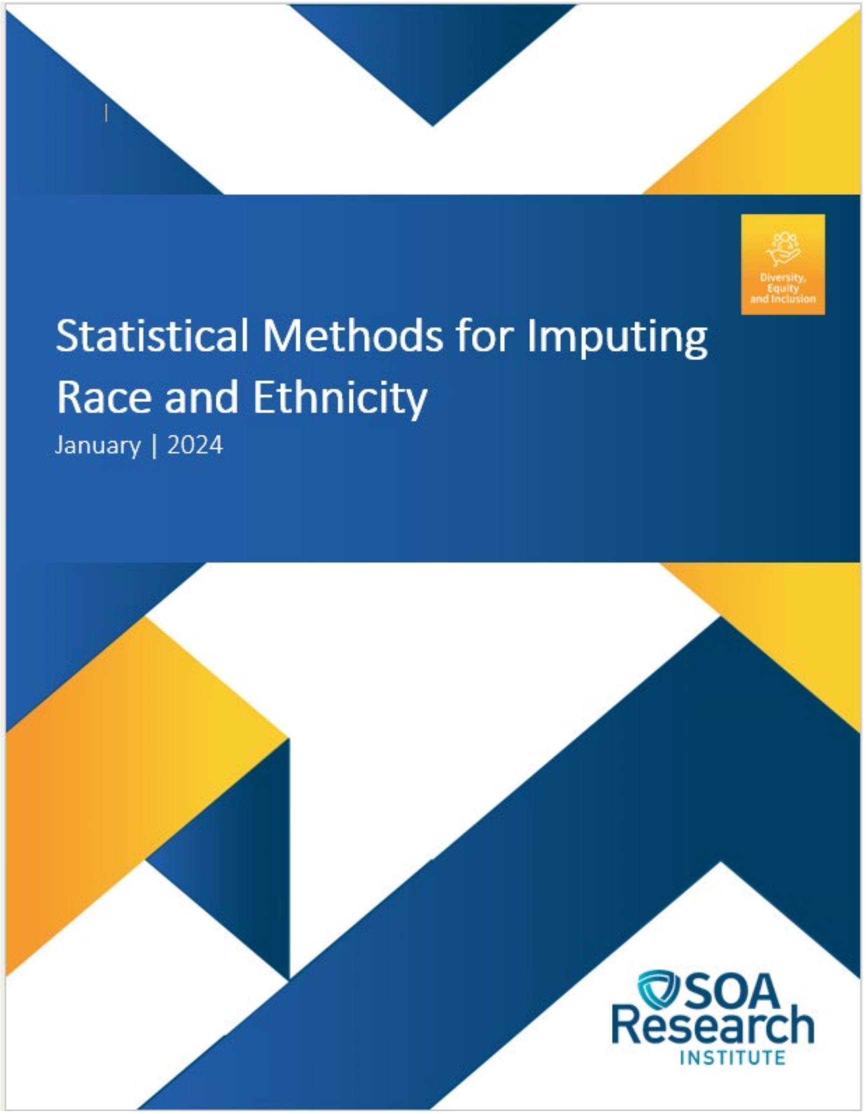
NAIC: Big Data and Artificial Intelligence (H) Working Group Meeting

Monday 7/29/2024 12:00 PM - 1:00 PM



A Summary of Research Sponsored by the Society of Actuaries on Statistical Methods for Imputing Race and Ethnicity

Presented by Dorothy L. Andrews, NAIC

AUTHORS
Larry Baeder
Erica Baird, PhD, FSA, MAAA
Peggy Brinkmann, FCAS, MAAA
Joe Long, ASA, MAAA
Caleb Stracke, ASA, MAAA
Kweweli Togba-Doya
Gabriele Usan
Natalie Weaver
Meseret Woldeyes, MS

SPONSOR Diversity, Equity, and Inclusion Research Advisory Council

NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS

Acknowledgments

The researchers' deepest gratitude goes to those without whose efforts this project could not have come to fruition: the Project Oversight Group and others for their diligent work overseeing, reviewing, and editing this report for accuracy and relevance.

Project Oversight Group members:

Dorothy Andrews, Ph.D., ASA, MAAA, CSPA
Brian Bayerle, FSA, MAAA
Stephen Cameron, FSA, MAAA
Amine Elmeghni, FSA, MAAA, MSc
Jean-Marc Fix, FSA, MAAA
Hannah Kraus, FSA, MAAA
Tim Luedtke, FSA, MAAA
Ian McCulla, FSA, MAAA
Andrew Melnyk, Credentials
Min Mercer, FSA
Murali Niverthi, FSA, MAAA
Renee West, FSA, MAAA

At the Society of Actuaries Research Institute:
Lisa Schilling, FSA, EA, FCA, MAAA, Senior Research Actuary

Society of Actuaries
Project Oversight Group (POG)

Agenda

1. Definitions
2. Statistical Approaches
3. Pre-Bayesian Methods
4. What is a Probability?
5. Bayes Theorem
6. Bayesian Methods
7. Required Data for BIFSG
8. Simple Example
9. Accuracy Concerns

COMMENTARY | HEALTH EQUITY

HEALTH AFFAIRS > VOL. 41, NO. 8: SPENDING, PAYMENT & MORE

COMMENTARY

Predicting Race And Ethnicity To Ensure Equitable Algorithms For Health Care Decision Making

Irineo Cabreros, Denis Agniel, Steven C. Martino, Cheryl L. Damberg, and Marc N. Elliott

AFFILIATIONS

PUBLISHED: AUGUST 2022 No Access https://doi.org/10.1377/hlthaff.2022.00095

NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS

Definitions

- Probabilistic Inference
- Statistical Inference
- Imputation
  - Indirect Estimation
  - Direct Estimation

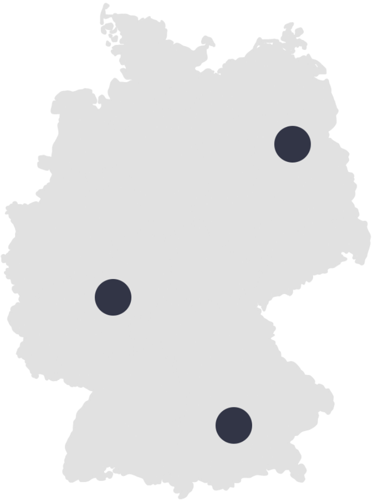
- Performance Metrics
  - Accuracy
  - Error Rate
  - False Positives
  - False Negatives
  - Precision
  - Sensitivity/Recall
  - Specificity/Selectivity
  - Receiver Operator Curve (ROC)
  - Area Under ROC Curve (AUC)
  - Concordance – C-Statistic

Statistical Approaches

Bayesian Statistics	Frequentist Statistics
Probability represents the degree of belief in a hypothesis; inferences are based on both data and prior beliefs.	Probability is used to describe the likelihood of an event occurring; inferences are made based on data alone.
Uses past hypotheses	No use of past hypotheses
No null hypothesis	Has null hypotheses
Experiment relies on past data and observations	Experiment relies frequency of repeated, random events
Subjectivity is permitted in testing and analysis	Subjectivity is NOT permitted in testing and analysis

Pre-Bayesian Methods

- Geocoding Only (GO)
- Surname Analysis (SA)
- Categorical Surname and Geocoding (CSG)



GEOCODING

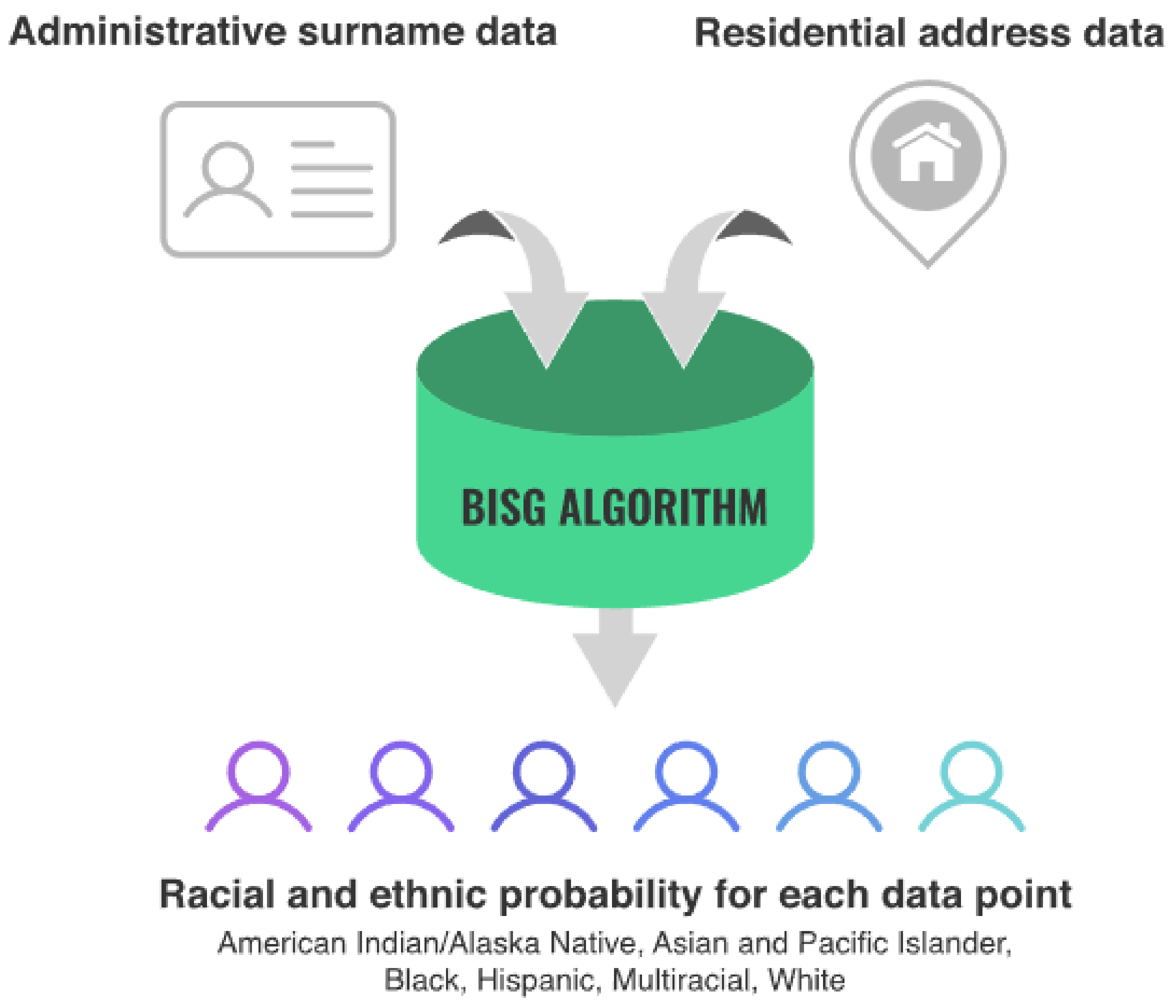
LATITUDE	LONGITUDE
48.1°N	11.6°E
50.1°N	13.4°E
52.5°N	8.7°E

Bayesian Methods

- Bayesian Surname Coding (BSG)
- Bayesian Improved Surname Geocoding (BISG)
- Medicare Bayesian Improved Surname Geocoding (MBISG)
- Bayesian Improved Surname Geocoding Extensions (BISGE)
- Bayesian Improved First Name Surname Geocoding (BISFG)
- Modified Bayesian Improved First Name Surname Geocoding (MBIFSG)
- Fully Bayesian Improved Surname Geocoding (fBISG)
  - With Zero-Count Correction
  - With Additional Surname
  - With First Name
  - With First and Middle Name
- Bayesian Instrumental Regression for Disparity Estimation (BIRDiE)

RAND CORPORATION OBJECTIVE ANALYSIS. EFFECTIVE SOLUTIONS.

B - Bayesian
I - Improved
S - Surname
G - Geocoding

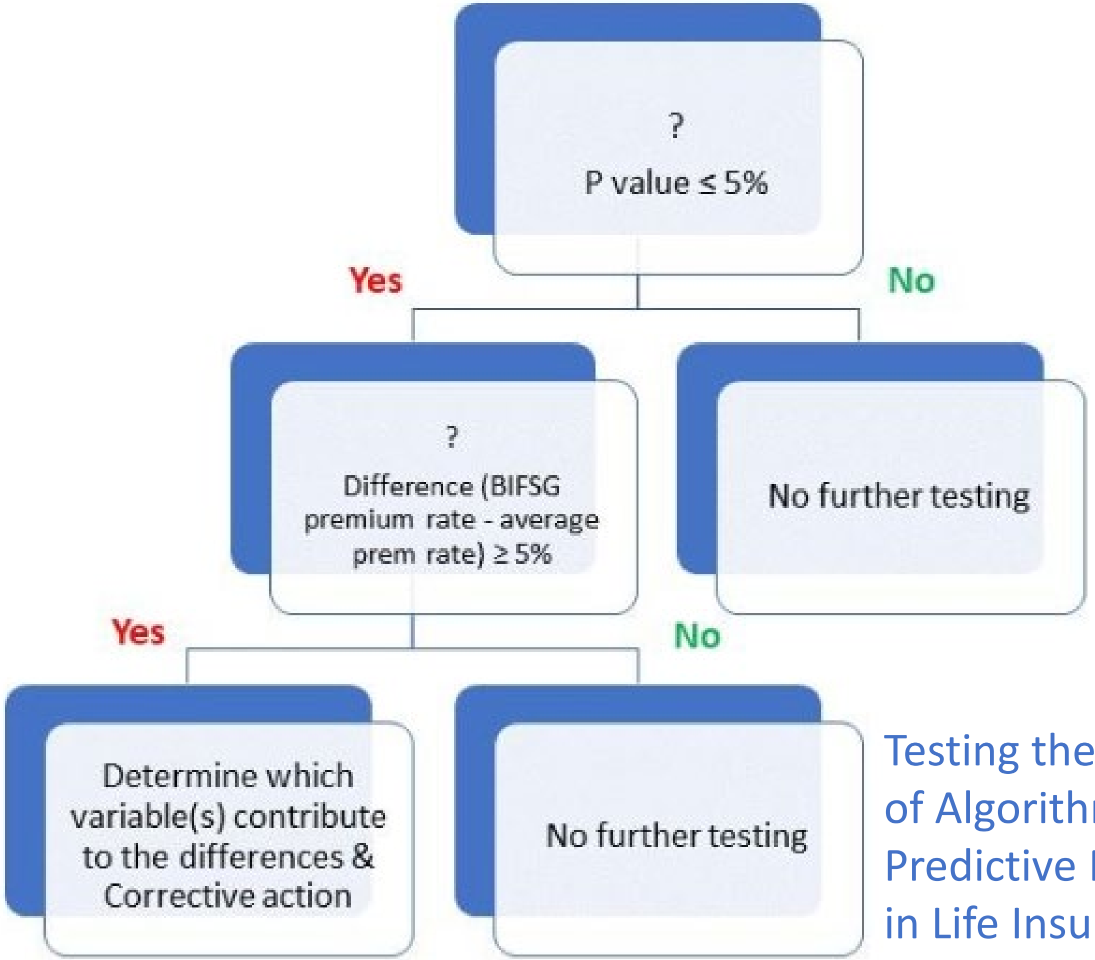


NAIC NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS

Consumer Financial Protection Bureau

B - Bayesian
I - Improved
F - First Name
S - Surname
G - GeoCoding

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



Testing the Fairness of Algorithms and Predictive Models in Life Insurance

What is a Probability of Event (x)?

$$Probability(x) = \frac{Number\ of\ times\ x\ Occurs}{All\ Possible\ Occurances}$$

$$= Proportion\ (x)$$

Bayes Theorem

(In Technicolor)

Likelihood

How probable is the evidence given that our hypothesis is true?

Prior

How probable was our hypotheses before observing the evidence?

$$P(H|e) = \frac{P(e|H) x\, P(H)}{P(e)}$$

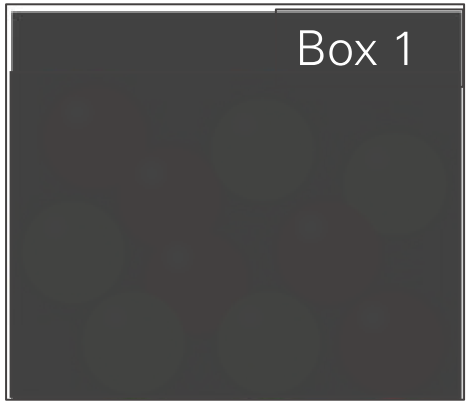
Posterior

How probable is our hypothesis given the observed evidence? (Not directly computable)

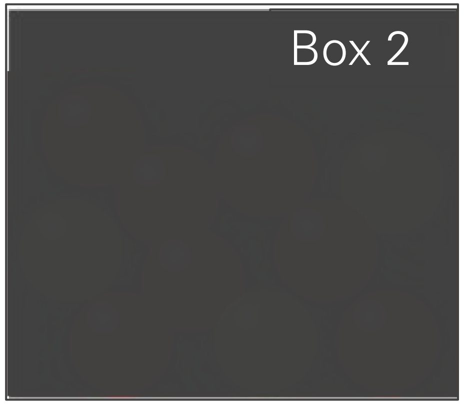
Marginal

How probable is the new evidence under all possible hypotheses?

Example of Bayes Theorem



10 Balls



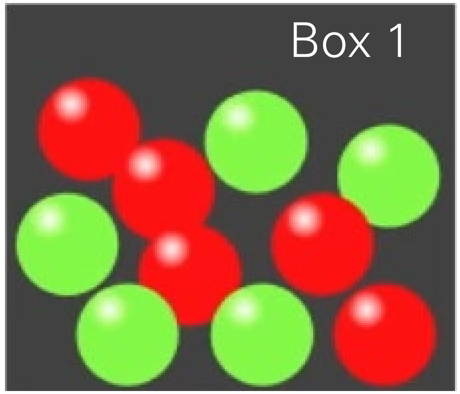
10 Balls

Scenario:
You are presented with a draw of a ball, and you are curious to know which box it came from knowing that each box is equally likely to have been selected.

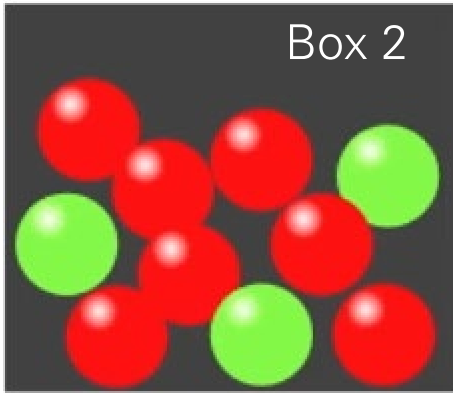
Question:
What is to probability the ball came from Box 1?

$$Prob(Box\ 1) = \frac{1}{2} = 0.5$$

Example of Bayes Theorem



5 Green 5 Red



3 Green 7 Red

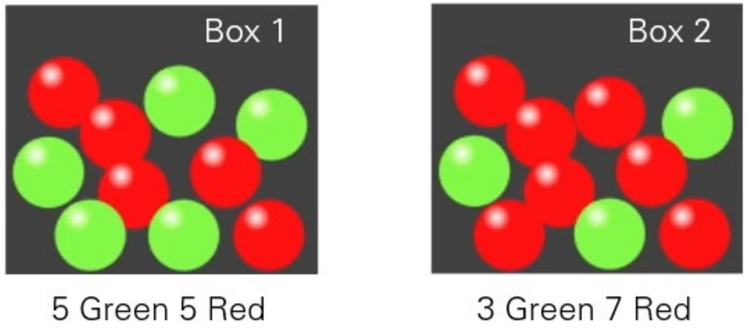
Scenario:
You are presented with a draw of a ball, and you are curious to know which box it came from knowing that each box is equally likely to have been selected.

New Information or Evidence:
The ball is Green.

Question:
Now what is the probability the ball came from Box 1?

$$Prob(Box\ 1|\ Green\ Ball)$$

Example of Bayes Theorem



Hypothesis 1: Ball came from Box 1

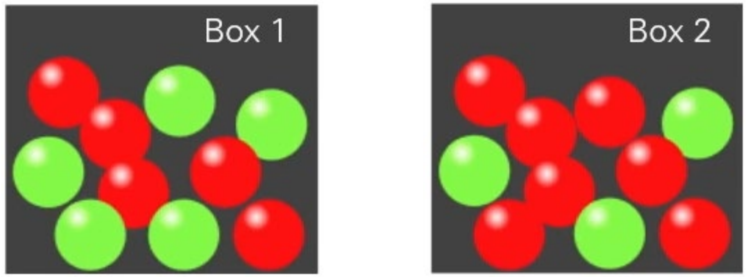
$$\underbrace{P(Box\ 1|Green\ Ball)}_{\{Posterior\}} = \frac{\overbrace{P(Green\ Ball\ |Box\ 1)}^{\{Likelihood\}} x\ \overbrace{P(Box\ 1)}^{\{Prior\}}}{\underbrace{P(Green\ Ball)}_{\{Marginal\}}}$$

Marginal Reflects Probability Over All Hypotheses:

- H1: Ball came from Box 1
- H2: Ball came from Box 2

NAIC NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS

Example of Bayes Theorem



5 Green 5 Red   3 Green 7 Red

Let's Calculate the Marginal First:

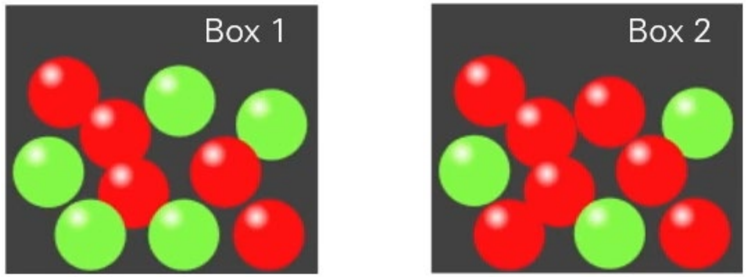
$$P(Green\ Ball) = P(Green\ Ball\ |Box1)\ x\ P(Box\ 1) + P(Green\ Ball\ |Box\ 2)\ x\ P(Box\ 2)$$

$$= (5/10) \times (1/2) + (3/10) \times (1/2)$$

$$= 5/20 + 3/20$$

$$= 8/20$$

Example of Bayes Theorem



5 Green 5 Red   3 Green 7 Red

Now we can calculate our probability of interest.

$$P(Box\ 1|Green\ Ball) = \frac{P(Green\ Ball\ |Box\ 1)x\ P(Box\ 1)}{P(Green\ Ball)} = \frac{5/20}{8/20}$$

$$= 5/8$$

$$= 0.625$$

$$P(Box\ 2|Green\ Ball) = 0.375$$

Consumer Financial Protection Bureau

B - Bayesian
I - Improved
F - First Name
S - Surname
G - GeoCoding

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

Hypotheses: (Race)

1. Hispanic
2. Asian/Pacific Islander
3. Black
4. Multiracial
5. White
6. American Indian/Alaska Native

Evidence: (aka Input)

- First Name
- Surname
- Geocoding

$$P(R_i|G,S,F) = \frac{P(R_i|S)P(G|R_i)P(F|Ri)}{\sum_{i=1}^{6} P(R_i|S)P(G|R_i)P(F|R_i)}$$

NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS

Example

First Name: Jose
Surname: Garcia
Geocoding: 63144

Surgeo

2010 Mortgage Data*

Probablities of First Name = Jose Given Race

White	Black	API	Native	Multiple	Hispanic
0.00258669	0.00123681	0.00337229	0.00753317	0.00252458	0.2001545

Census 2010 Data

Probablities of Race Given Surname = Garcia

White	Black	API	Native	Multiple	Hispanic
0.0538	0.0045	0.0141	0.0047	0.0026	0.9203

Census 2010 Data

Probablities of Zip Code = 63144 Given Race

White	Black	API	Native	Multiple	Hispanic
0.000039	0.000007	0.000037	0.000005	0.000025	0.000005

* Tzioumis, K. (2017), "Demographic Aspects of First Names," *Scientific Data*, forthcoming. The first name list is available at: https://dx.doi.org/10.7910/DVN/TYJKEZ

NAIC NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS

Example

First Name: Jose
Surname: Garcia
Geocoding: 63144

Surgeo

Marginal Probabilities

White	Black	API	Native	Multiple	Hispanic
5.36E-09	3.76E-11	1.76E-09	1.73E-10	1.64E-10	8.66E-07

BIFSG Probabilities

White	Black	API	Native	Multiple	Hispanic
0.006137	0.000043	0.002013	0.000198	0.000187	**0.991421**

NAIC Staff Favorites

What race does BIFSG infer for them?

Miguel Romero
66216

Scott Sobel
29016

Dorothy Andrews
28226

NAIC Staff Favorites

What race does BIFSG infer for them?

Hispanic (99.5%)

White (99.9%)

White (78.5%)

NAIC Staff Favorites

Let's Look at the Data!

There is no way Dorothy would have been classified as Black by BIFSG!

Probablities of First Name

First Name	White	Black	API	Native	Multiple	Hispanic
Dorothy	**0.8286**	0.1318	0.0167	0.0035	0.0023	0.0171
Scott	0.9831	0.0027	0.0087	0.0006	0.0006	0.0043
Miguel	0.0616	0.0057	0.0113	0.0011	0.0000	0.9202

Probablities of Surname

Surname	White	Black	API	Native	Multiple	Hispanic
Andrews	**0.7178**	0.2158	0.0078	0.0109	0.0220	0.0257
Sobel	0.9571	0.0059	0.0065	0.0029	0.0029	0.0247
Miguel	0.0865	0.0050	0.0130	0.0069	0.0037	0.8850

Probablities of Zip Code

Zip Code	White	Black	API	Native	Multiple	Hispanic
28226	**0.800990**	0.067108	0.041519	0.001964	0.015633	0.072785
29016	0.661801	0.276570	0.012521	0.002688	0.013438	0.032983
66216	0.822316	0.051765	0.041412	0.003287	0.019432	0.061789

NAIC NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS

Accuracy of BIFSG

Identified Concerns

- Suffers from Majoritarian Bias (MB):
  Overstating the probabilities that non-White individuals are White.

- MB => Smaller Disparity Differences Than Exist
- Blacks with High Income, High Education => White

- Violation of Conditional Independence
- Biased Weights Toward Subgroups
- High Accuracy for Self-Report (SR) White/Hispanic
- Low Accuracy for SR Black, Native, ANHPI, Other

- Disproportionately High Probably to Whites
- Disproportionately Low Probably to Non-Whites

- More Attribute Data Can Improve the Method

Statistical Bias in Racial and Ethnic Disparity Estimates Using BIFSG

41 Pages • Posted: 19 Mar 2024

Elena Derby
Government of the United States of America - Joint Committee on Taxation

Connor Dowd
Government of the United States of America - Joint Committee on Taxation

Jacob Mortenson
Joint Committee on Taxation, US Congress

Date Written: February 20, 2024

Abstract

Bayesian Improved First Name and Surname Geocoding (BIFSG) is a widely used method for inferring race and ethnicity in data when this information is not available. It is well known that the assumptions underlying BIFSG can fail, but the effects of these failures on estimation by race and ethnicity are not well understood. In this paper we combine U.S. administrative tax data with data containing race and ethnicity to assess statistical bias in estimates of differences between racial/ethnic groups. We find that BIFSG suffers from majoritarian bias, overstating the probabilities that non-White individuals are White. When using these probabilities to estimate disparities between groups, BIFSG estimates understate differences in various outcomes between White and non-White taxpayers, in some cases reversing the direction of the disparity.

Derby, Elena and Dowd, Connor and Mortenson, Jacob, Statistical Bias in Racial and Ethnic Disparity Estimates Using BIFSG (February 20, 2024). Available at SSRN: https://ssrn.com/abstract=4733299 or http://dx.doi.org/10.2139/ssrn.4733299

NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS

Imputation Packages

- Surgeo (Python)
- Ethnicolr (Python)
- Wru (R)
- BIRDie
- Rethnicity

Predictive Modeling Imputation Methods

- Regression
- Natural Language Processing
- Multinomial Regression
- Multinomial Regression with Elastic Net Penalty
- Random Forests
- K-Nearest Neighbors
- Gradient Boosted Decision Trees

Questions