Using Machine Learning Small Area Estimation to Generate Precise Estimates of Financial Inclusion

Jonathan Hersh, PhD (Chapman Argyros School of Business)

Lucia Martin Rivero (IDB)

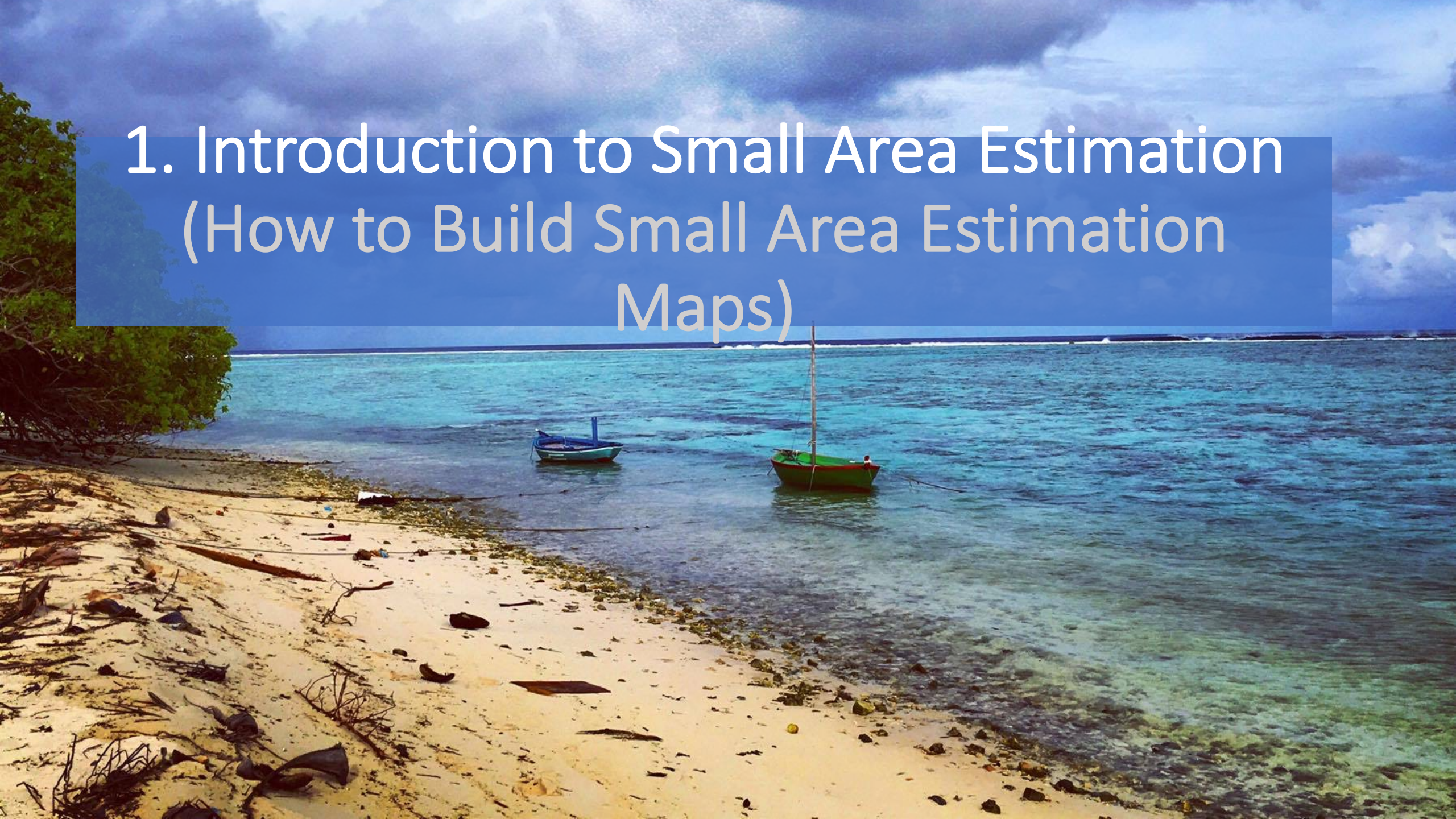
3/17/2021

Outline

1. Introduction to Small Area Estimation
2. Data
3. Machine Learning Model to Predict Household FI
4. Aggregation of Census Predictions to Generate Enumeration District FI measures
5. ED Characteristics Associated with FI
6. Policy Recommendations to Increase Financial Inclusion

This Paper

- April 2019 Labor Force Survey includes a Financial Inclusion (FI) module that presents FI at district averages
- We can use Machine Learning based Small Area Estimation to generate Enumeration District level picture of Financial Inclusion
- We focus on four metrics of FI that measure: banking access, barriers to banking access, online bank usage, and formal lending
- We use a random forest model to estimate household characteristics of FI and use these to generate ED level averages of FI
- ED Level Analysis Shows:
    - Significant sub-District level heterogeneity in FI
    - ED Poverty is the strongest predictor of FI
    - Ethnic Minorities and Foreign Born Much Less Likely to have FI

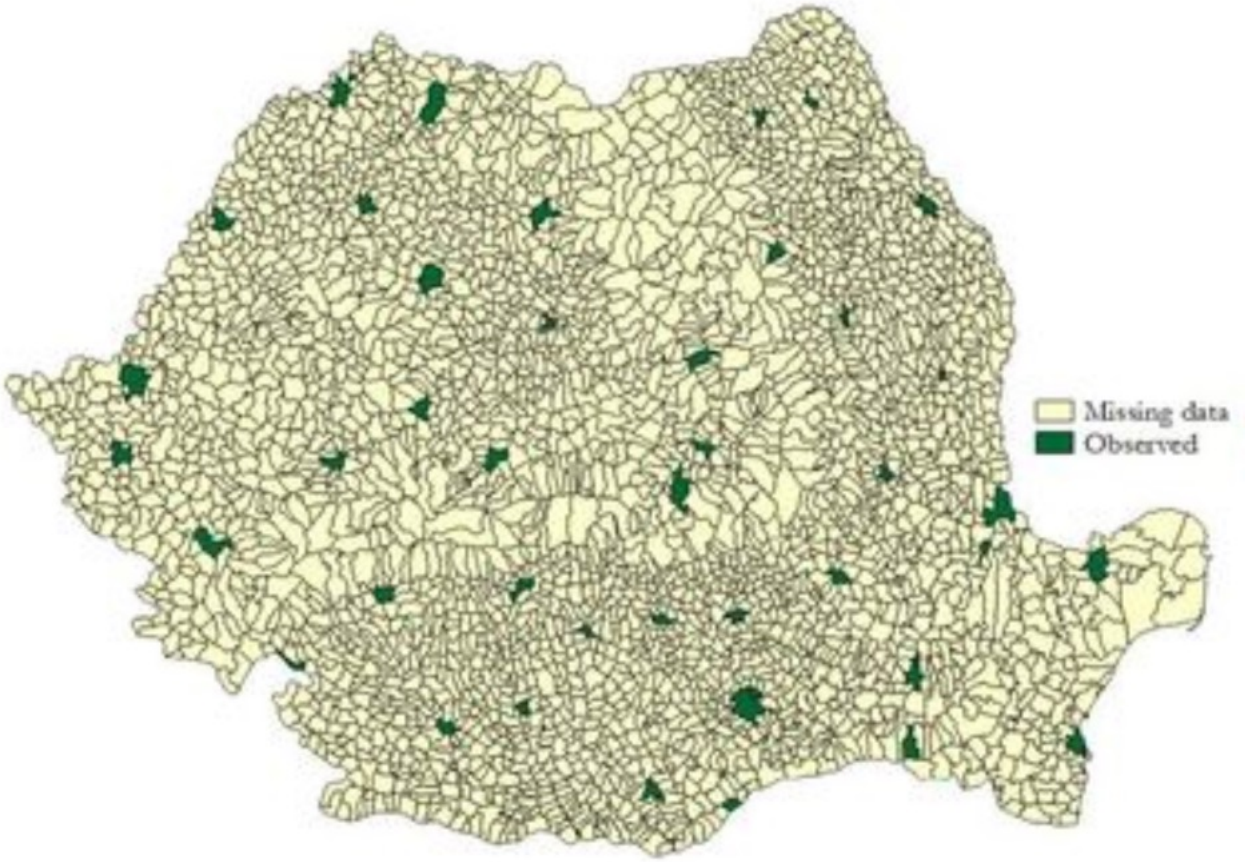


1. Introduction to Small Area Estimation (How to Build Small Area Estimation Maps)

Why is Small Area Estimation Useful?

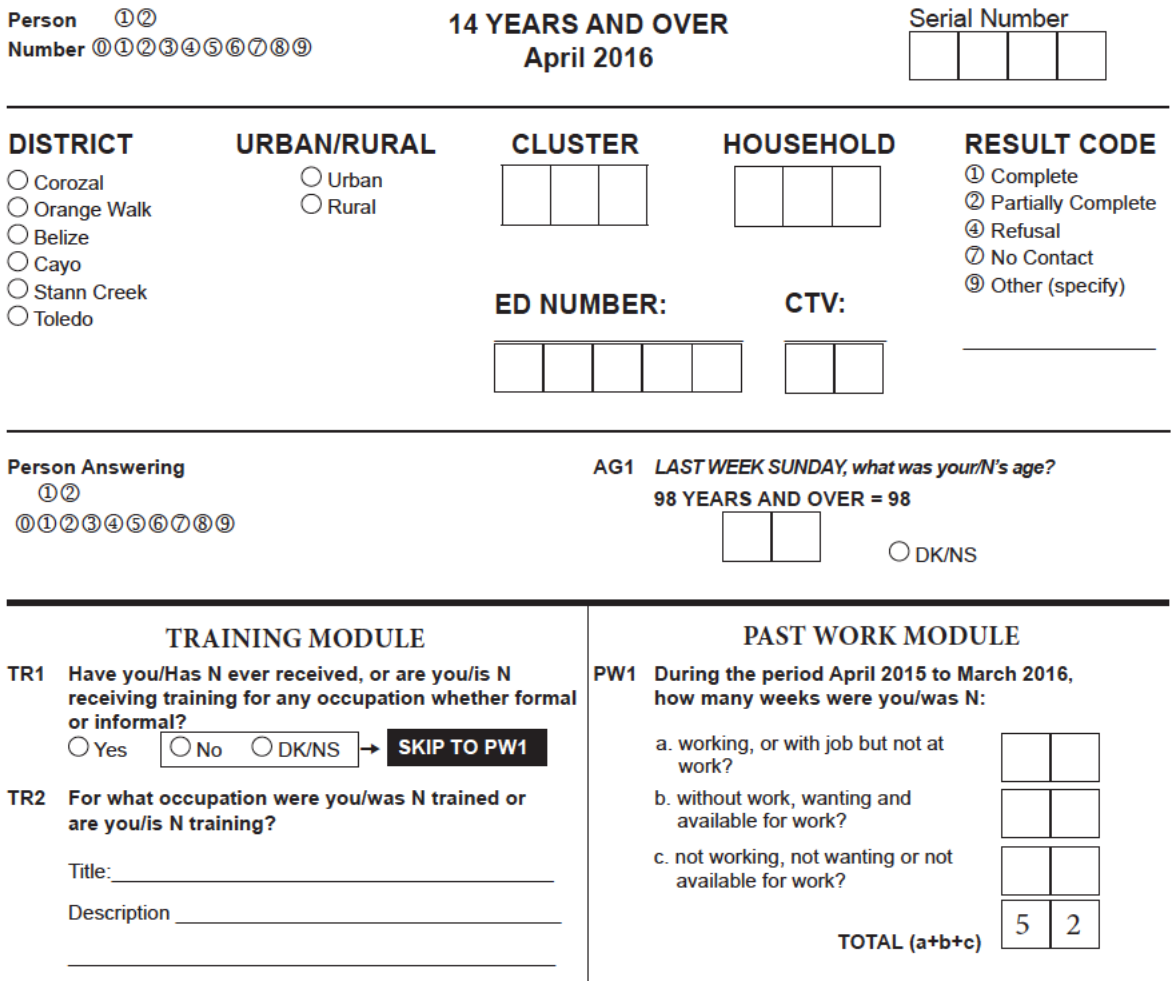
- Typical household surveys:
  - Have very low sample size
  - Usually do not cover every geographic area or subgroup of interest

- Often we want to learn survey information for important sub-groups such as for specific areas



Why is Small Area Estimation Useful?

- Typical household surveys:
  - Have very low sample size
  - Usually do not cover every geographic area
- Often we want to learn survey information for important sub-groups such as for specific areas



Person Number ①②
⓪①②③④⑤⑥⑦⑧⑨

14 YEARS AND OVER
April 2016

Serial Number

DISTRICT
○ Corozal
○ Orange Walk
○ Belize
○ Cayo
○ Stann Creek
○ Toledo

URBAN/RURAL
○ Urban
○ Rural

CLUSTER

HOUSEHOLD

RESULT CODE
① Complete
② Partially Complete
④ Refusal
⑦ No Contact
⑨ Other (specify)

ED NUMBER:

CTV:

Person Answering
①②
⓪①②③④⑤⑥⑦⑧⑨

AG1 *LAST WEEK SUNDAY, what was your/N's age?*
98 YEARS AND OVER = 98
○ DK/NS

TRAINING MODULE

TR1 **Have you/Has N ever received, or are you/is N receiving training for any occupation whether formal or informal?**
○ Yes ○ No ○ DK/NS → **SKIP TO PW1**

TR2 **For what occupation were you/was N trained or are you/is N training?**
Title:
Description

PAST WORK MODULE

PW1 **During the period April 2015 to March 2016, how many weeks were you/was N:**
a. working, or with job but not at work?
b. without work, wanting and available for work?
c. not working, not wanting or not available for work?
TOTAL (a+b+c) 5 2

Small Area Estimation

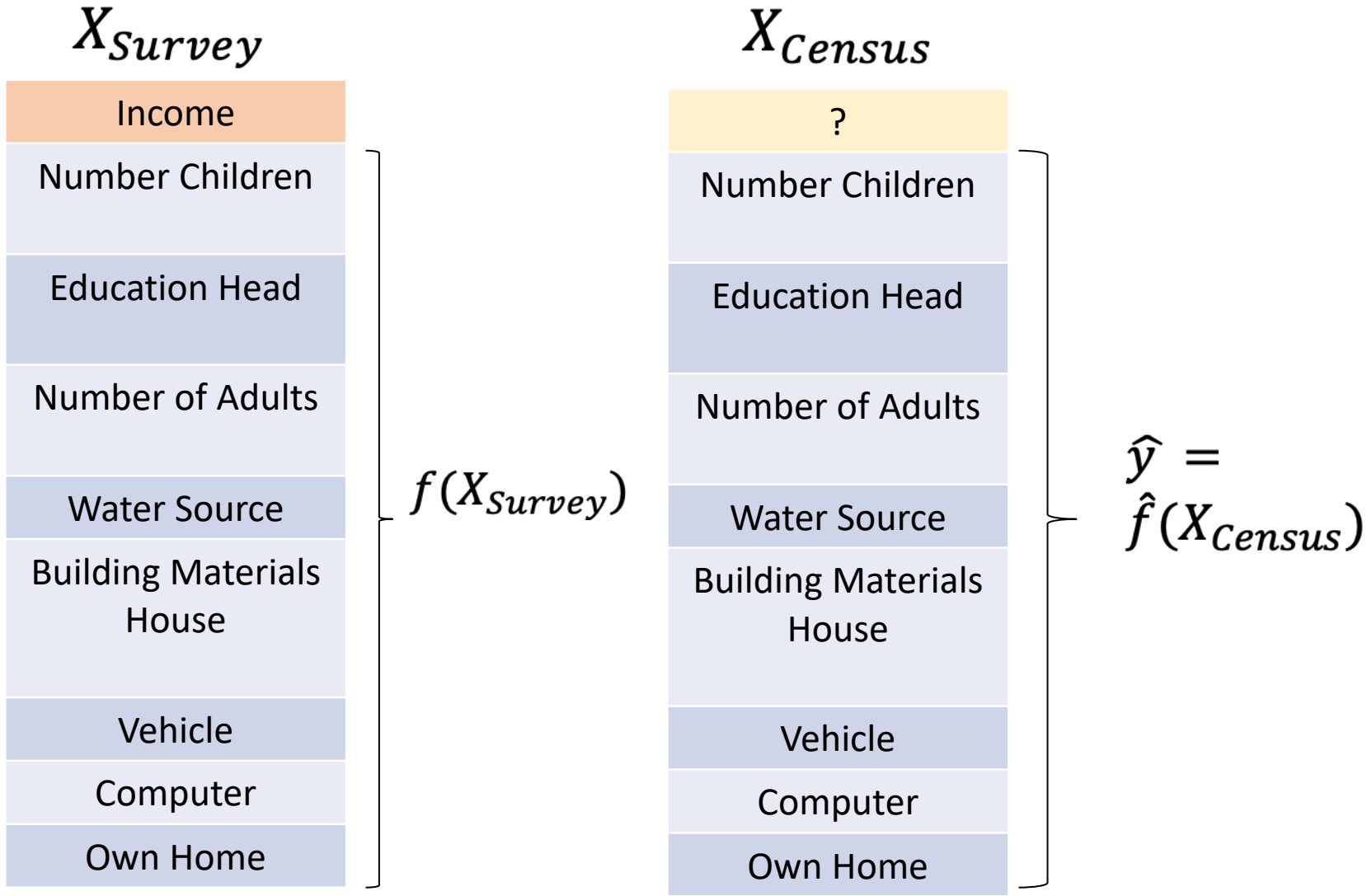
- One solution recognizes that there are often many common variables shared in census and surveys

X_{Survey}
Income
Number Children
Education Head
Number of Adults
Water Source
Building Materials House
Vehicle
Computer
Own Home

X_{Census}
?
Number Children
Education Head
Number of Adults
Water Source
Building Materials House
Vehicle
Computer
Own Home

Small Area Estimation

- One solution recognizes that there are common variables in census and surveys
- We can fit a household model relating survey characteristics and the outcome of interest (FI measures) and use this relationship to predict FI in the Census, allowing us to aggregate to the subgroup of interest.



Elbers, Lanjouw, Lanjouw (2003)

- Seminal paper is Elbers Lanjouw, Lanjouw (2003) -> simulation method for errors.

- In my own work, using machine learning directly can lead to better outcomes if number of variables is very large.

- One may still apply simulated standard errors to prediction following machine learning.

Econometrica, Vol. 71, No. 1 (January, 2003), 355–364

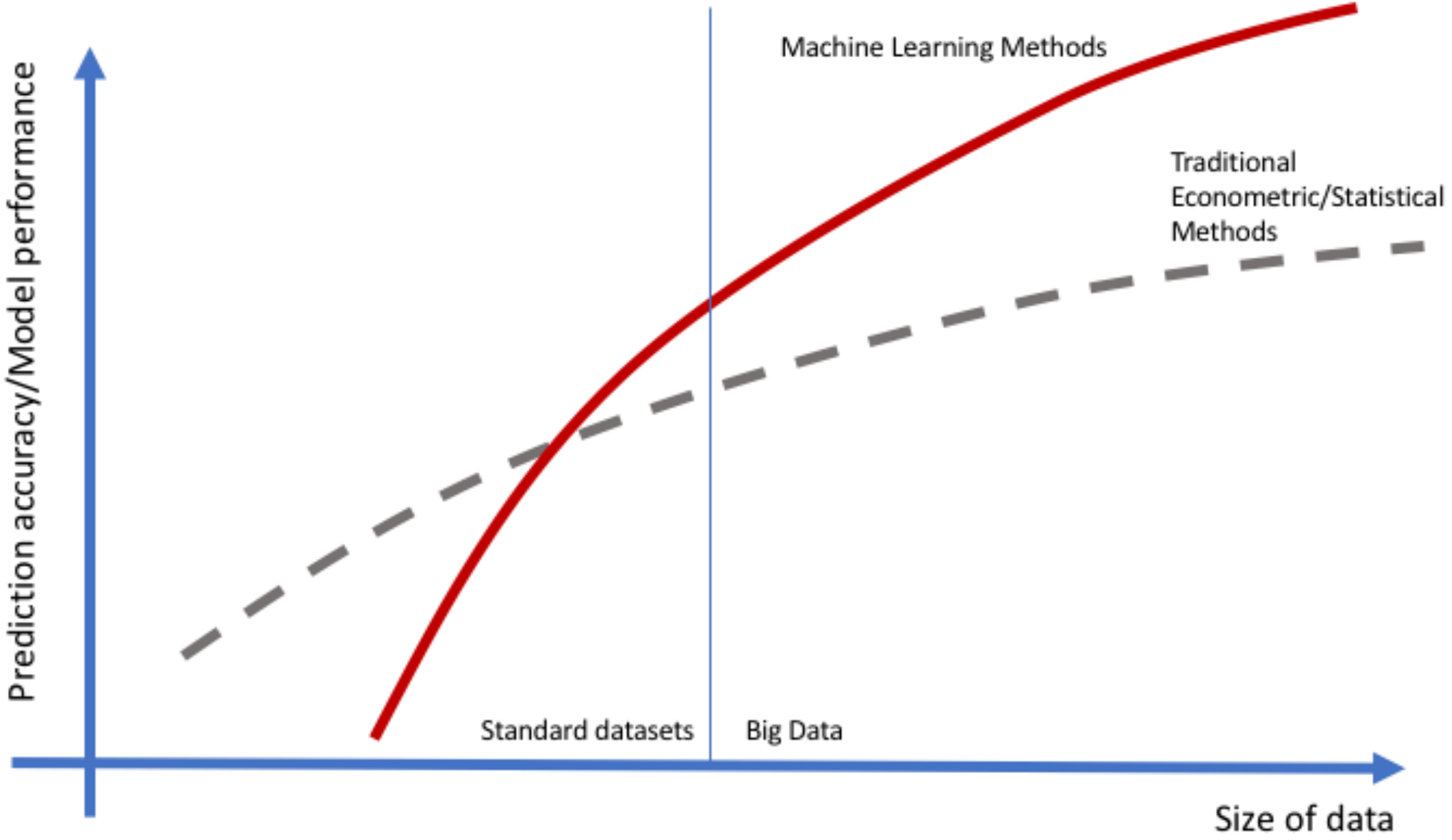
MICRO-LEVEL ESTIMATION OF POVERTY AND INEQUALITY

BY CHRIS ELBERS, JEAN O. LANJOUW, AND PETER LANJOUW[1]

1. INTRODUCTION

RECENT THEORETICAL ADVANCES have brought income and wealth distributions back into a prominent position in growth and development theories, and as determinants of specific socio-economic outcomes, such as health or levels of violence. Empirical investigation of the importance of these relationships, however, has been held back by the lack of sufficiently detailed high quality data on distributions. Household surveys that include reasonable measures of income or consumption can be used to calculate distributional measures, but at low levels of aggregation these samples are rarely representative or of sufficient size to yield statistically reliable estimates. At the same time, census (or other large sample) data of sufficient size to allow disaggregation either have no information about income or consumption, or measure these variables poorly. This note outlines a statistical procedure to combine these types of data to take advantage of the detail in household sample surveys and the comprehensive coverage of a census. It extends the literature on small area statistics (Ghosh and Rao (1994), Rao (1999)) by developing estimators of population parameters that are nonlinear functions of the underlying variable of interest (here unit level consumption), and by deriving them from the full unit level distribution of that variable.

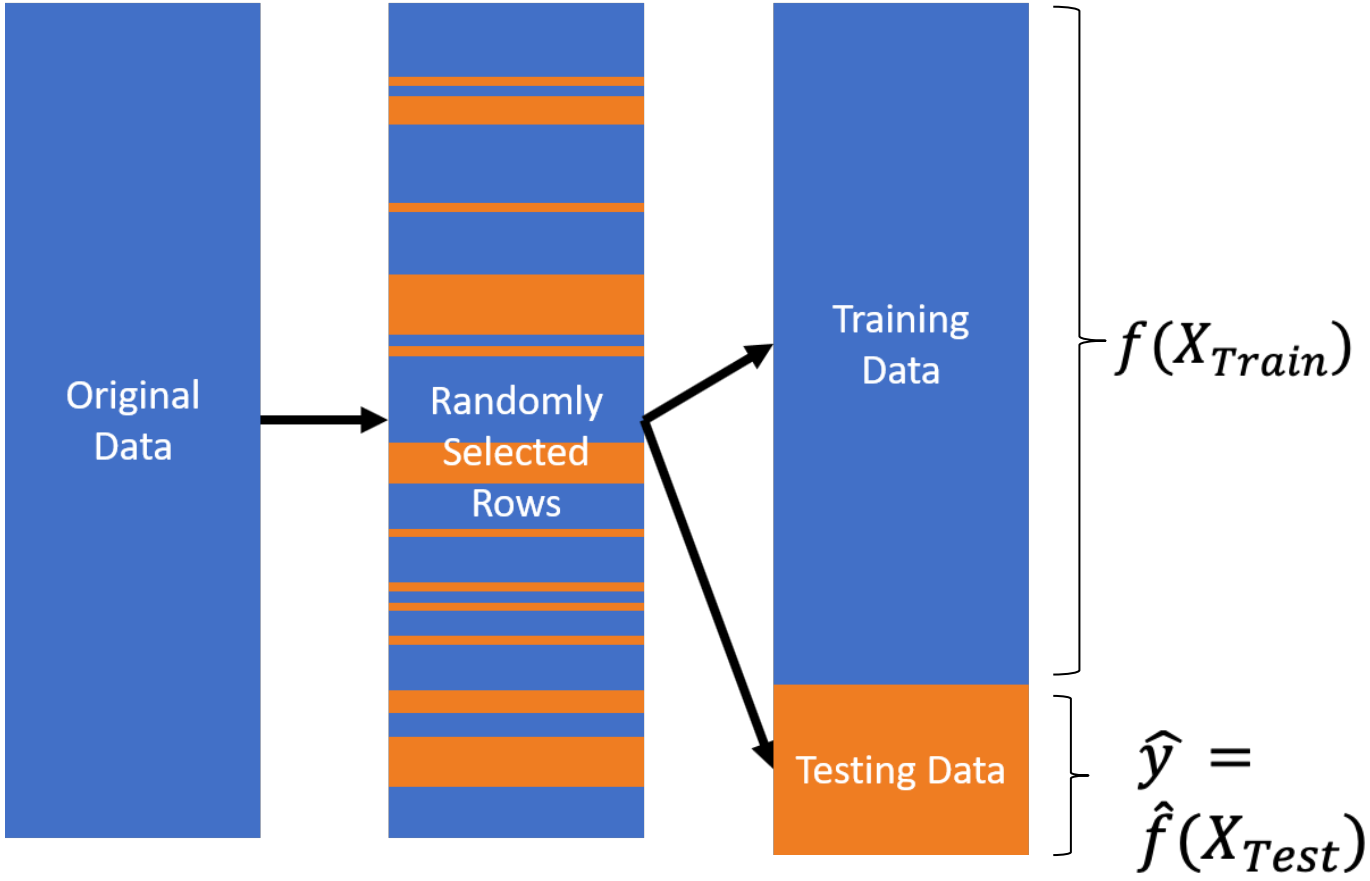
In examples using Ecuadorian data, our estimates have levels of precision comparable to those of commonly used survey based welfare estimates—but for populations as small as 15,000 households, a 'town.' This is an enormous improvement over survey based estimates, which are typically only consistent for areas encompassing hundreds of thousands, even millions, of households. Experience using the method in South Africa, Brazil, Panama, Madagascar, and Nicaragua suggest that Ecuador is not an unusual case (Alderman et al. (2002), and Elbers et al. (2002)).



Source: Harding and Hersh (2018) “Big Data in Economics”

Model Validation: Only Use "Test" Set

- We're also going to be very careful about testing and training split
- Recall: training data: data used to fit model
- Testing data: data used to validate model
- We split out data into 75% training and 25% test



2. Data

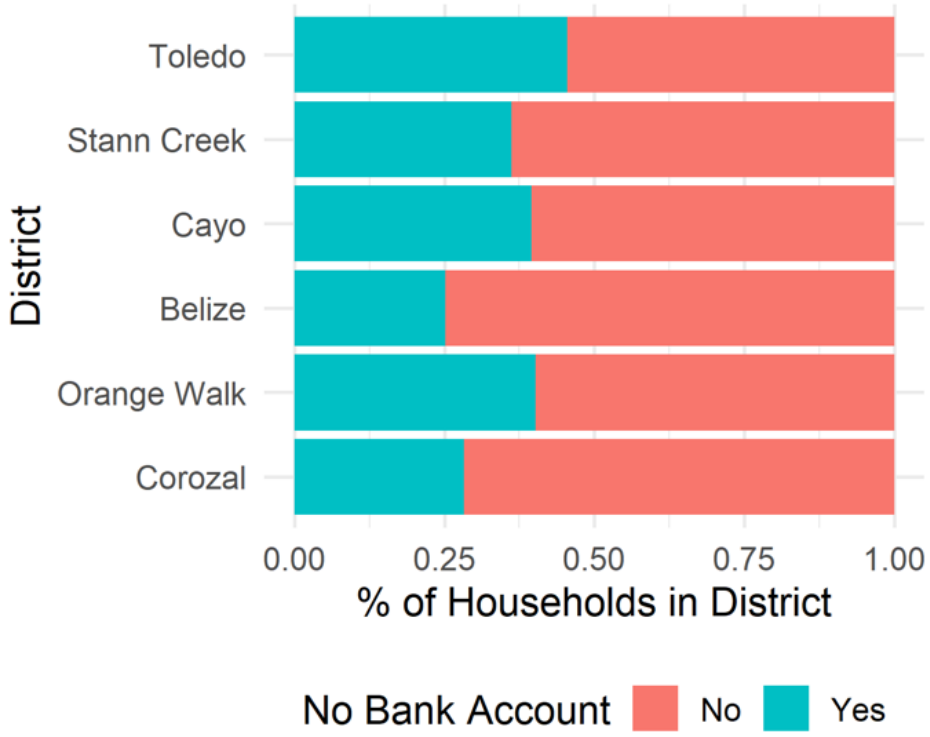
April 2019 Wave Labor Force Survey

Variable Name	Mean	Std. Dev
Household Financial Inclusion Metrics		
Household does not have a bank account	33%	(0.47)
Primary reason for household not having a bank account because not enough money	23%	(0.42)
Anyone in household has ever used online banking	10%	(0.30)
Anyone in household has borrowed formally	22%	(0.41)
Household-Level Characteristics		
Household is in an urban area	48%	(0.50)
Household owns home	61%	(0.49)
Household rents home	27%	(0.44)
House out walls made of poor material	31%	(0.46)
House floors made of poor material	37%	(0.48)
Toilet not in septic or sewer	23%	(0.42)
House not on electric grid	93%	(0.25)
Number of bedrooms in house	2.31	(1.03)
Household has an air conditioner	10%	(0.31)
Household has a refrigerator	80%	(0.40)
Household has a microwave	44%	(0.50)
Household has a washing machine	80%	(0.40)
Household has a stereo	60%	(0.49)
Household has a DVD player	28%	(0.45)
Household has a TV	80%	(0.40)
Household has a cellphone	94%	(0.23)
Household has a computer	37%	(0.48)
Household has a vehicle	41%	(0.49)
Household has cable	47%	(0.50)
Household has internet	59%	(0.49)
Number of household members	3.62	(2.06)
Number of children in household	1.29	(1.49)
Number of dependents in household	1.68	(1.47)
Number of adults in household	2.34	(1.19)
Total household size (persons)	10.51	(10.88)
Years of education head of household	8.22	(5.30)
Total Households (N)		2171

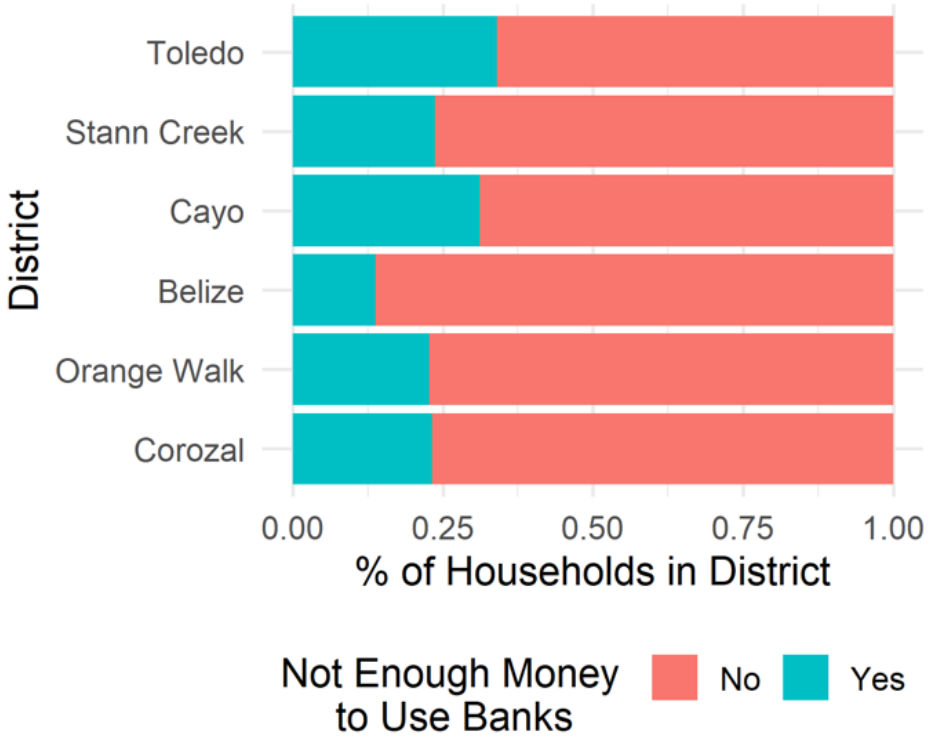
FI measures to predict using household characteristics

Household variables to build household model of FI

Measures of Financial Inclusion – Unbanked + Barriers to Banking

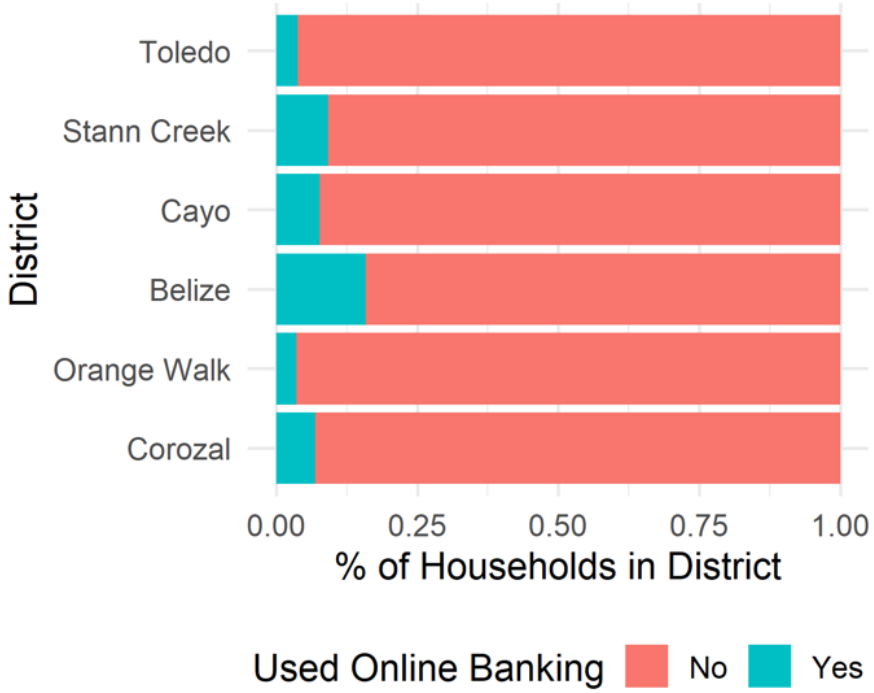


1) Does anyone in the household have an account at a credit union or a bank? (FI module question 1)

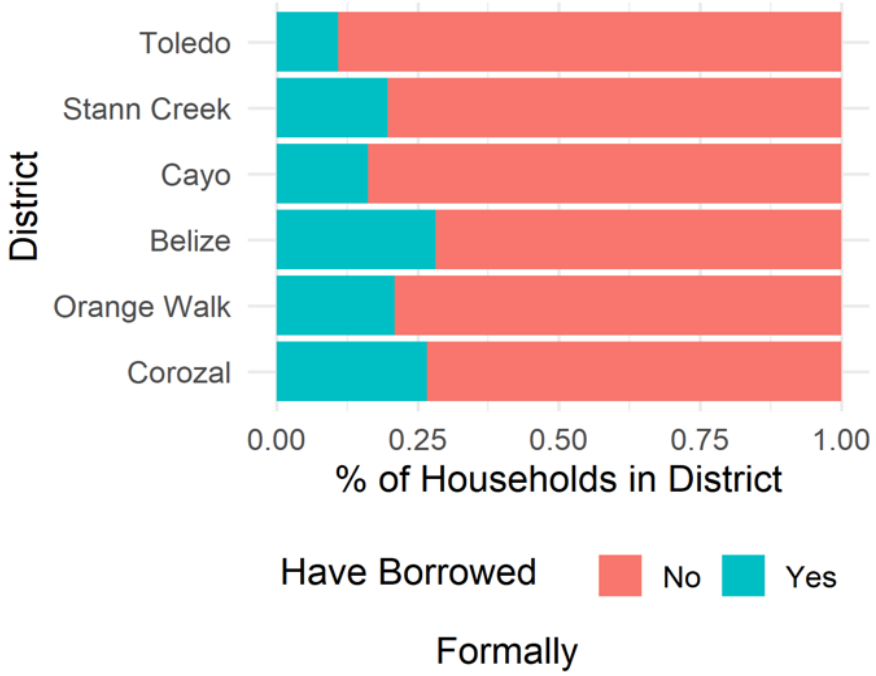


2) Is the reason you do not have an account at a bank or credit union because you don't have enough money to use them? (FI module question 3-F)

Measures of Financial Inclusion – Online Banking and Formal Borrowing



3) In the past 12 months have you ever used Internet / online banking? (FI module question 4-E)



4) In the past 12 months, have you borrowed any money from a bank, credit union or another type of formal institution? (FI module question 7)

3. Machine Learning Model to Predict Household Financial Inclusion

Modeling Approach

2010 Belize Census
- 75,000 households

Labor Force Survey
- April 2019 Wave
- 2,216 unique households

Household level outcome to model: One of four Financial Inclusion Question

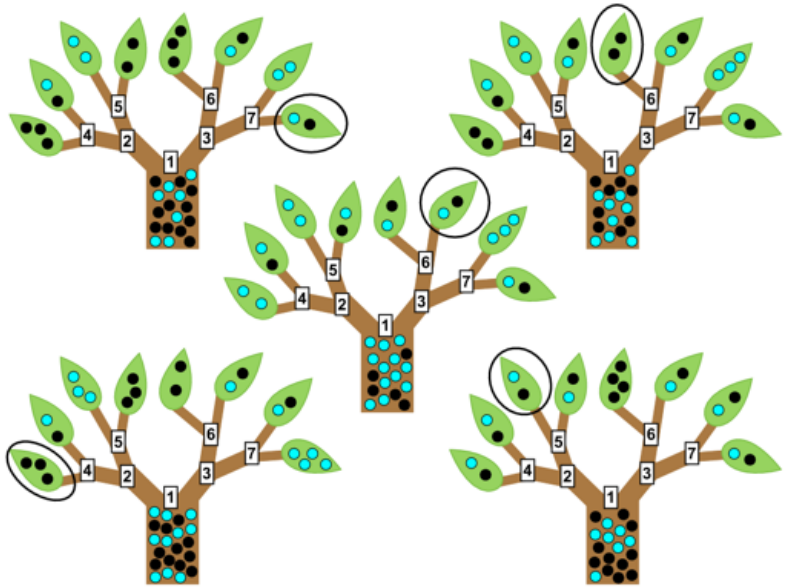
Test/Validation approach
- 75% data training (estimation) sample and 25% test (validation)

Models
- Random Forest each made up of 500 decision trees

Intersection of Census and LFS Surveys had 26 transformed variables

Random Forests

- Random forests are a slight trick to bagging that highly improves predictive power
- **Many trees do poorly because the stepwise greedy algorithm doesn't fully explore variable and parameter space**
- Random forests is like bagging, only each time a split in a tree is considered, a random selection of m predictors is chosen as split candidate
- A fresh set of m predictors is taken at each split.



Random Forest Model: Many Decision Trees

B bootstrap sampes of data (~ 1000)

…

Bootstrap 1

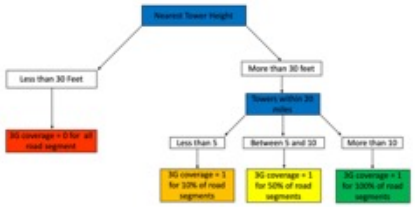
Bootstrap b

Bootstrap 1000

Tree 1

Tree b

Tree 1000



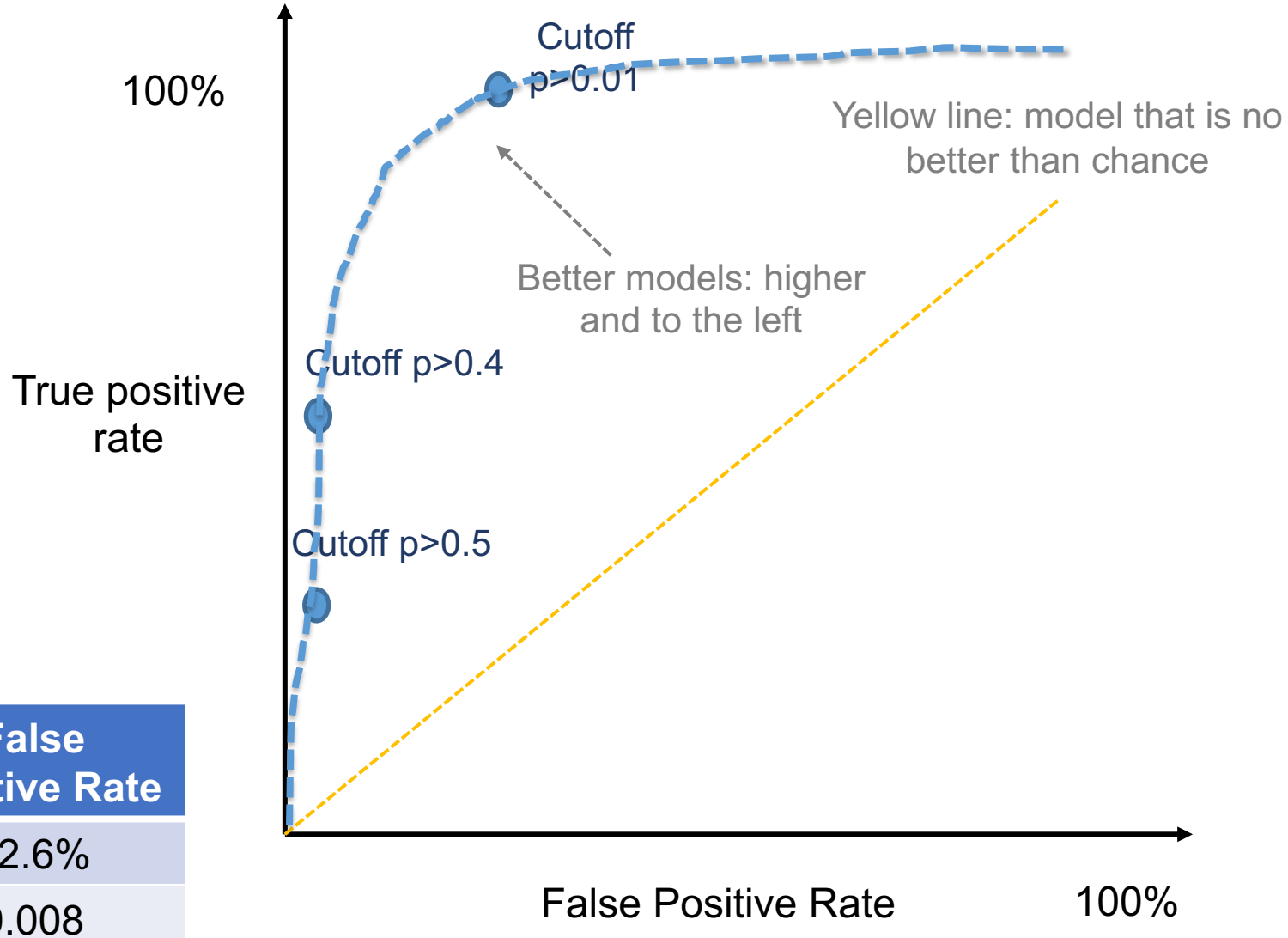
Vote household i has FI access = 0 or 1

Vote household i has FI access = 0 or 1

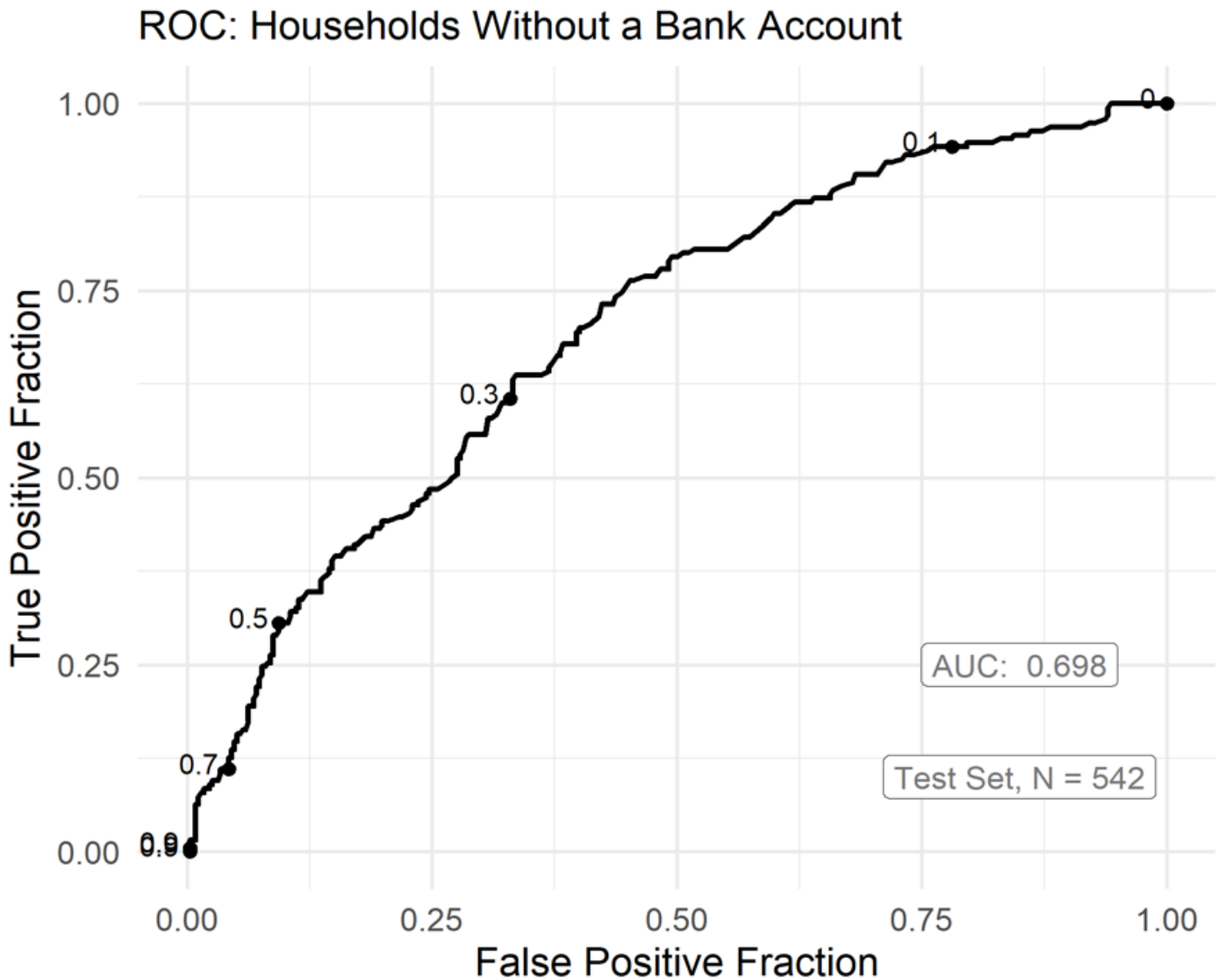
Vote household i has FI access = 0 or 1

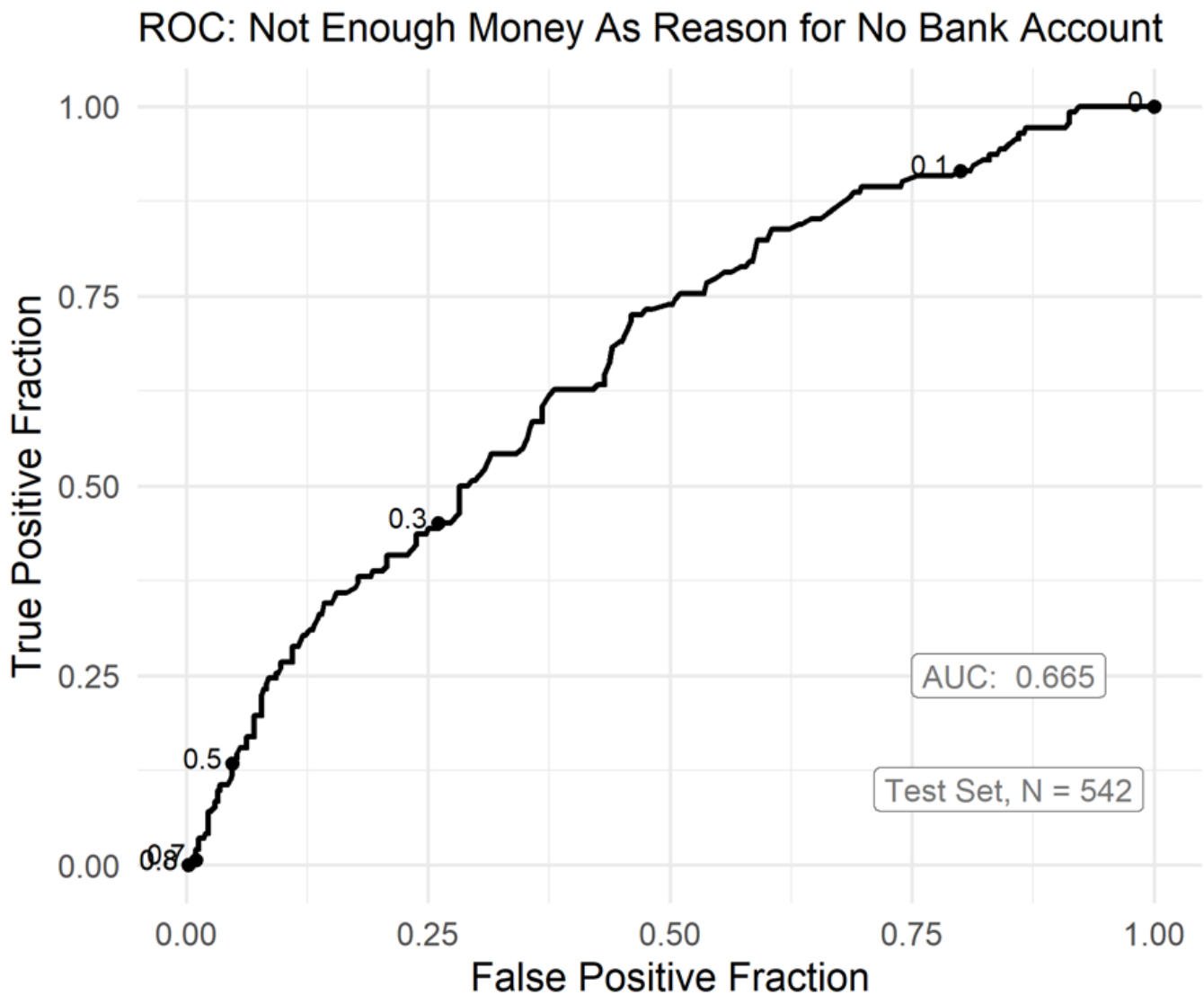
Binary Model Accuracy Measure: ROC Curve

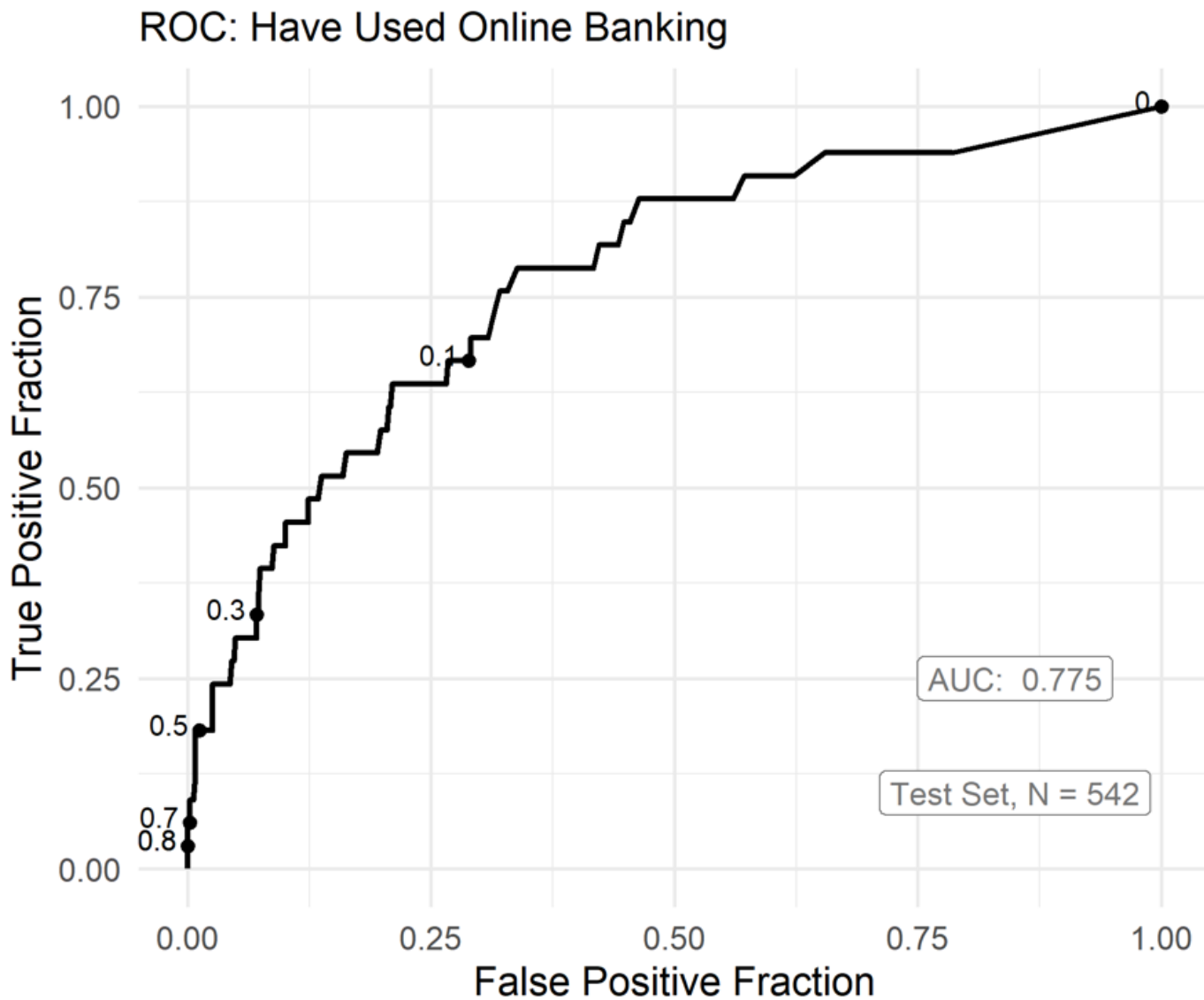
- Models return household probability of FI access
- Can we show consequences of False Positives and Negatives as we vary the cutoff probability to assign classes?
- That is a ROC (Receiver Operator Curve) plot

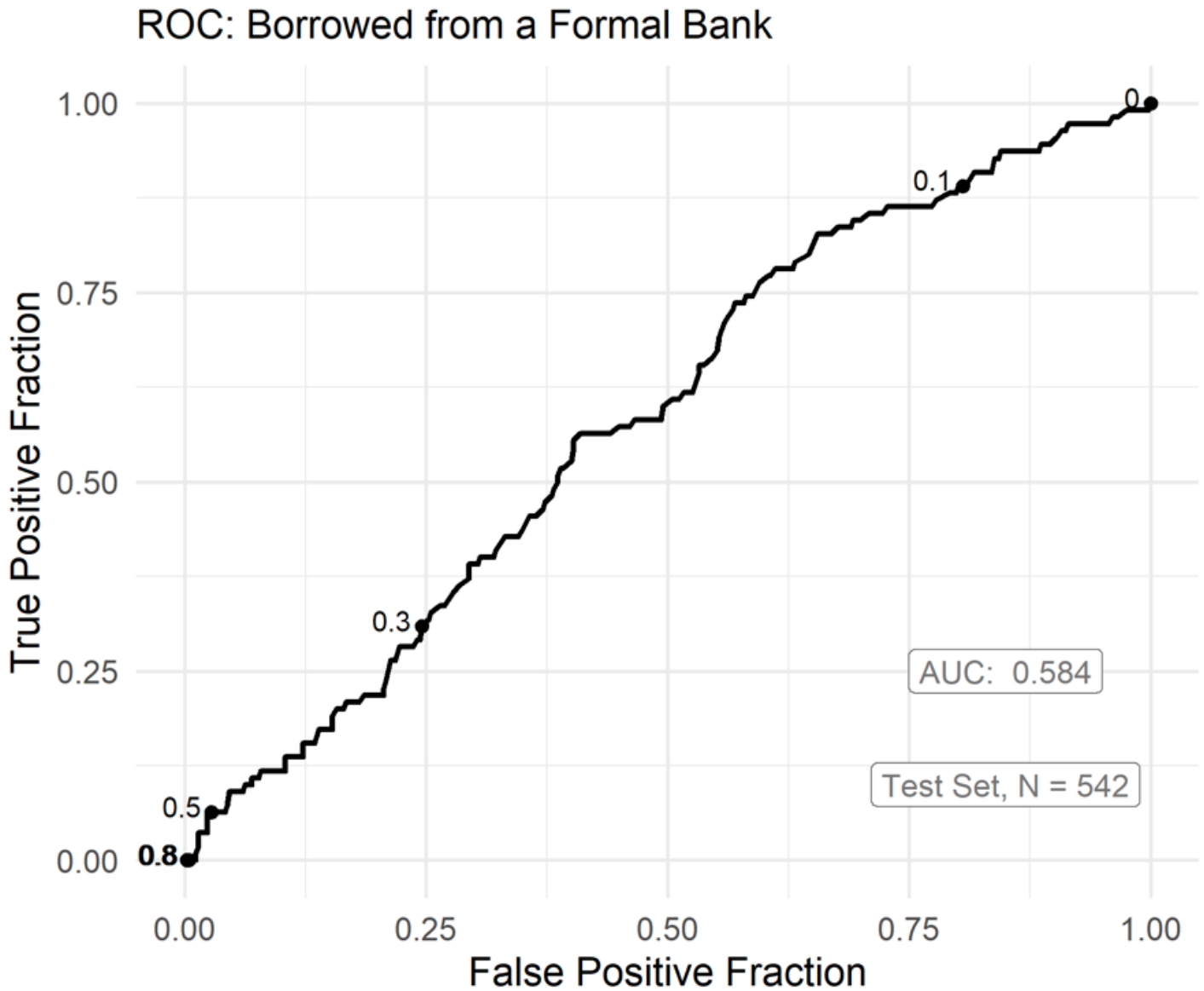


Probability Cutoff	True Positive Rate	False Positive Rate
0.01	100%	22.6%
0.4	57%	0.008
0.5	21.4%	0.004%
0.6	21.4%	0.002%


ROC: Households Without a Bank Account
True Positive Fraction
False Positive Fraction
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
0
0.1
0.3
0.5
0.7
AUC: 0.698
Test Set, N = 542


ROC: Not Enough Money As Reason for No Bank Account
True Positive Fraction
False Positive Fraction
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
0
0.1
0.3
0.5
0.7
0.8
AUC: 0.665
Test Set, N = 542

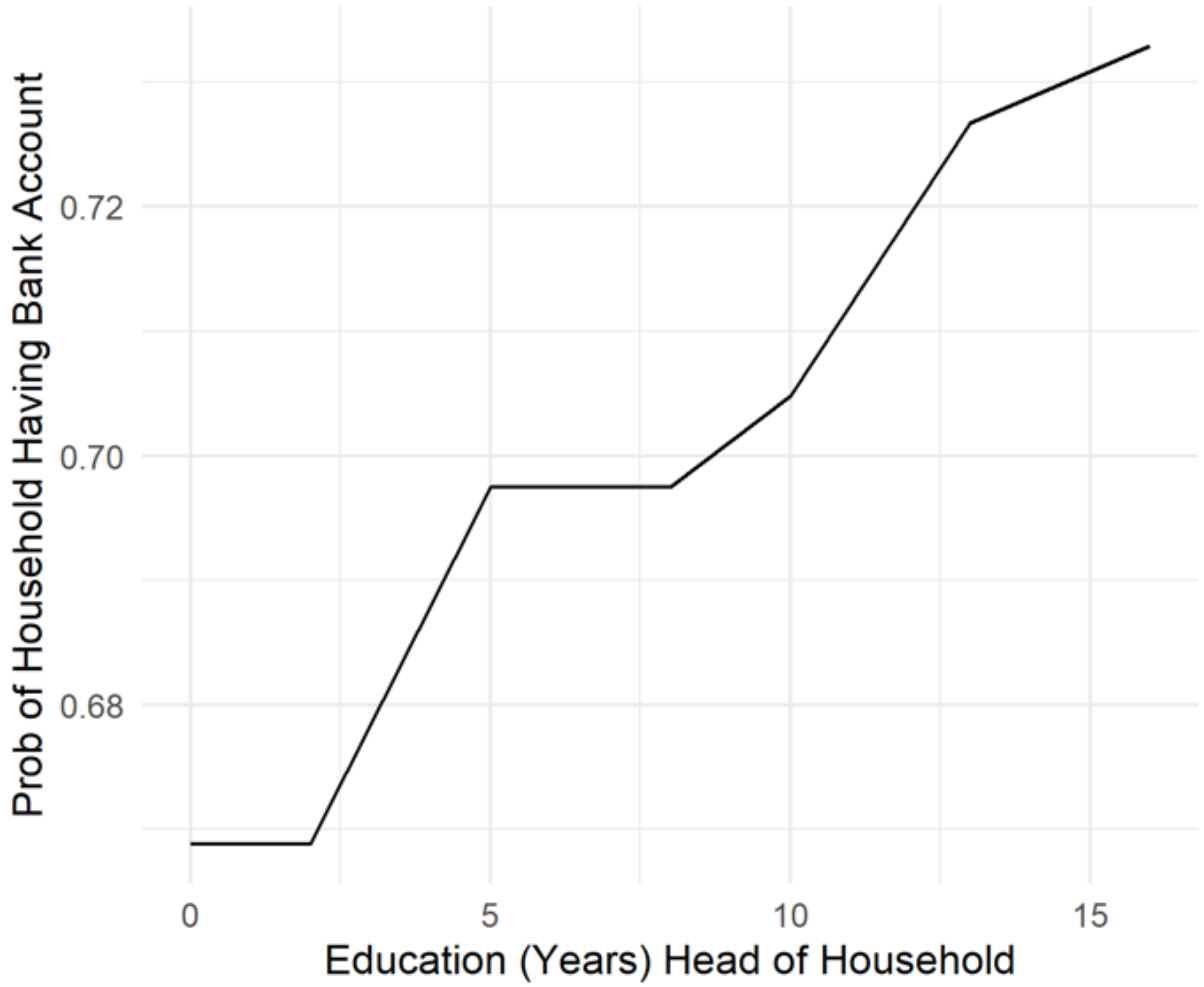

ROC: Have Used Online Banking
True Positive Fraction
False Positive Fraction
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
0
0.1
0.3
0.5
0.7
0.8
AUC: 0.775
Test Set, N = 542


ROC: Borrowed from a Formal Bank
True Positive Fraction
False Positive Fraction
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
0
0.1
0.3
0.5
0.8
AUC: 0.584
Test Set, N = 542

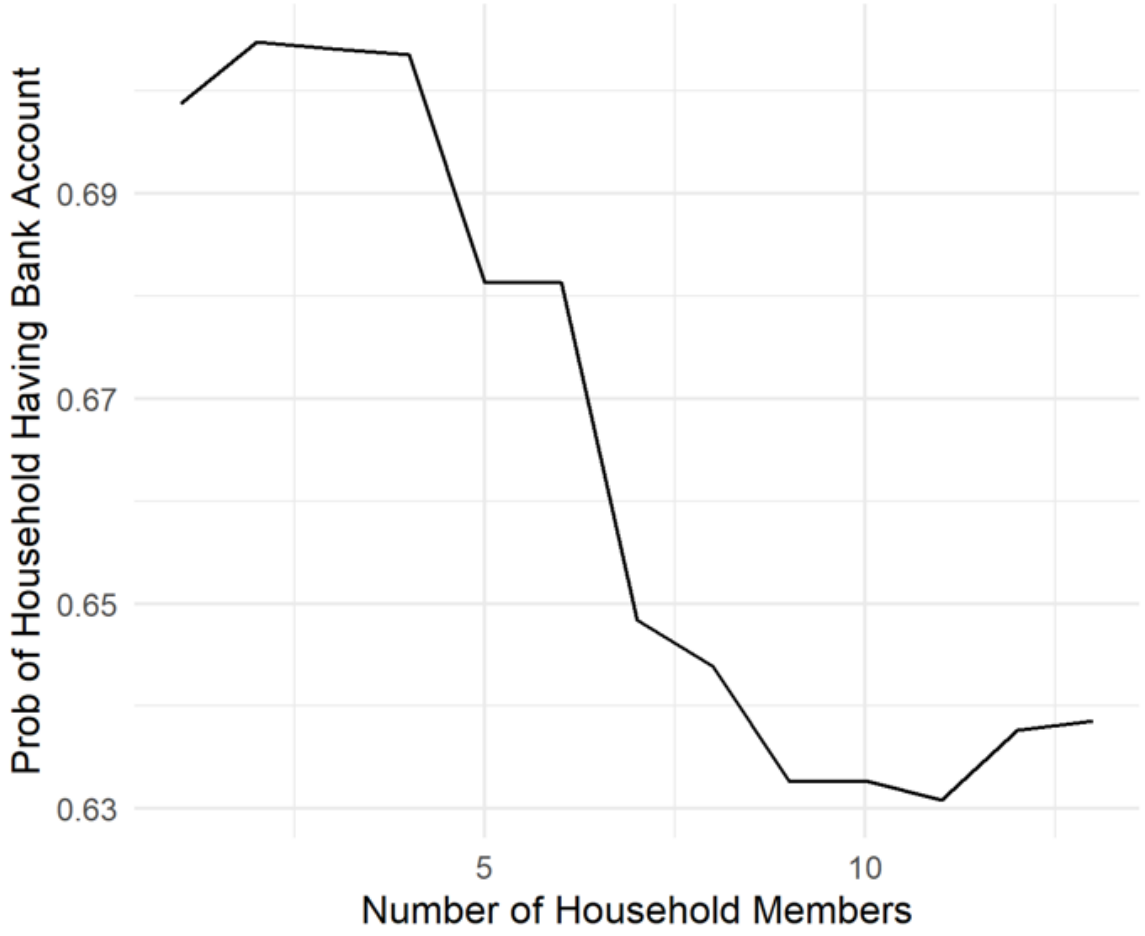
Overall Model Accuracy: Area Under ROC Curve

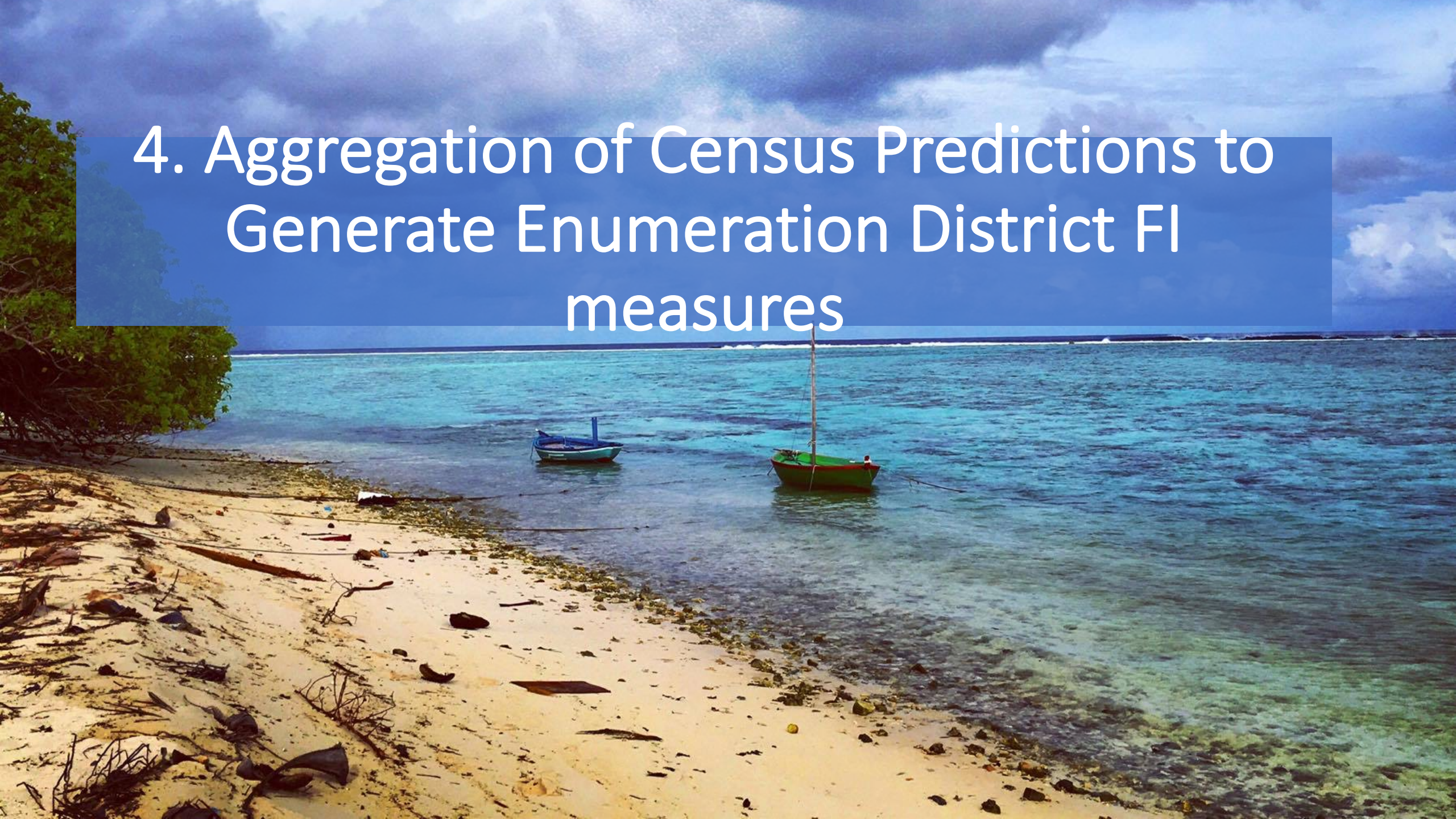
AUC (Test Set)	Outcome Variable
0.698	Household does not have a bank account
0.665	Primary reason for household not having a bank account because not enough money
0.775	Anyone in household has ever used online banking
0.584	Anyone in household has borrowed formally

Head of Household Education As Predictor of Bank Account Access:



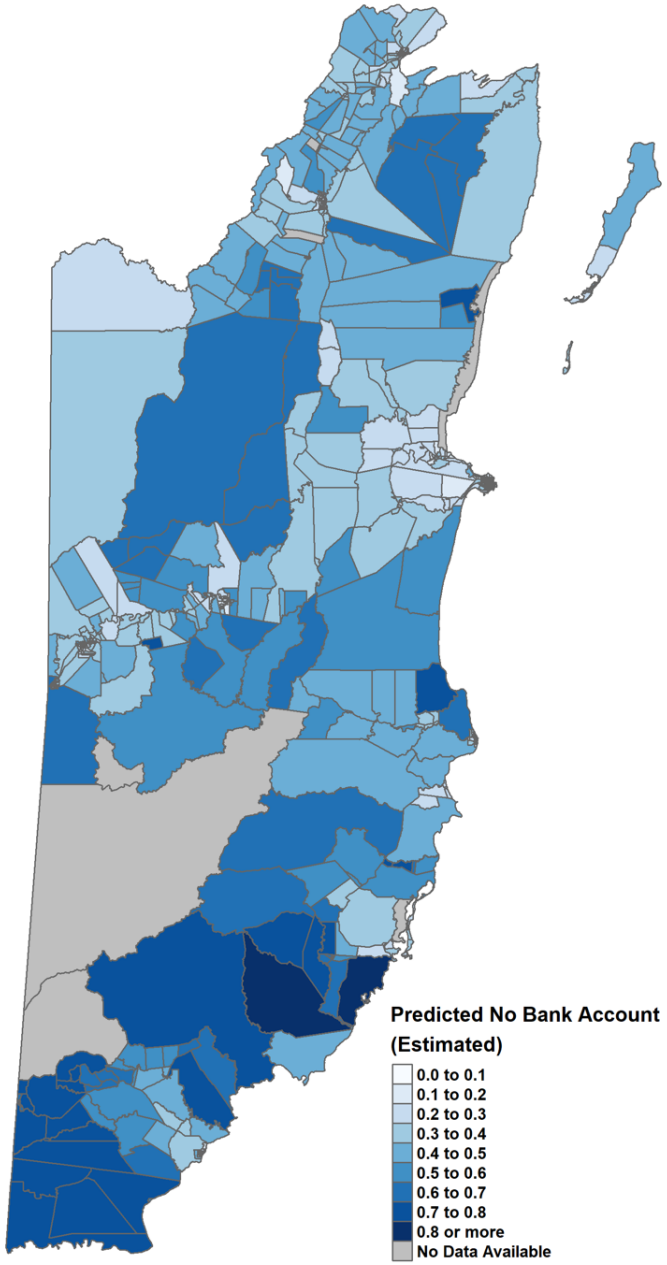
Number of Household Members As Predictor of Bank Account Access:



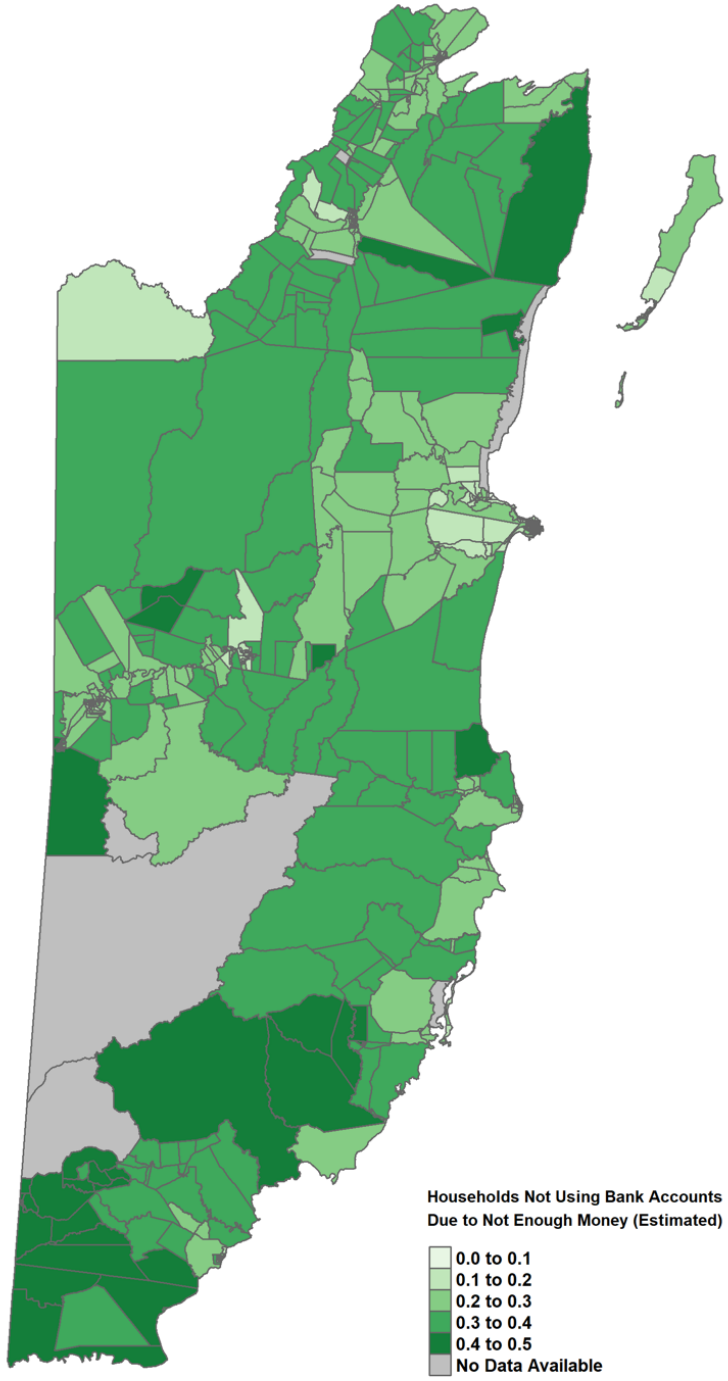


4. Aggregation of Census Predictions to Generate Enumeration District FI measures

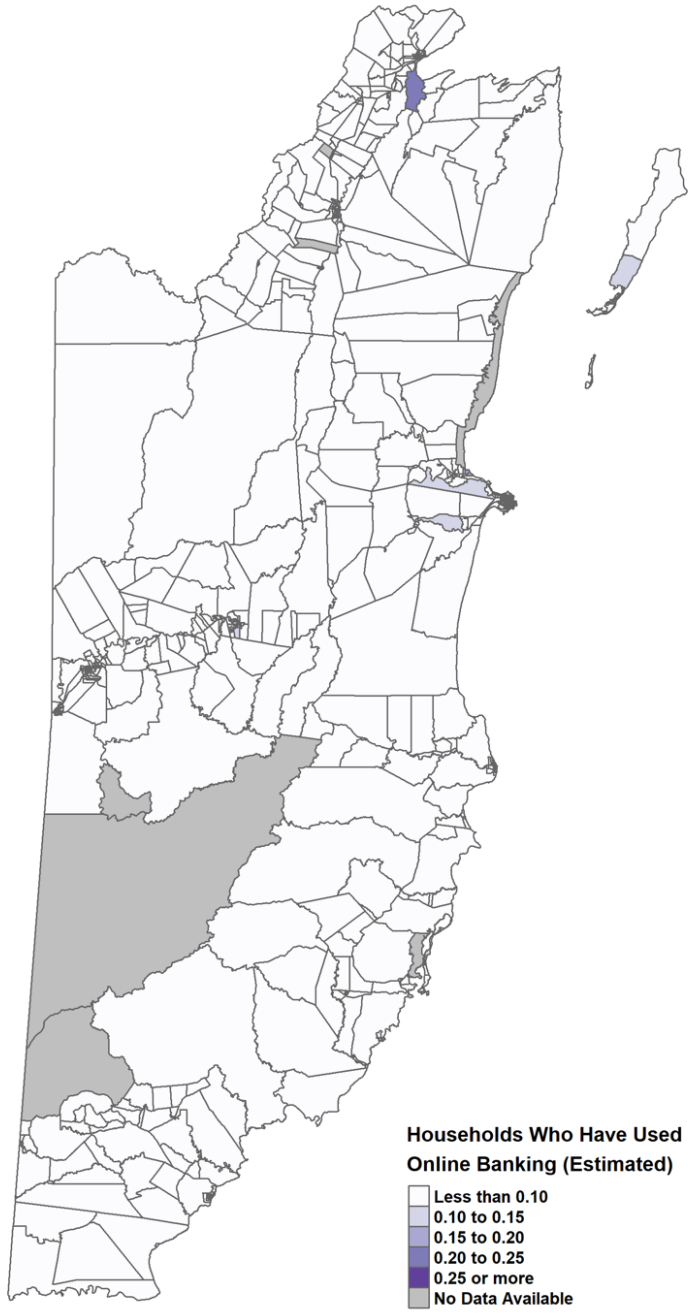
Estimated % of Households in ED Without Bank Accounts



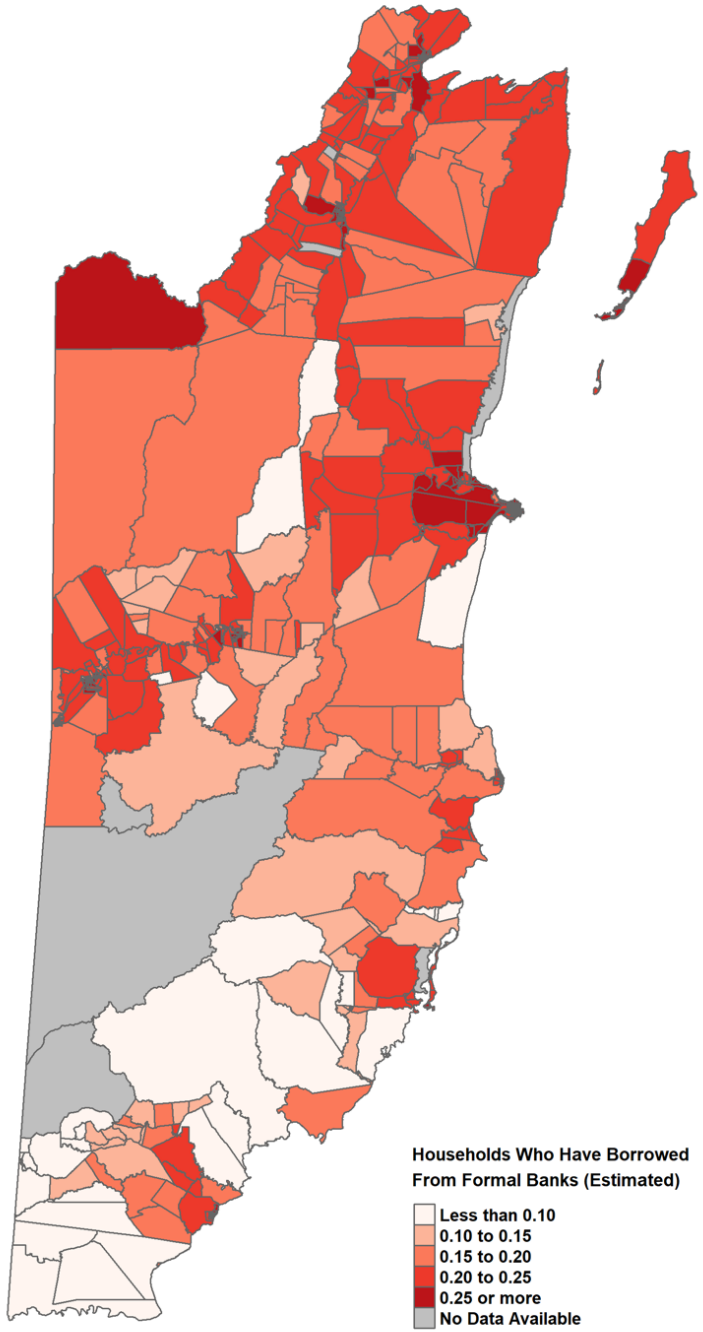
Estimated % of Households in ED Who State They Do Not Have a Bank Account Due to Not Enough Money



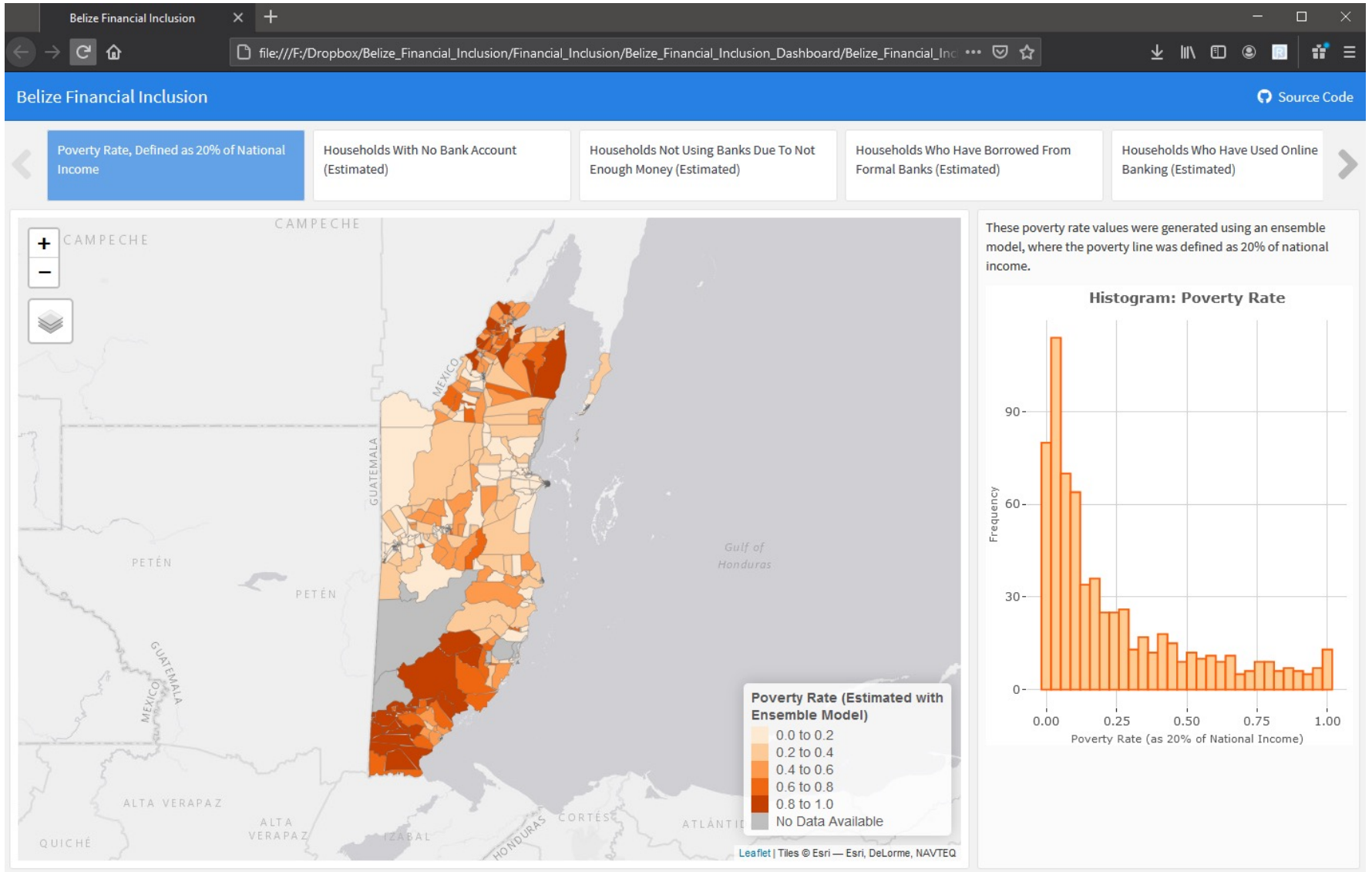
Estimated % of Households in ED Who Have Used Online Banking



Estimated % of Households in ED Who Have Borrowed Formally



Dashboard To Interactively View Results



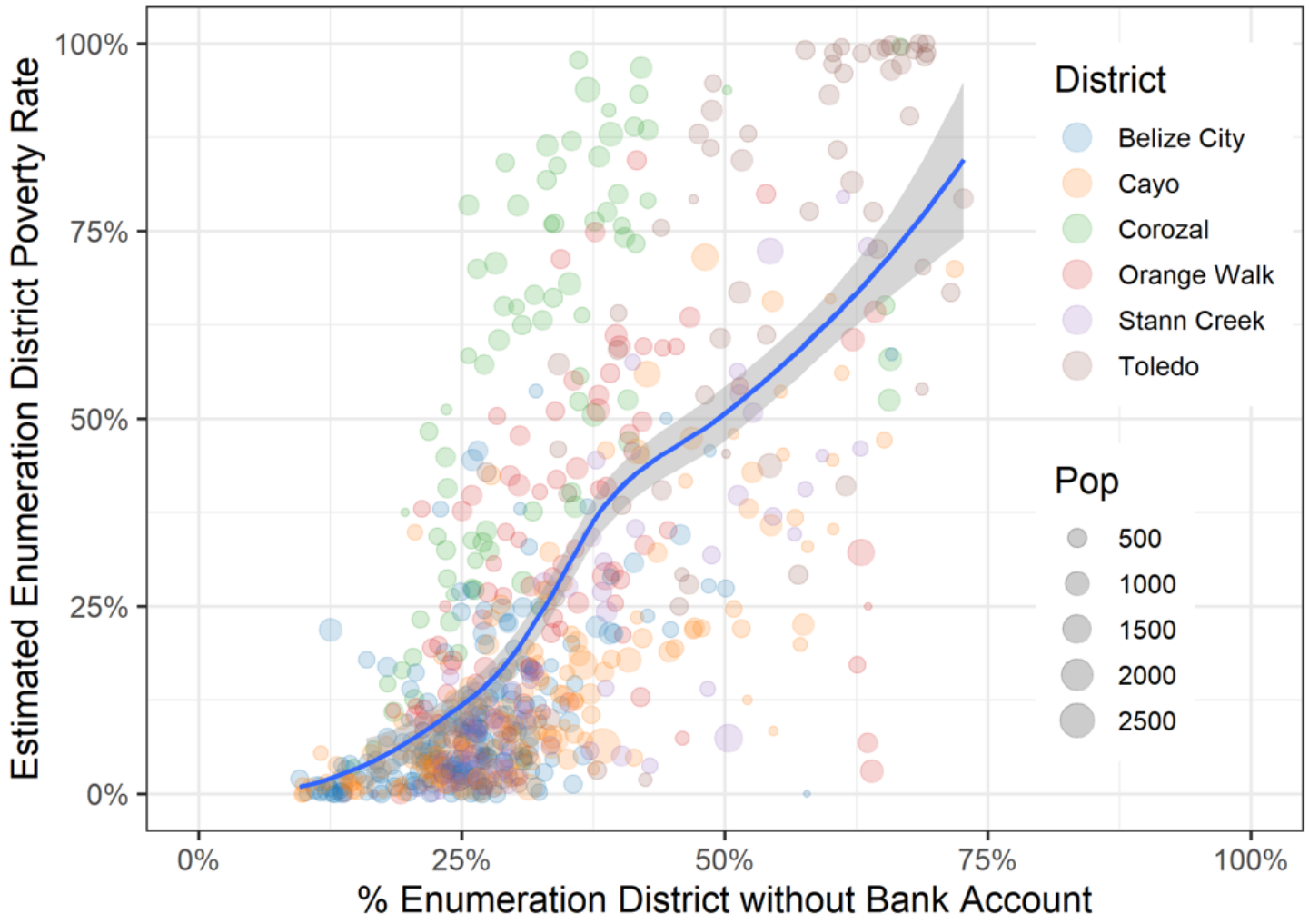
Click Link Below

flexdashboard_final.html

5. ED Characteristics Associated with FI

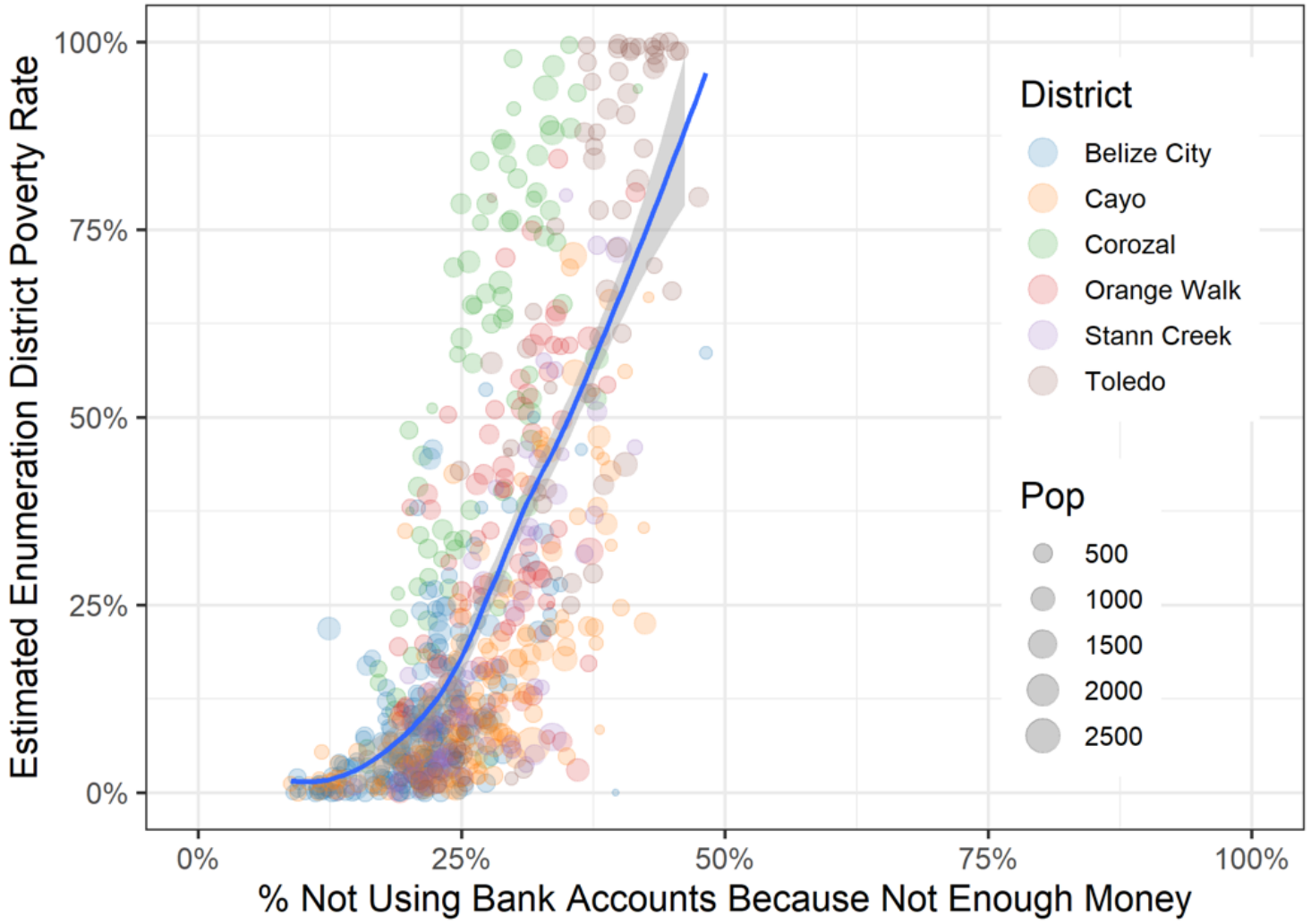
Dependent Variable:	% of households in Enumeration District that have bank accounts	% of households in Enumeration District who do not have a bank account due to not having enough money
Urban = 1	0.027*** (0.009)	-0.037*** (0.008)
Religion = Pentecostal	-0.041 (0.029)	0.047* (0.027)
Religion = Roman Catholic	0.099*** (0.027)	-0.094*** (0.025)
Religion = Seventh Day Adventist	0.102*** (0.039)	-0.064* (0.037)
Fixed Phone = 1	0.291*** (0.025)	-0.269*** (0.023)
Log of ED Population	0.040 (0.052)	0.218*** (0.048)
Language = English	0.165*** (0.020)	-0.163*** (0.019)
Language = Spanish	0.007 (0.046)	-0.007 (0.043)
Language = Garifuna/Yucatec/Ketchi/Mopan	**-0.461*** (0.045)**	**0.371*** (0.042)**
Language = German	**-0.459*** (0.044)**	**0.175*** (0.041)**
Ethnicity = Creole	**-0.243*** (0.039)**	**0.192*** (0.036)**
Ethnicity = Ketchi Maya	0.063* (0.034)	-0.047 (0.032)
Ethnicity = Mestizo	**-0.181*** (0.045)**	**0.211*** (0.042)**
Born in Belize	**0.284*** (0.032)**	**-0.121*** (0.030)**
Poverty Rate (20%)	**-0.245*** (0.021)**	**0.189*** (0.019)**
Constant	0.415*** (0.050)	0.355*** (0.046)
Observations	684	684
R2	0.813	0.805
Adjusted R2	0.809	0.8
Residual Std. Error (df = 668)	0.092	0.085
F Statistic (df = 15; 668)	193.518***	183.501***

Relationship Between Enumeration District Poverty and Unbanked



- Higher poverty EDs tend to have greater share of unbanked

Relationship Between Enumeration District Poverty and Unbanked



- Even stronger relationship between ED poverty and % of households in an ED who state they don't have a bank account because they don't have enough money

6. Policy Recommendations to Increase FI

Key Policy Recommendations

1. **Significant sub-District level heterogeneity in FI**
    - More study is needed. Is this supply or demand driven?
    - If Supply, consider more financial access points.
    - If Demand, consider geographically targeted subsidies

2. **ED Poverty is the strongest predictor of FI**
    - Consider targeted financial products to the poor with zero or negative cost (subsidy)

3. **Ethnic Minorities and Foreign Born Much Less Likely to have FI**
    - Consider targeted outreach in German, Garifuna, Yucatec, Ketchi, or Mopan.

6. **F**eedback and Questions