

# Using Social Media Advertising Data to Map Poverty of Migrants and Refugees

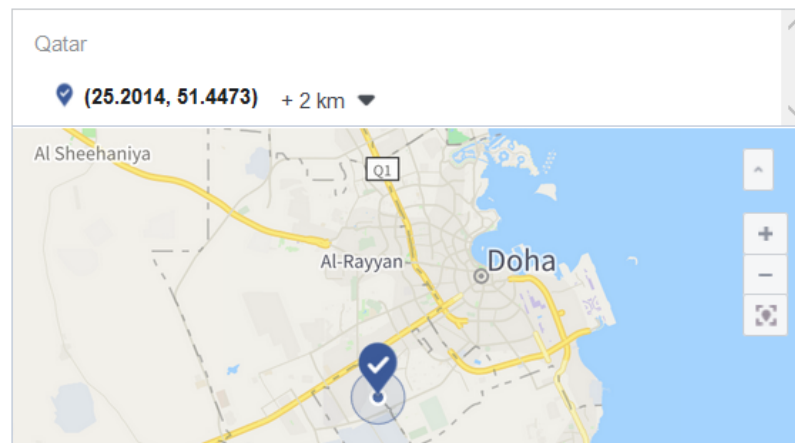
Ingmar Weber

Regional Workshop on Poverty Measurement  
in the Era of Big Data

December 22-23, 2020

**Audience**Define who you want to see your ads. [Learn more.](#)

Locations ⓘ

**Audience size**

Your audience selection is broad. This requires a large budget.

Potential reach: 42,000 people ⓘ

**Estimated daily results**

Reach ⓘ

**2.7K-17K**

# LIVE DEMO

Gen

e Network  
ement,  
hance.

Behaviours &gt; Ex-pats

**Lived in India (formerly Ex-pats – India)****Lived in Nepal (formerly Ex-pats – Nepal)**

Add demographics, interests or behaviours

Suggestions

Browse

Detailed targeting ⓘ

and MUST ALSO match at least ONE of the following ⓘ

Behaviours &gt; Mobile Device User &gt; All Mobile Devices by Operating System

**Facebook access (mobile): Android devices**

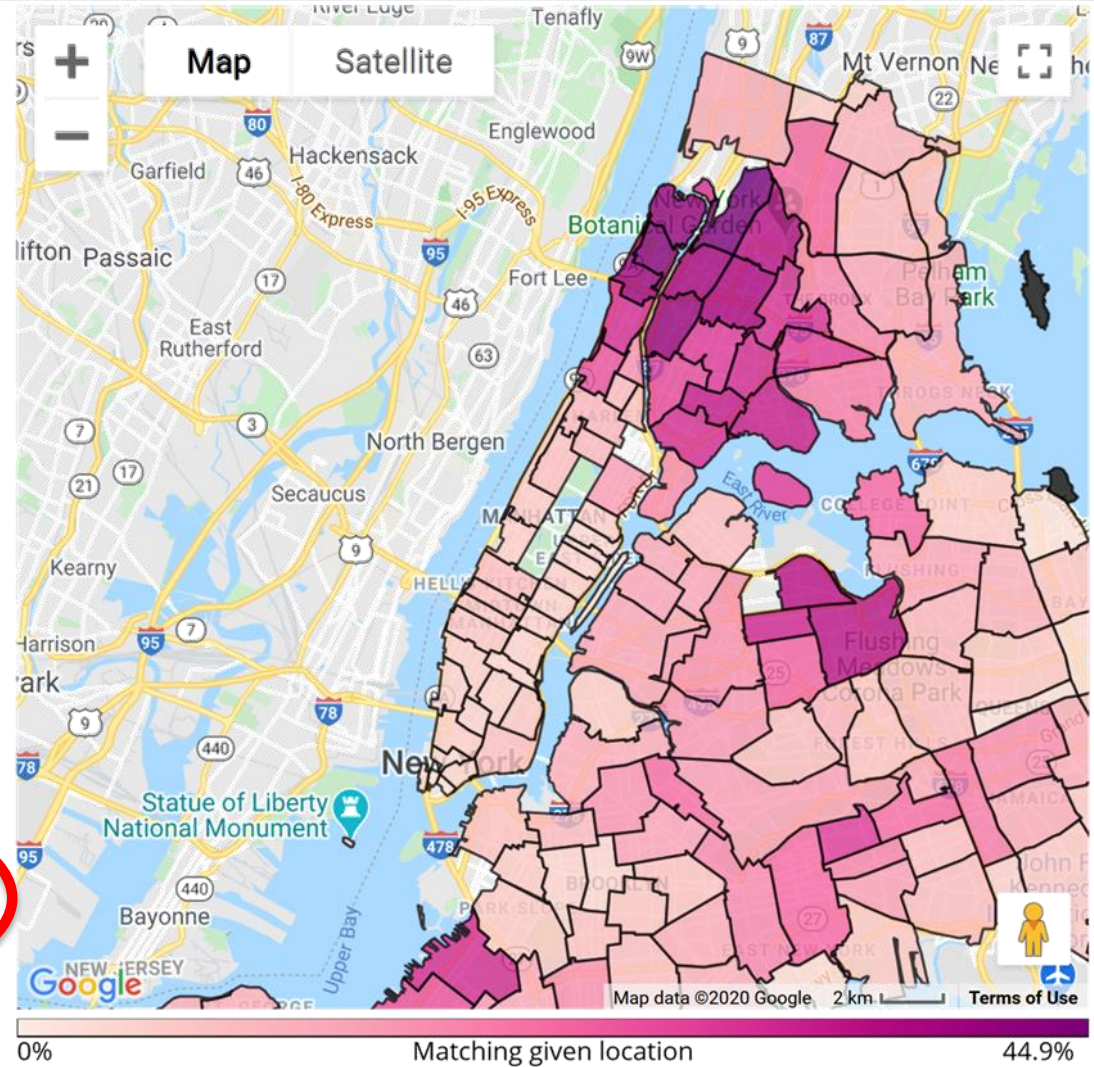
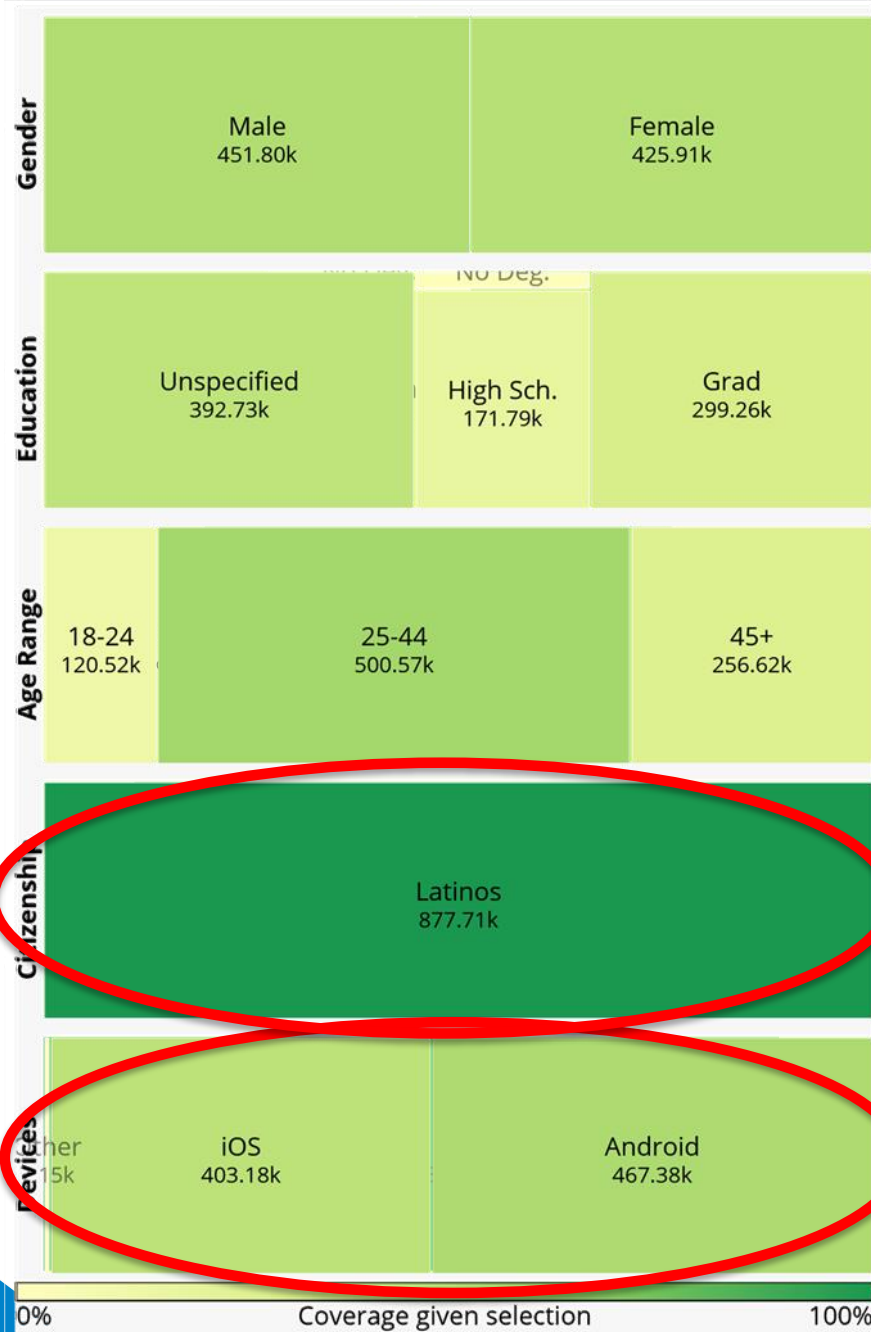
Add demographics, interests or behaviours

Suggestions

Browse

The accuracy of estimates is based on factors such as past campaign data, the budget you've entered and market data. Numbers are provided to give you an idea of performance for your budget, but are only estimates and don't guarantee results.

[Were these estimates helpful?](#)



All Locations

No Locations

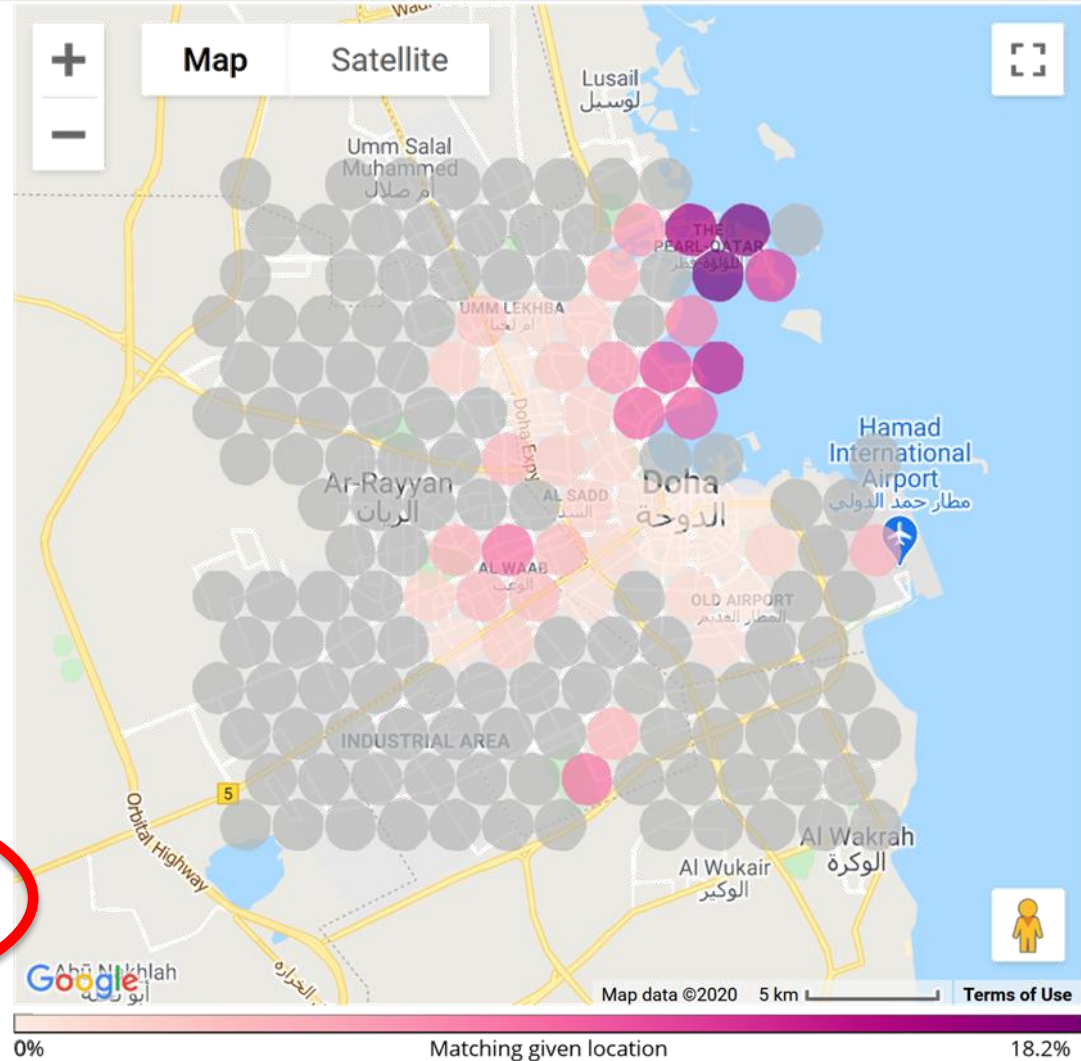
Share what you see:



<https://fb-nyc.qcri.org>







<https://fb-doha.qcri.org>



# VALIDATING IN A NON-REFUGEE SETTING



# Ground Truth and Training Data



- Representative sample of ~40 households for each “cluster” (PH: n=1.2k, IN: n=28k)
- Target variable: asset ownership based “Wealth Index”
- Philippines: 2017 & India: 2015-2016 vs. FB features from 2019

## Sampling noise

Wealth index depends on particular households

Expected  $R^2 = .955/.973$  (PH/IN, bootstrap estimate)

## Spatial perturbation

True location is  $(x,y)$ , but reported at  $(x',y')$

Protects privacy

Expected  $R^2 = .885/.860$  (PH/IN, simulations)

## Combined

**Expected  $R^2 = .845/.838$  (PH/IN)**

“Expected upper bound”, “explainable variance”

# Regression Setup

The DHS Wealth Index is the target ( $y$ )

Use the FB data as features ( $x$ )

Try to find a function  $f$  such as  $y \approx f(x)$

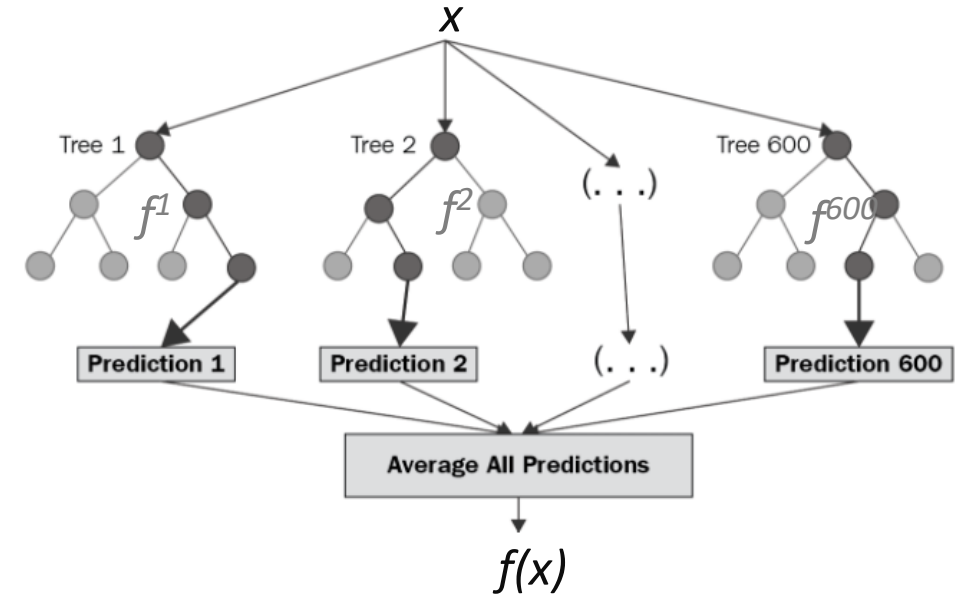
Use a gradient boosted regression forest as  $f$

Worked better than linear methods

Fitted and evaluated using 10-fold cross validation.

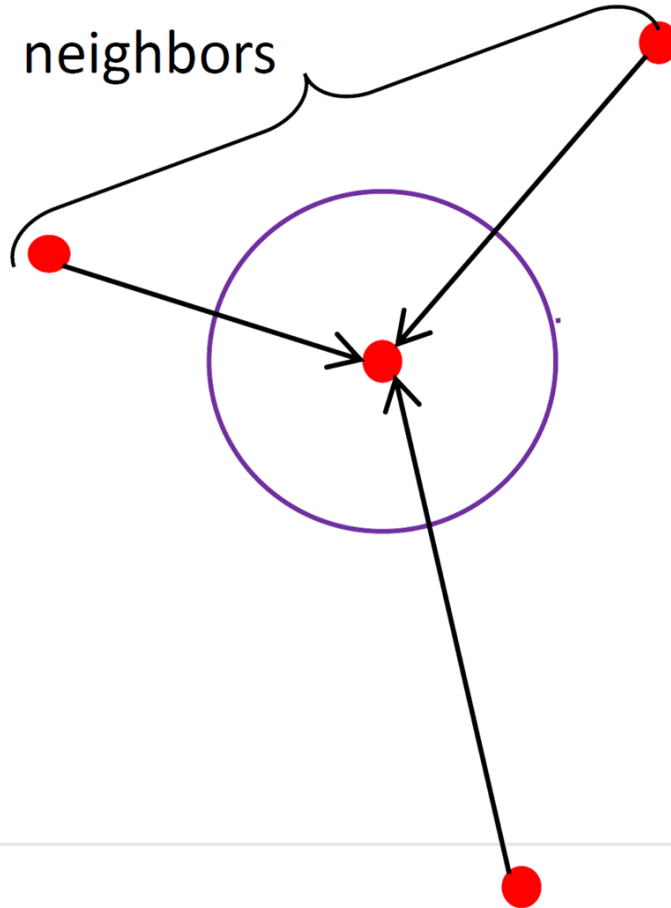
Implemented using the “gbm” package in R

<https://cran.r-project.org/web/packages/gbm/index.html>



# Baseline: Interpolation with Nearest Ground Truth

Average wealth  
index of  $k$  nearest  
neighbors



	$k = 1$	$k = 3$	$k = 5$	$k = 10$
Philippines	0.597	0.686	0.687	0.681
India	0.739	0.793	0.796	0.788

Pearson's  $r$  correlation





# Results

This uses the Wealth Index of nearby locations as features.

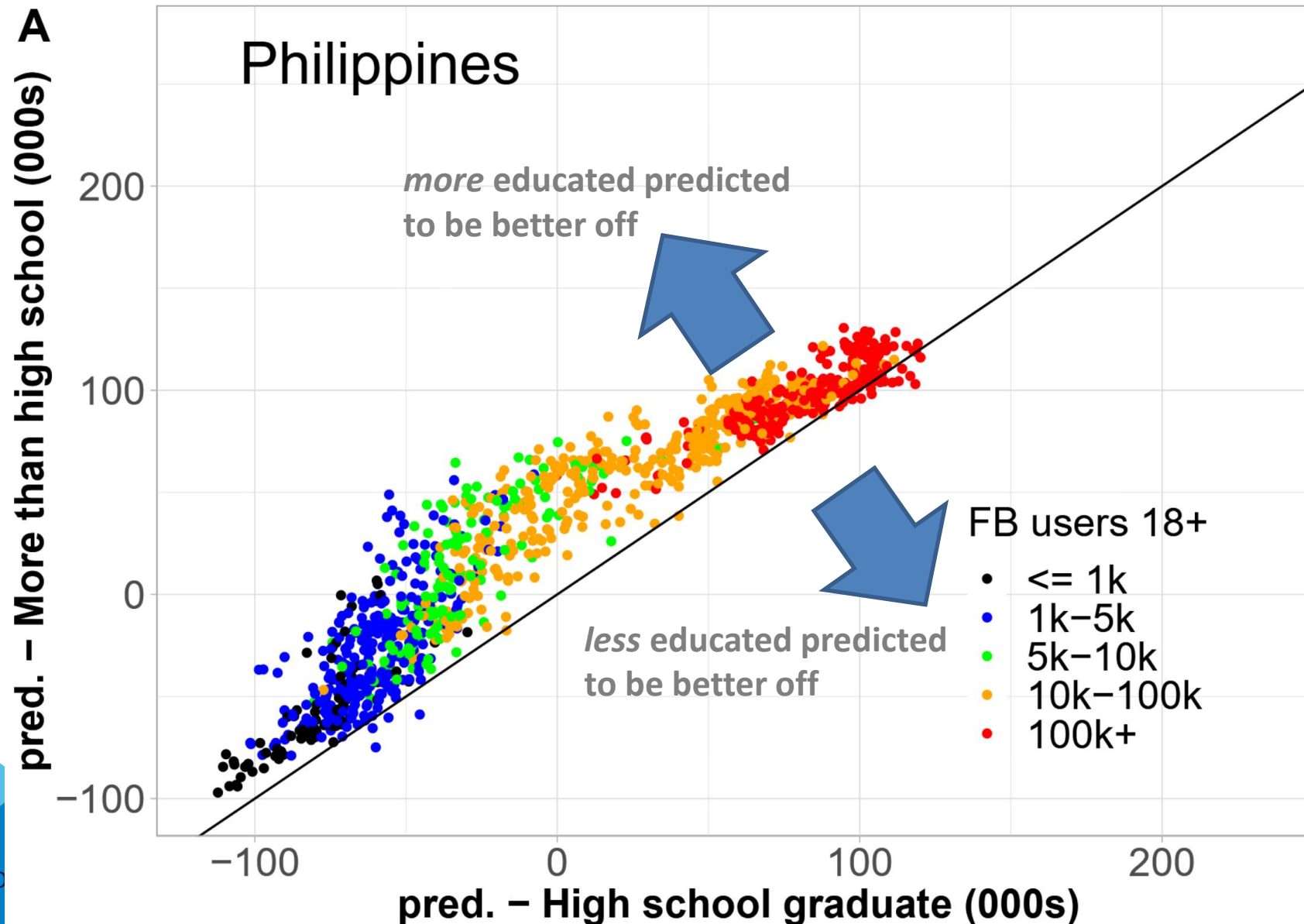
This includes the FB penetration, computed using high resolution settlement layer information.

This is a dummy variable “is the location part of [name of region]”.

Model features						
	Interpolated DHS Wealth Index		X			X
	Facebook features			X	X	X
	Log population density				X	X
	Regional indicators				X	X
Philippines ( $N = 1205$ )	$R^2$	0.480	0.608	0.627	0.630	
	RMSE	50,983	44,218	43,099	42,965	
India ( $N = 28,043$ )	$R^2$	0.652	0.563	0.691	0.728	
	RMSE	46,810	52,502	44,149	41,394	

Recall: upper bound  $R^2 = .845/.838$  (PH/IN) (due to noise)

# Education-Level-Disaggregated Predictions



# MONITORING THE VENEZUELAN EXODUS



Ingmar Weber (179...)

Ad Set Name ⓘ Young, male Venezuelans living in Norte de Santander ⚙️

Switch to

✓ Campaign

Objective ✓

Ad Set

Audience ✓

Placements ✓

Budget & Schedule ✓

Ad

Identity

Format

Media

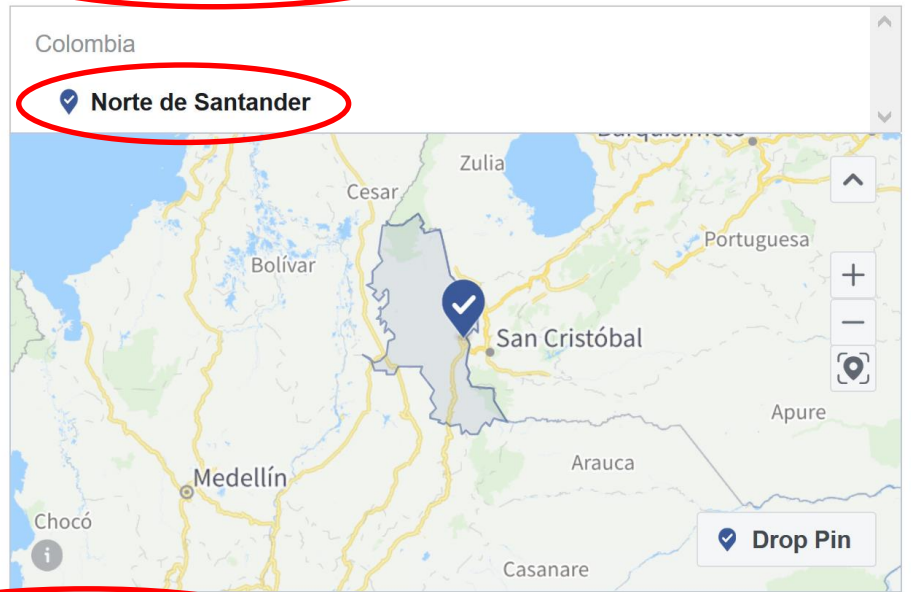
Text & Links

Languages

Tracking

Locations ⓘ People living in this location ▼

📍 Norte de Santander



Age ⓘ 13 ▼ - 24 ▼

Gender ⓘ All Men Women

Detailed Targeting ⓘ Include people who match ⓘ

Behaviors > Expats

Lived in Venezuela (Formerly Expats - Venezuela)

Audience Size



Your audience selection is fairly broad.

Potential Reach: 16,000 people ⓘ

Estimated Daily Results

Reach ⓘ  
4.9K - 7.7K

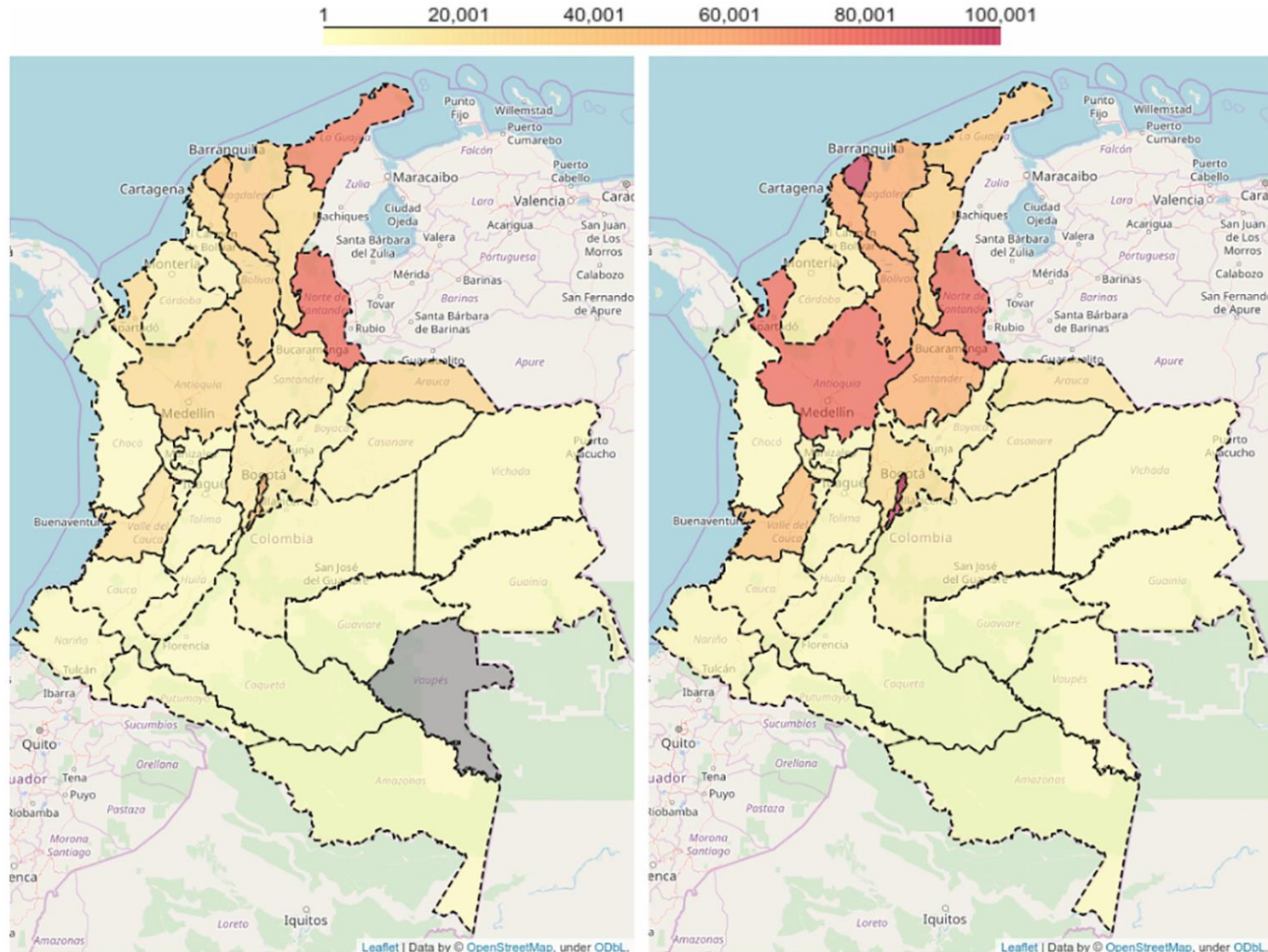
The accuracy of estimates is based on factors like past campaign data, the budget you

Size: 4,854,832  
Behaviors > Expats > Lived in Venezuela (Formerly Expats - Venezuela)

Description: People who used to live in Venezuela who now live abroad



# Validation w/ (Few) Available Data



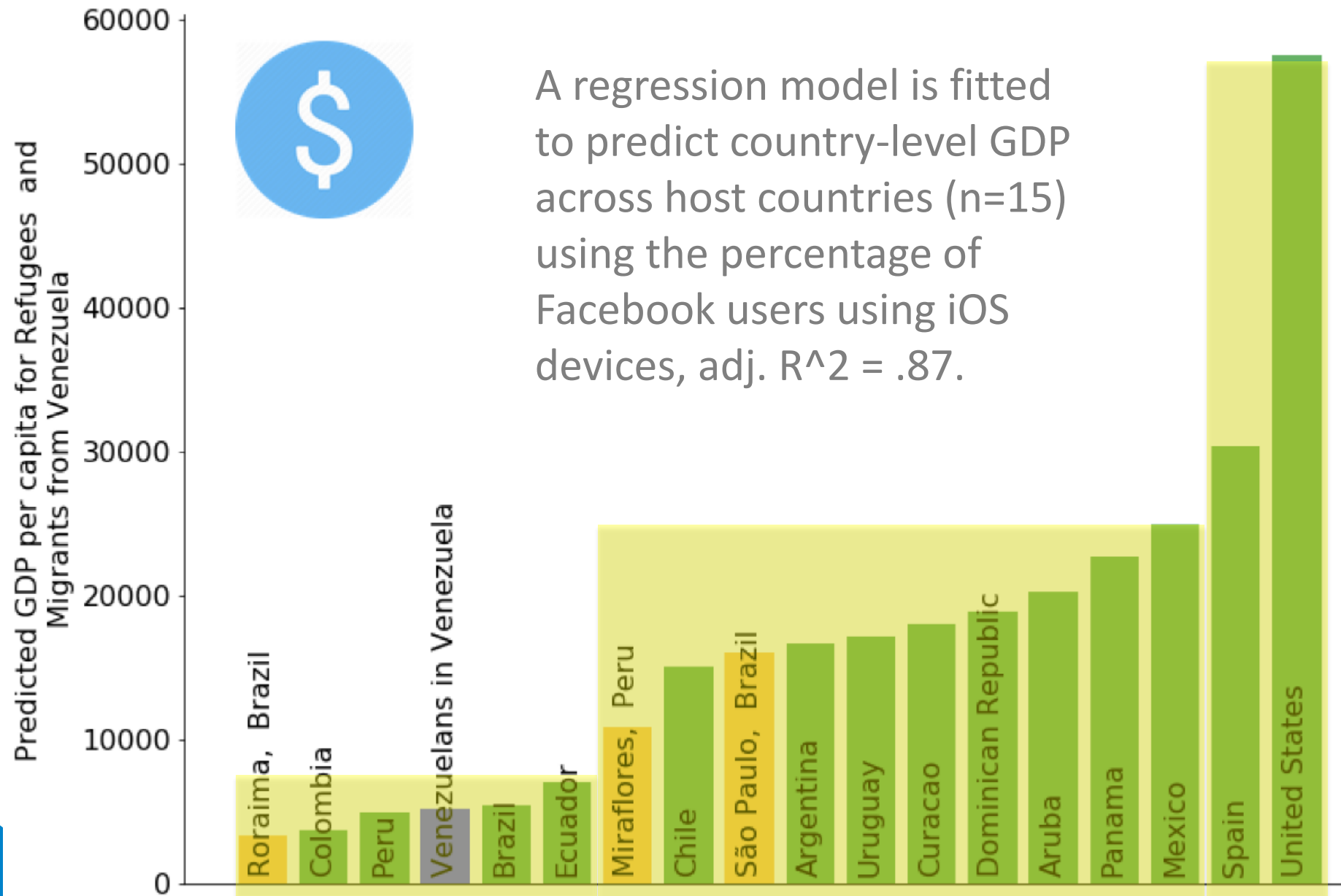
(a) RAMV - April 6 to June 8, 2018

(b) Facebook - June 8, 2018





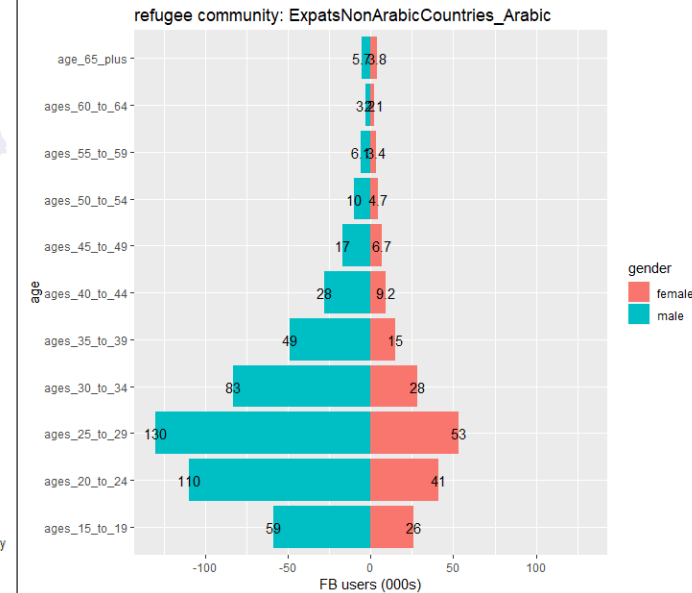
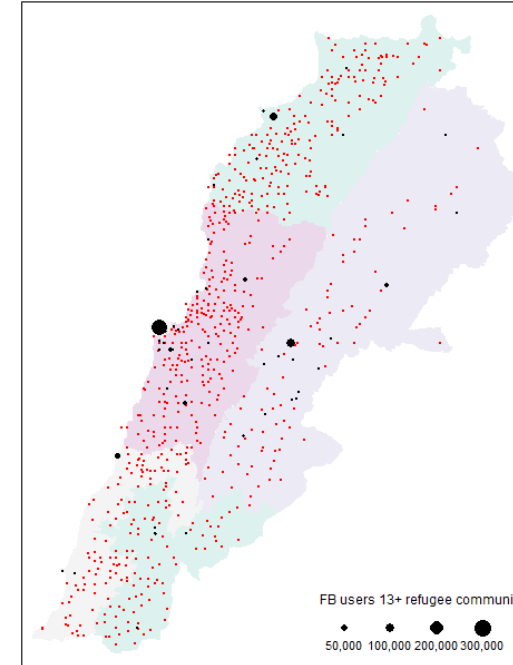
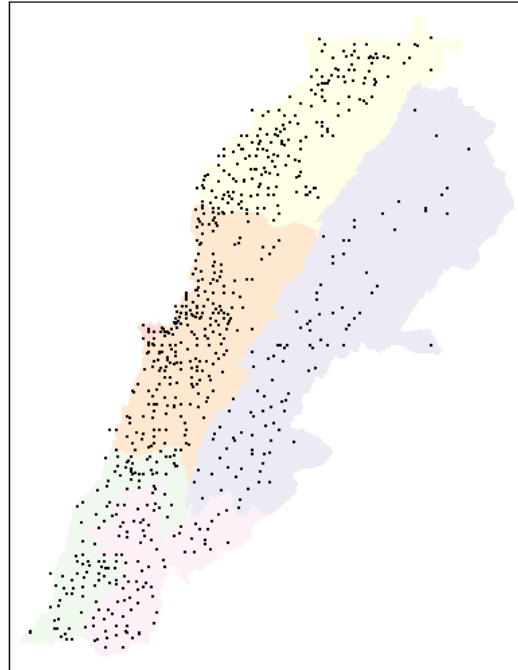
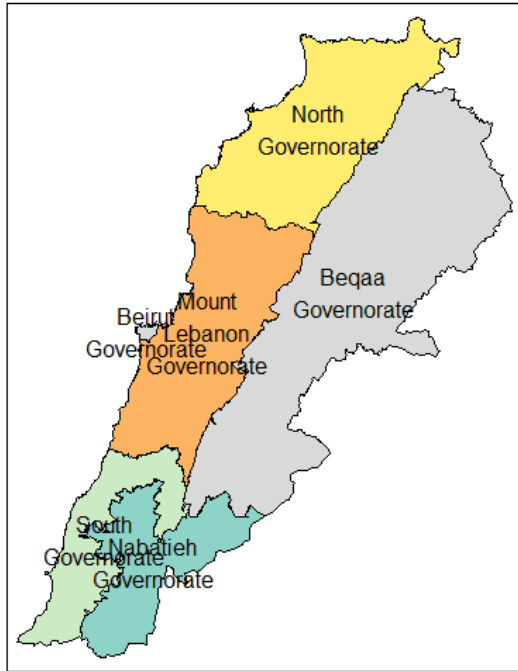
# Predicted Income Based on OS



# ASSESSMENT OF SYRIAN REFUGEES' SOCIO-ECONOMIC SITUATION IN LEBANON



# Collecting Facebook Audience Estimates



For 6 governorates and 770 cities

720k FB users “lives abroad” + AR

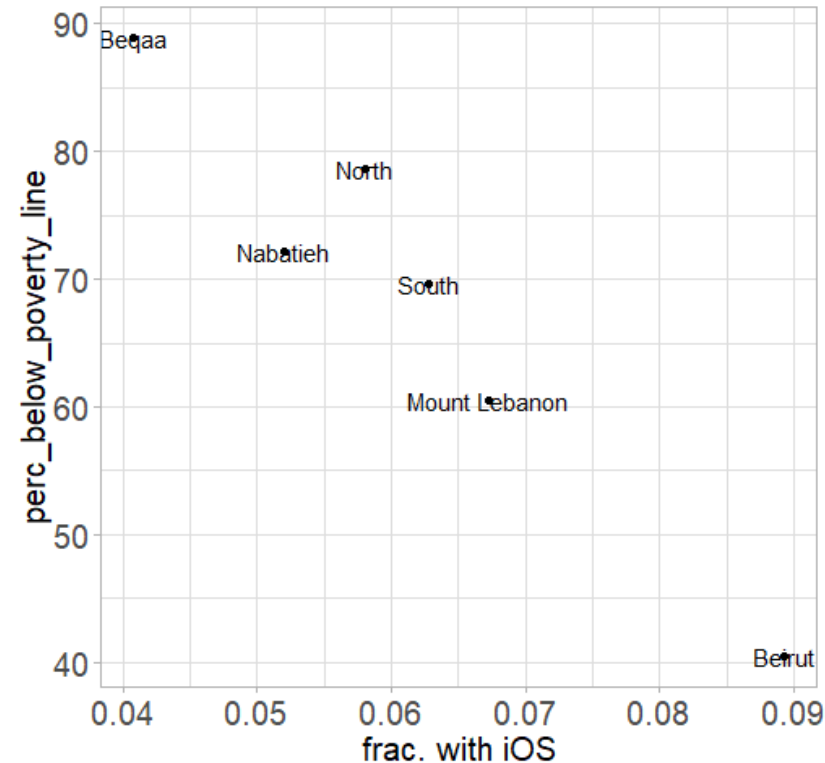
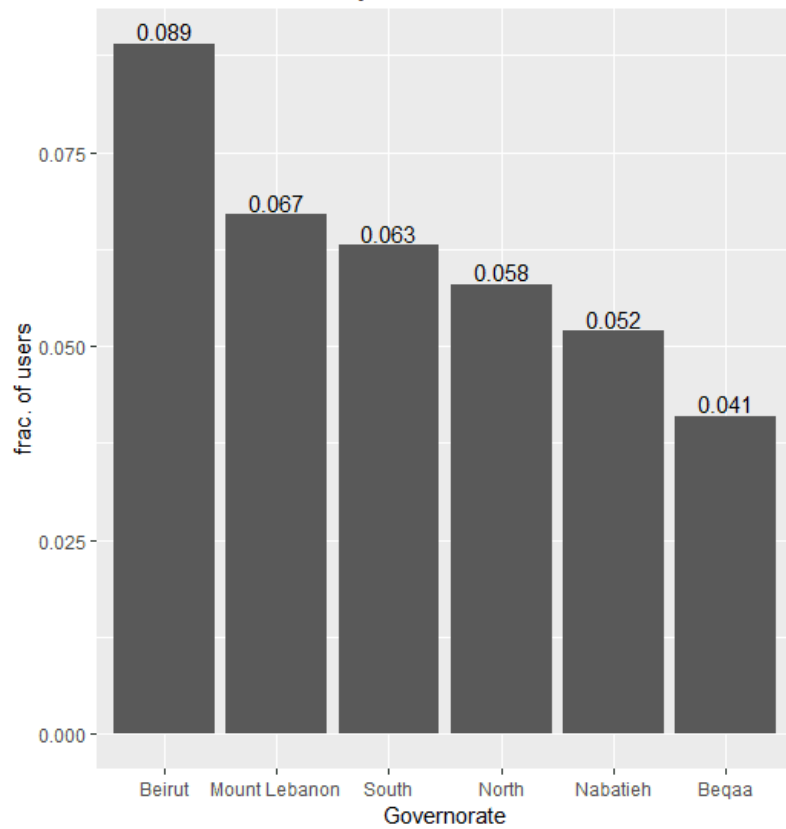
3.5M FB users overall

Only 157 cities with > 1000 FB users

Strong gender bias

# Is OS-Type Predictive of Poverty?

Users 13+ with iOS by Governorate



Predicting: % below poverty line

Model variable	CV performance (LOOCV)
	R <sup>2</sup>
% iOS device users	<b>0.895</b>
% high-end phones (iphones/galaxy) users	<b>0.678</b>



# CLOSING THOUGHTS



# Strengths and Opportunities

- + (Almost) Real-time data
- + Can be collected at no cost
- + Anonymous and aggregate data
- + Directly measures one type of asset ownership
- + Could be used for connectivity mapping (4G)
- + Can be disaggregated by age, gender, ...
- + Complementary strength to satellite imagery
- + Upwards development likely detectable

# Limitations and Challenges

- Black box for how attributes are inferred
- Cross-context comparison risky due to bias
- Usage patterns change over time
- No historic data available
- Sparsity constraints (< 1000 FB users)
- Not all countries of origin are supported
- Downwards development likely not detectable
- Risk of misuse

# Worth Exploring: Passive -> Active

- So far only “passive” use of advertising platform
  - Collect aggregate user data without users’ direct involvement
- Explore “active” use for targeted surveys
  - Show an advertisement message promoting a survey

Clearly not a representative sample, but quota sampling (age, gender, education level, ...) is possible. Facebook-users-vs.-non-users bias remains though, and people without internet access are excluded.

# Thanks!

[iweber@hbku.edu.qa](mailto:iweber@hbku.edu.qa)

<https://ingmarweber.de/publications/>