# Informing Preference Heterogeneity with Stated Preferences or Passive Geolocation

**ART FORUM 2019**

Paul Johnson (Dynata), Marc Dotson (BYU), Adam Smith (UCL)

June 13, 2019

# Passive Geolocation Data

# Passive and Stated Location Data Both Have Strengths

## Passive Data

- Less Recall Error
- Great at Answering the What
- Large Volume of Data
- Less Expensive per Data Point
- Aggregated Trends/Segments

## Stated Data

- Less Measurement Error
- Great at Answering the Why
- Better Structured Data
- Purposefully Collected
- Individual Level Data

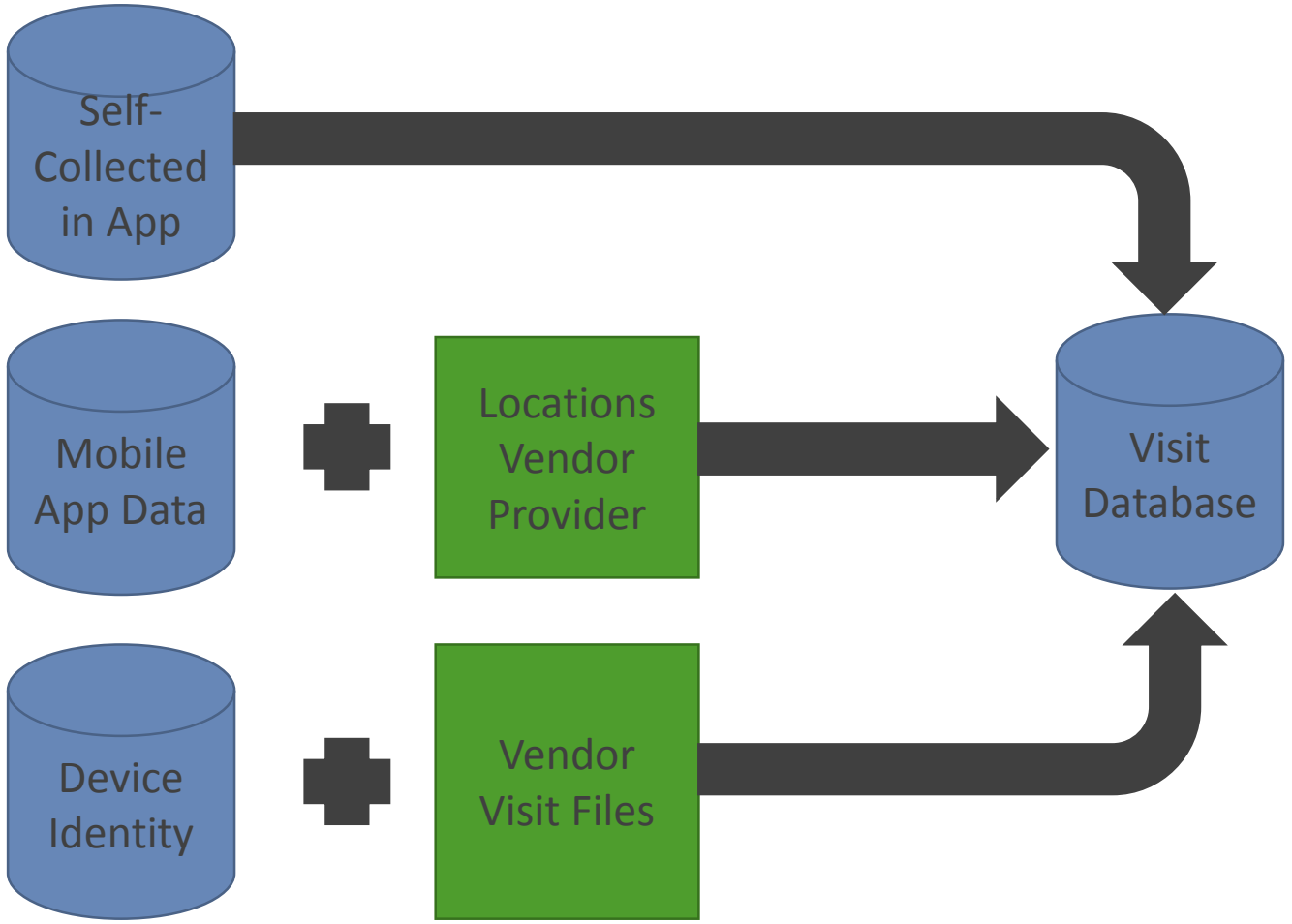
**Can we combine the two together to predict better?**

# Geolocation Data Not All Created Equal

**Requires**
- Locations Predefined
- Panelist Downloaded App
- Panelist Compliance with Location Monitoring

**Requires**
- Panelist Downloaded App
- Happen to be Measured at Location
- Vendor Contracts

**Requires**
- Permissioned Panel
- Vendor Contracts



# Examples of Polygon Testing

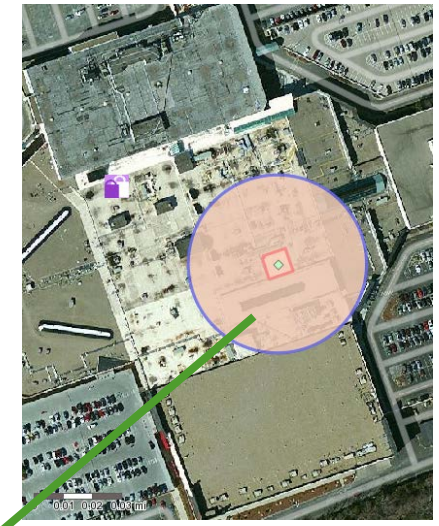
verizon

Stand Alone Location

2009 North
Main Street
Crossville, TN

Generally okay but gets passing road

Inside a Mall



Westfield Mall
5065 Main Street
Trumbull, CT

False positives in half the mall

Shared Parking Lot

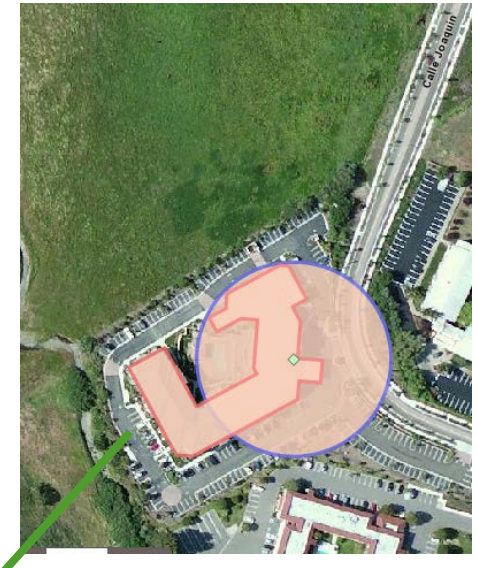

1130 Levis
Commons Blvd
Perrysburg, OH

Incorrectly Captures Highway Nearby

dynata

# Examples of Polygon Testing

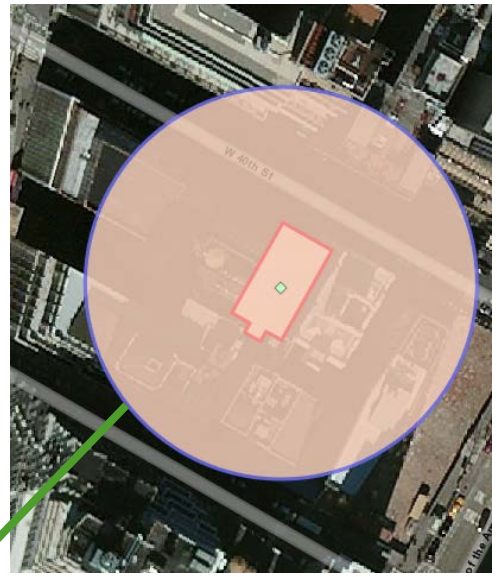
Stand Alone Location



1605 Calle Joaquin Road
San Luis Obispo, CA

Doesn't Include Half the Hotel

Urban Area



114 West 40th Street,
NYC, NY

Lots of False Positives

Resort Area

99751 Overseas Hwy
Key Largo, FL

False Positives in Road Nearby

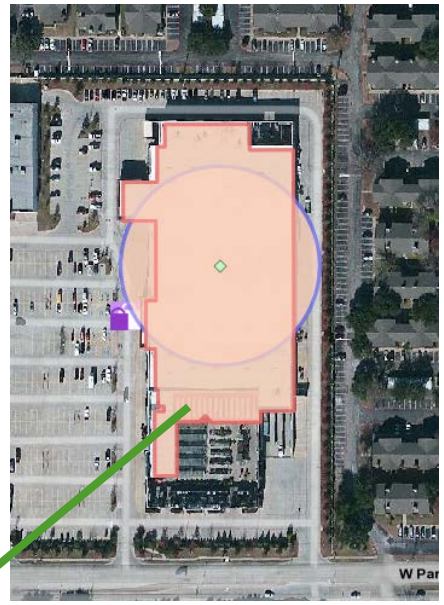
# Examples of Polygon Testing

Stand Alone Location

1551 Froom Ranch Road
San Luis Obispo, CA

Doesn't Even Cover All Entrances

Suburban Area



1801 West Parker Road
Plano, TX

Significant Amount Not Covered

Near Highway

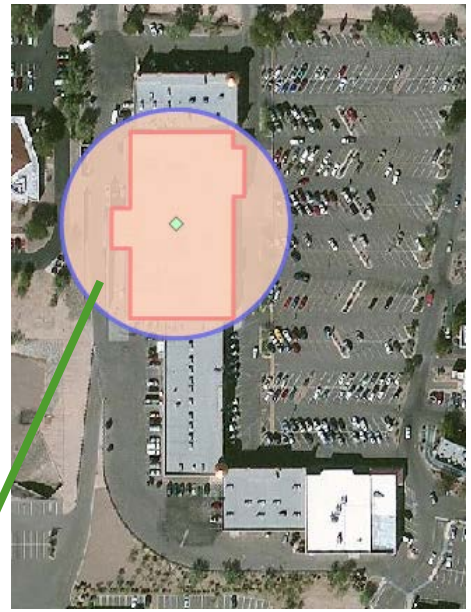
117 Main Street
Derby, CT

Reasonable but May Miss Some Visits

# Examples of Polygon Testing

Walmart

Stand Alone Location



7001 Concourse Parkway
Douglasville, GA

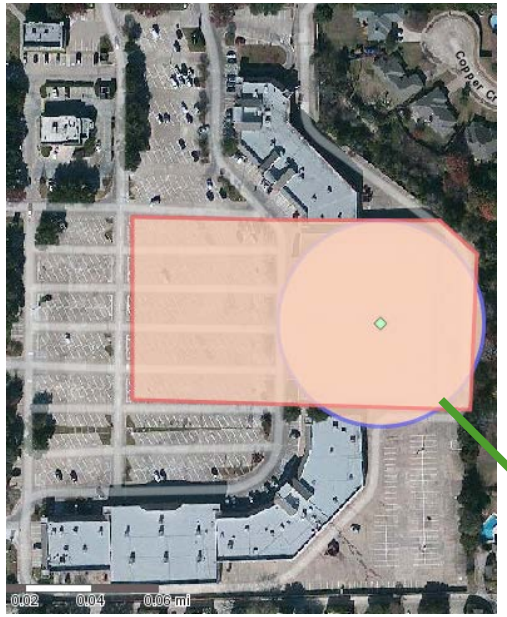
Incorrectly Captures Back Road

Strip Mall Example 1

7951 North Oracle Road
Oro Valley, AZ

Doesn't Capture Full Store

Strip Mall Example 2



3100 Custer Road
Plano, TX

Polygon Drawn Incorrectly

# Three Rounds of Testing on Visit Vendors

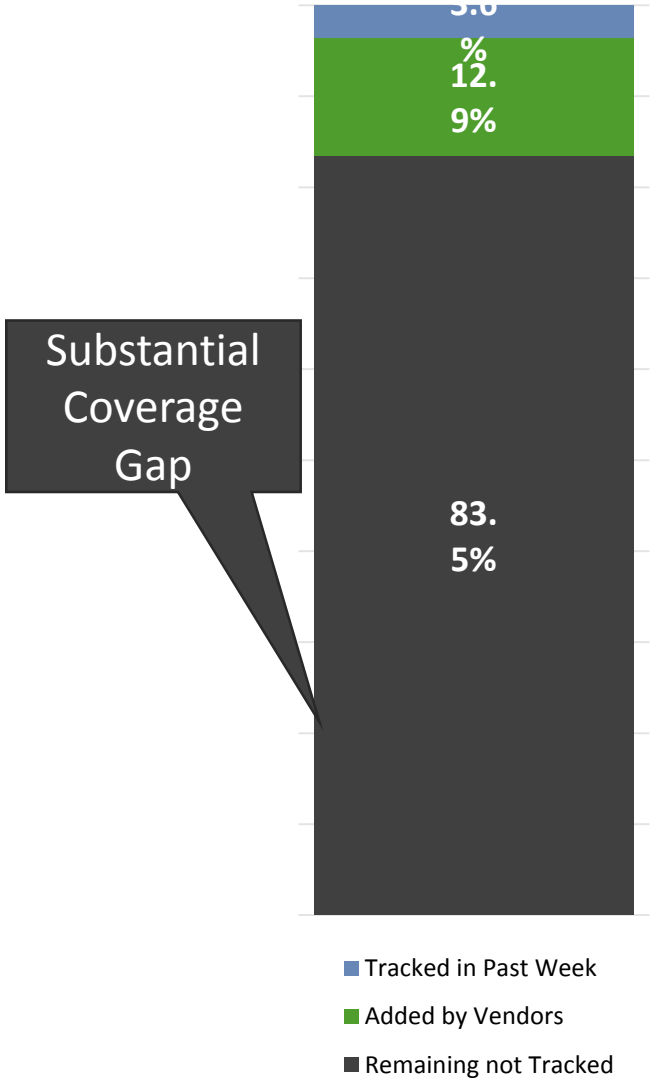
1. How many people match? (Breath)

2. How close does the raw data match? (Accuracy)

3. How many more completed surveys can I do? (Depth)

# Vendors do Add Substantial Volume, but Have Gaps

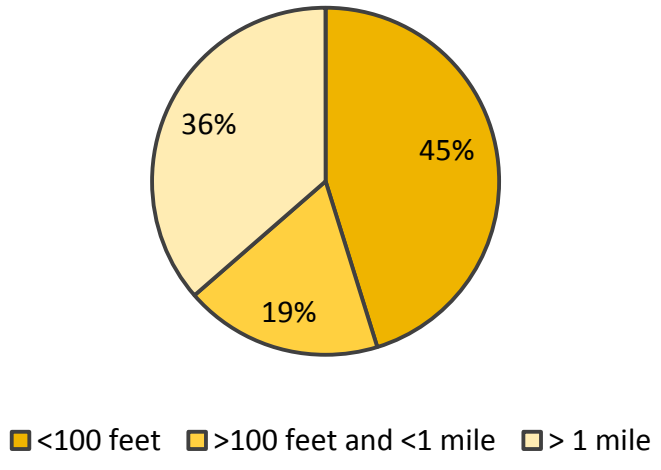
| Vendor | Devices | Match Rate | Matches | % Unique |
|---|---|---|---|---|
| **Vendor 1** | 18,235,193 | 0.23% | 42,811 | 9.8% |
| **Vendor 2** | 224,922,908 | 0.06% | 140,171 | 65.4% |
| **Vendor 3** | 100,700,000 | 0.14% | 140,480 | 30.1% |
| **Vendor 4** | 131,105,306 | 0.14% | 185,592 | 35.5% |
| **Vendor 5** | 6,000,002 | 0.27% | 16,135 | 46.8% |



# Definite Differences in Data Collected from Vendors

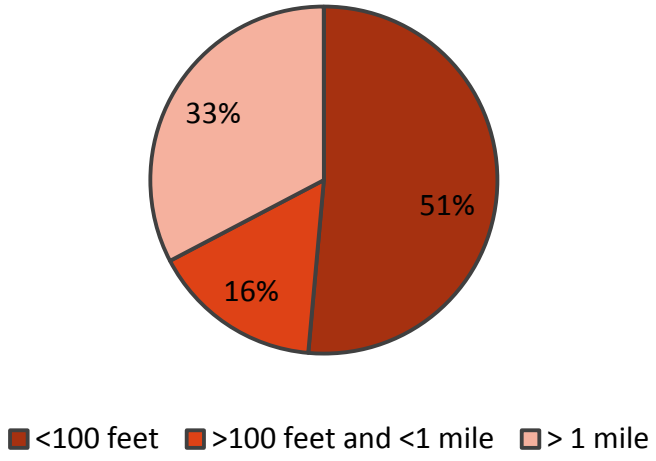
## Vendor 1

- 1 week of data
- 50K MAIDS
- Timestamp Matches within 1 minute
- 2.4 million matches (300K per day)



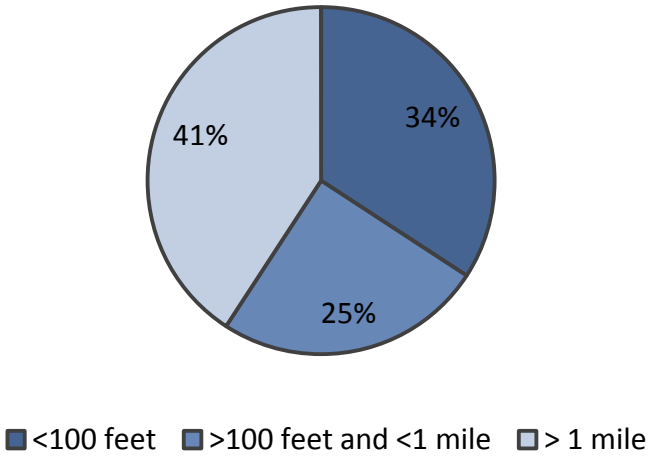

## Vendor 2

- 1 month of data
- 4K MAIDS
- Timestamp Matches within 1 minute
- 188K matches (6K a day)

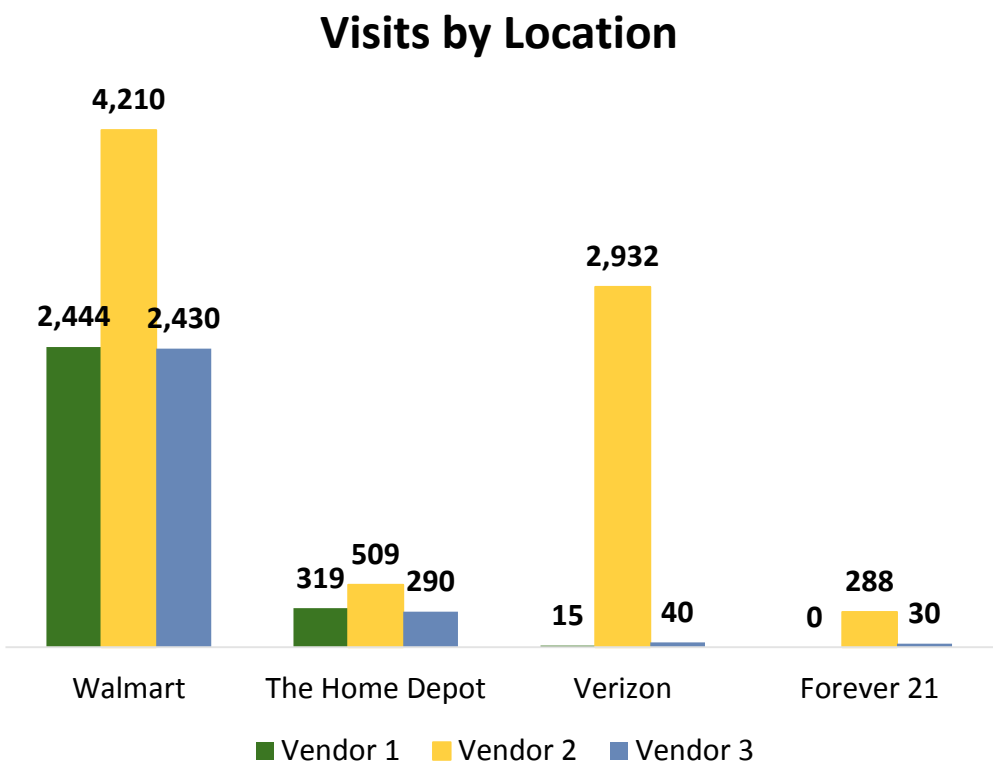


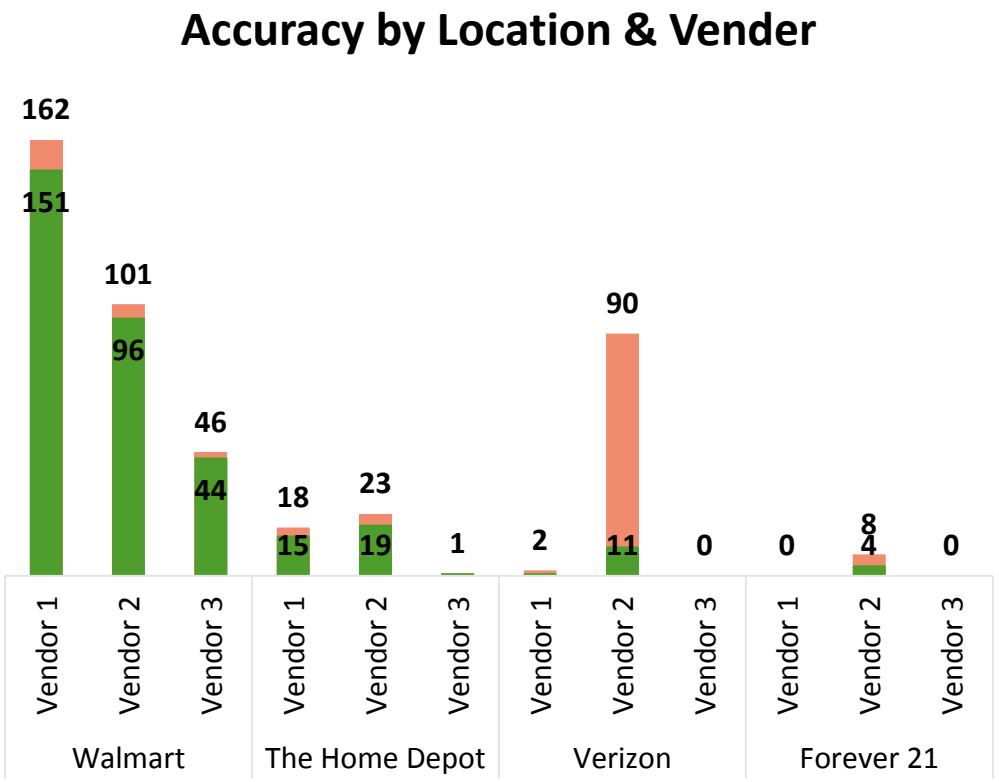

## Vendor 3

- 2 days of data
- 50K MAIDS
- Timestamp Matches within 1 minute
- 280K matches (140K per day)



# ‘Easy’ Locations High Confirmation Rate
# ‘Hard’ Locations Lack Data or Inaccurate



# Summary of Geolocation Data

## Tips for Passive Data Usage

1. Be Realistic of Competing Interests for Feasibility vs Accuracy
2. Don’t Ignore Measurement Error Even After Vendor Cleaning
3. Confirm with Survey Data When Possible
4. Make Sure Any Individual Linking is Legally Permissioned
5. Have a Predefined Use for the Data

# Our Study

# Passive and Stated Location Data Both Have Strengths

## Passive Data

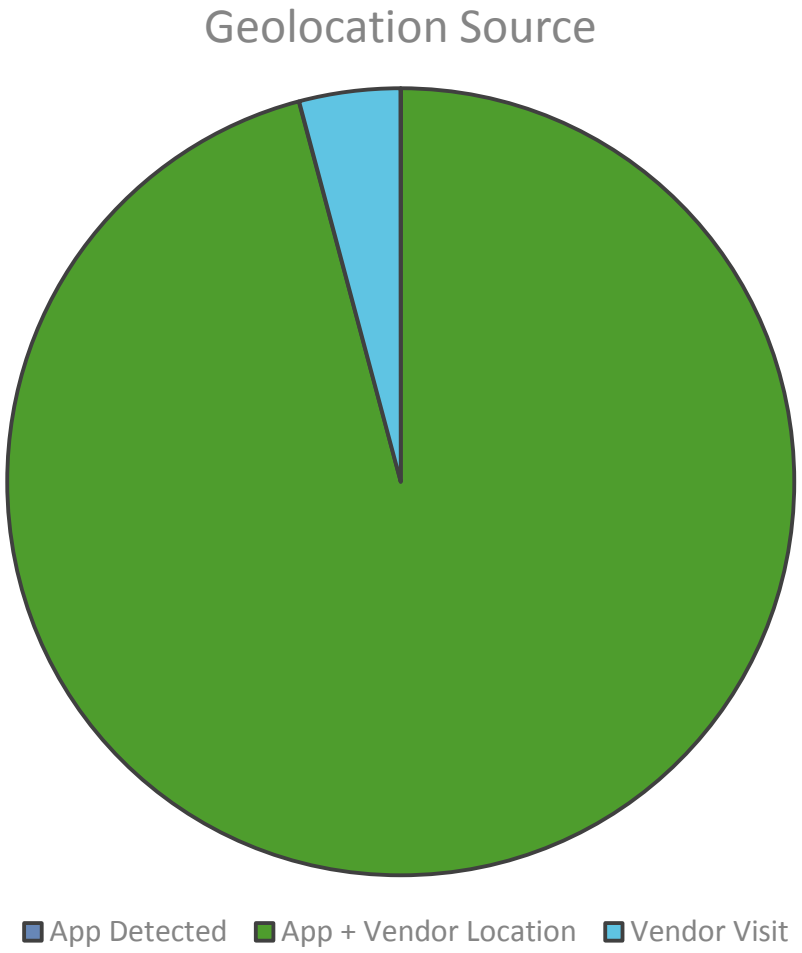
- 784 respondents
- Visits to Branded Dealerships
- Past 6 months
- 10,793 branded locations

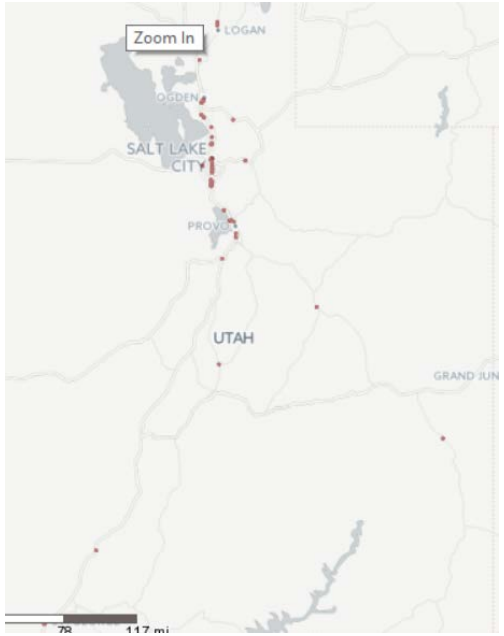
## Stated Data

- 784 respondents
- 12 conjoint tasks, 8 attributes
- Stated Brand, Price, Car Type Preference
- Demographics

**All tied at the individual level!**

# Checking Passive Geolocation Data

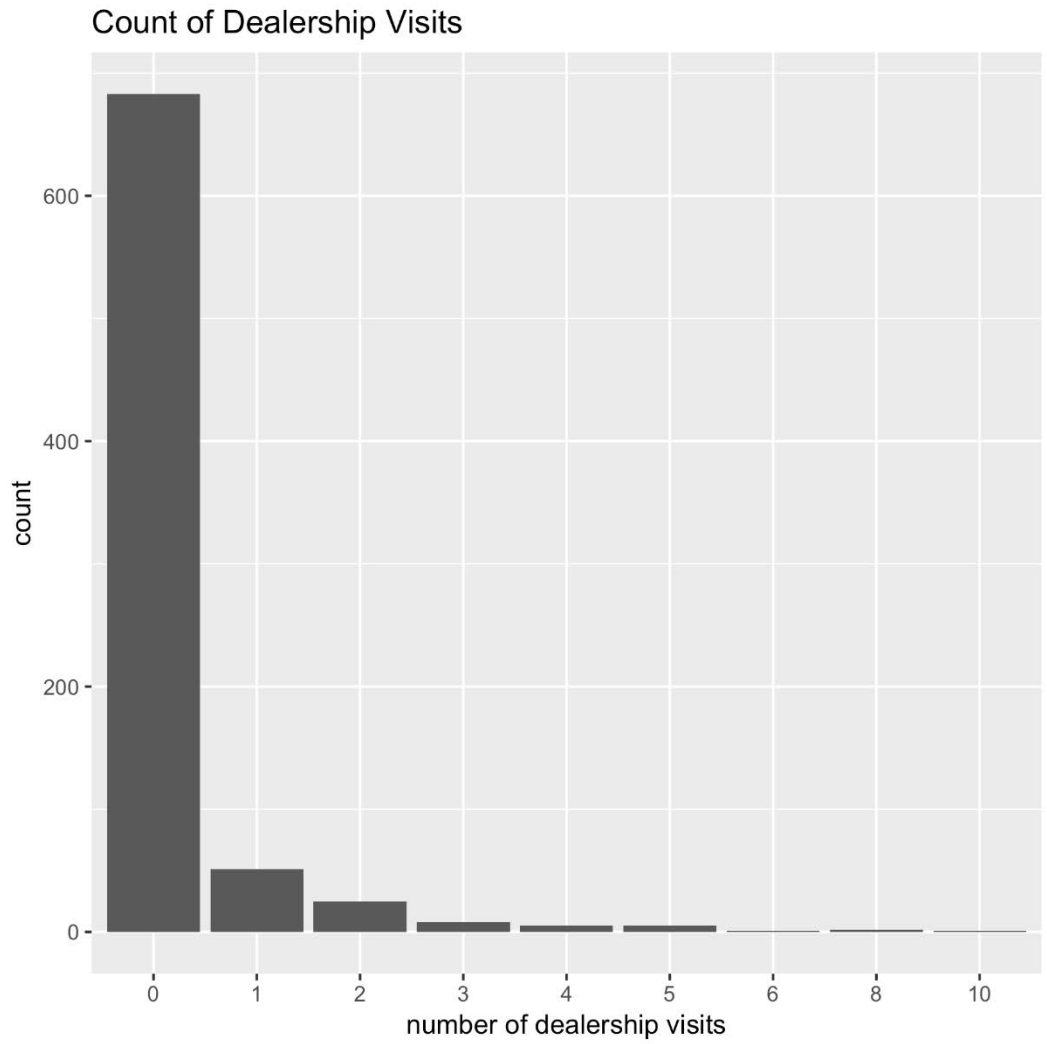


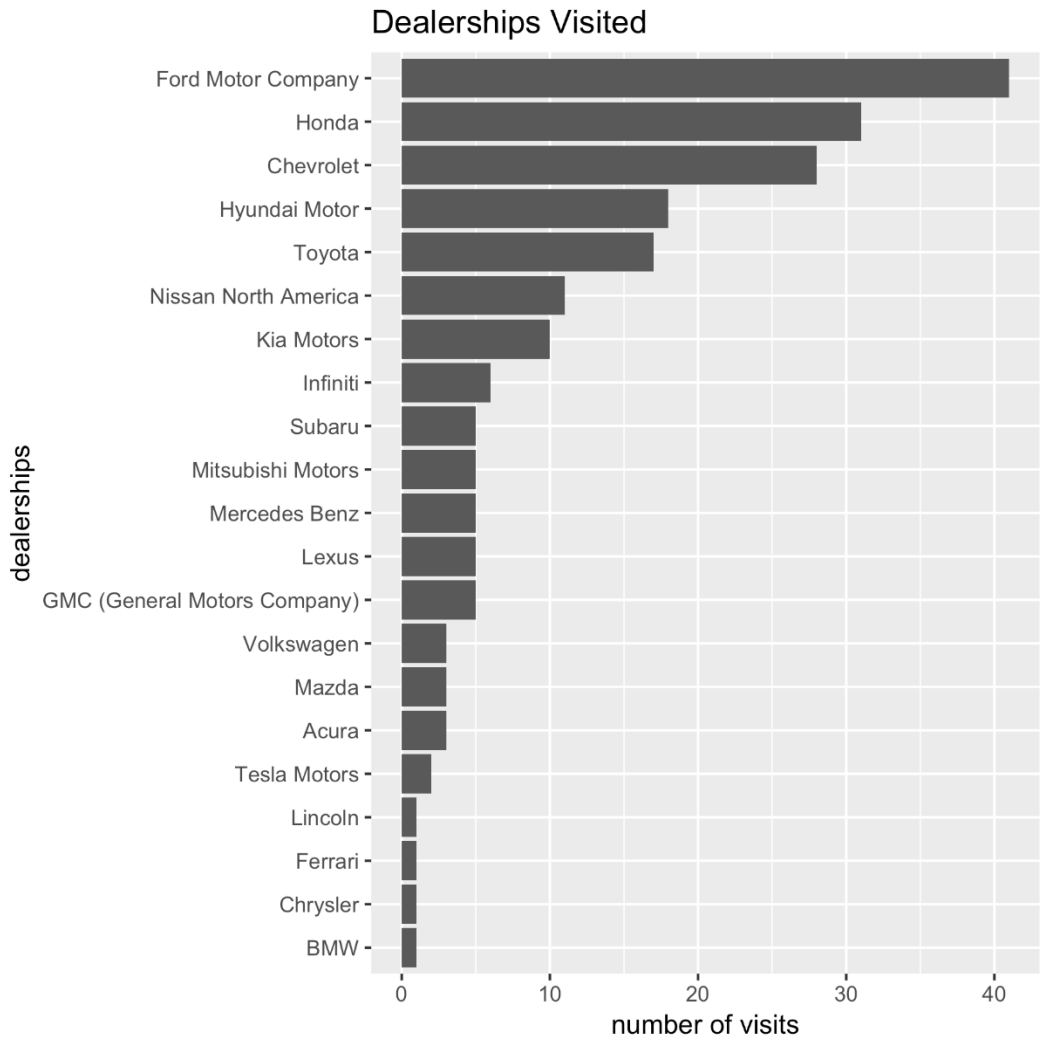






# Preliminary Results

# Dealership Visits



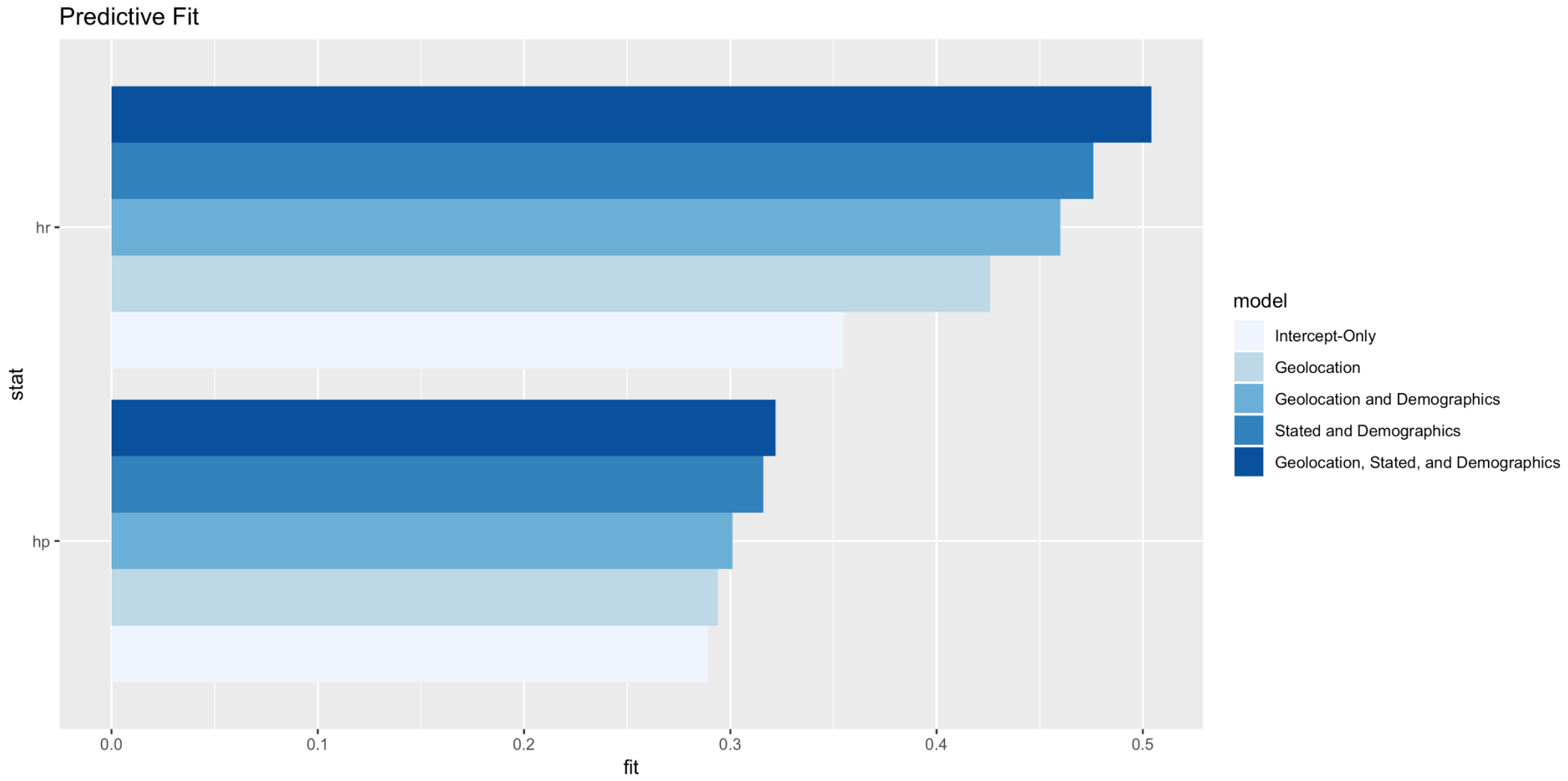



# Model Fit Table

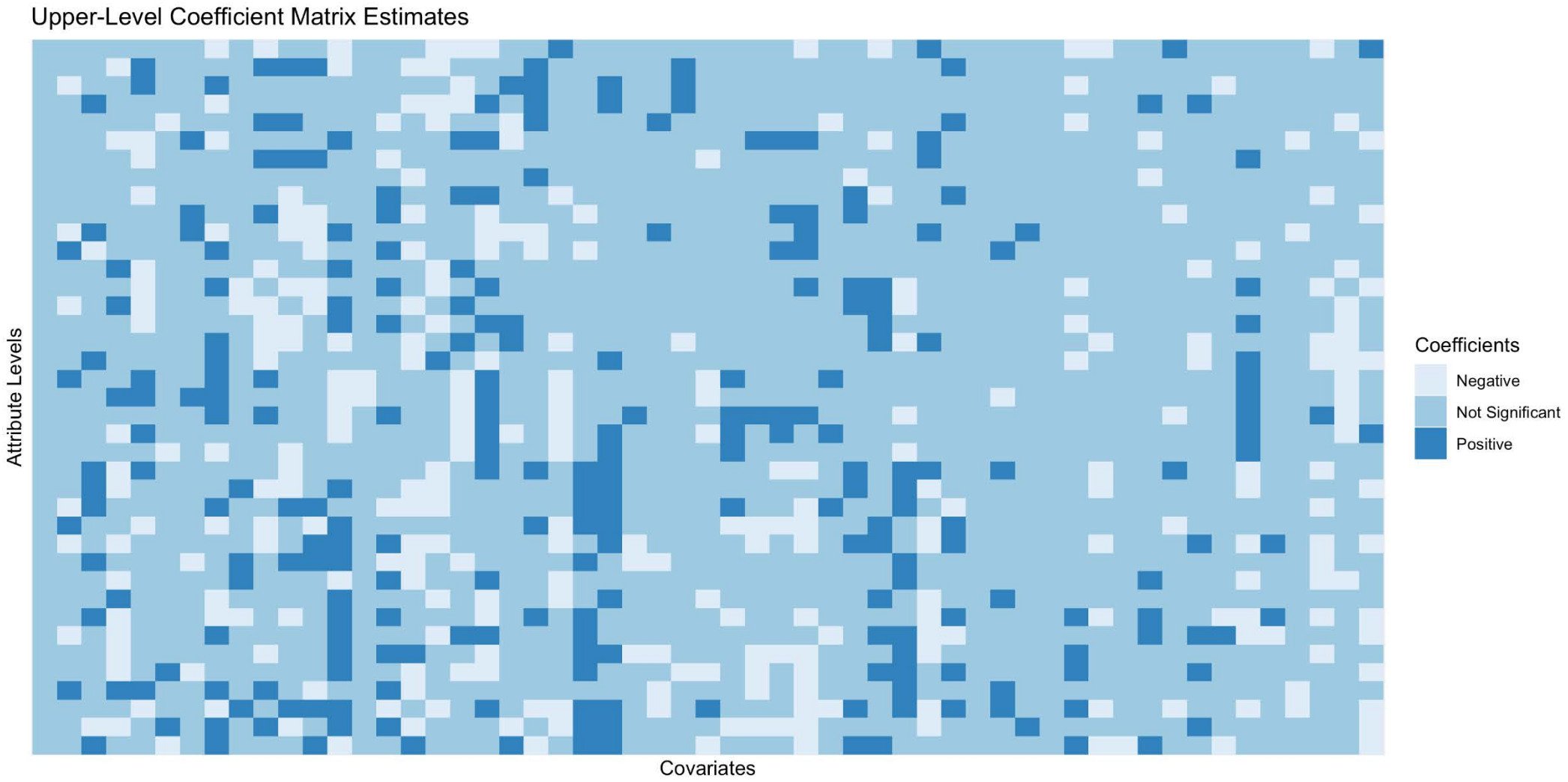
## Using Geolocation Improves Fit

| | lmd | dic | hr | hp |
|---|---|---|---|---|
| Intercept-Only | -4270 | 16061 | 0.355 | 0.289 |
| Geolocation | -4254 | 15933 | 0.426 | 0.294 |
| Geolocation and Demographics | -4124 | 15686 | 0.460 | 0.301 |
| Stated and Demographics | -4051 | 14777 | 0.476 | 0.316 |
| Geolocation, Stated, and Demographics | -3941 | 14629 | 0.504 | 0.322 |

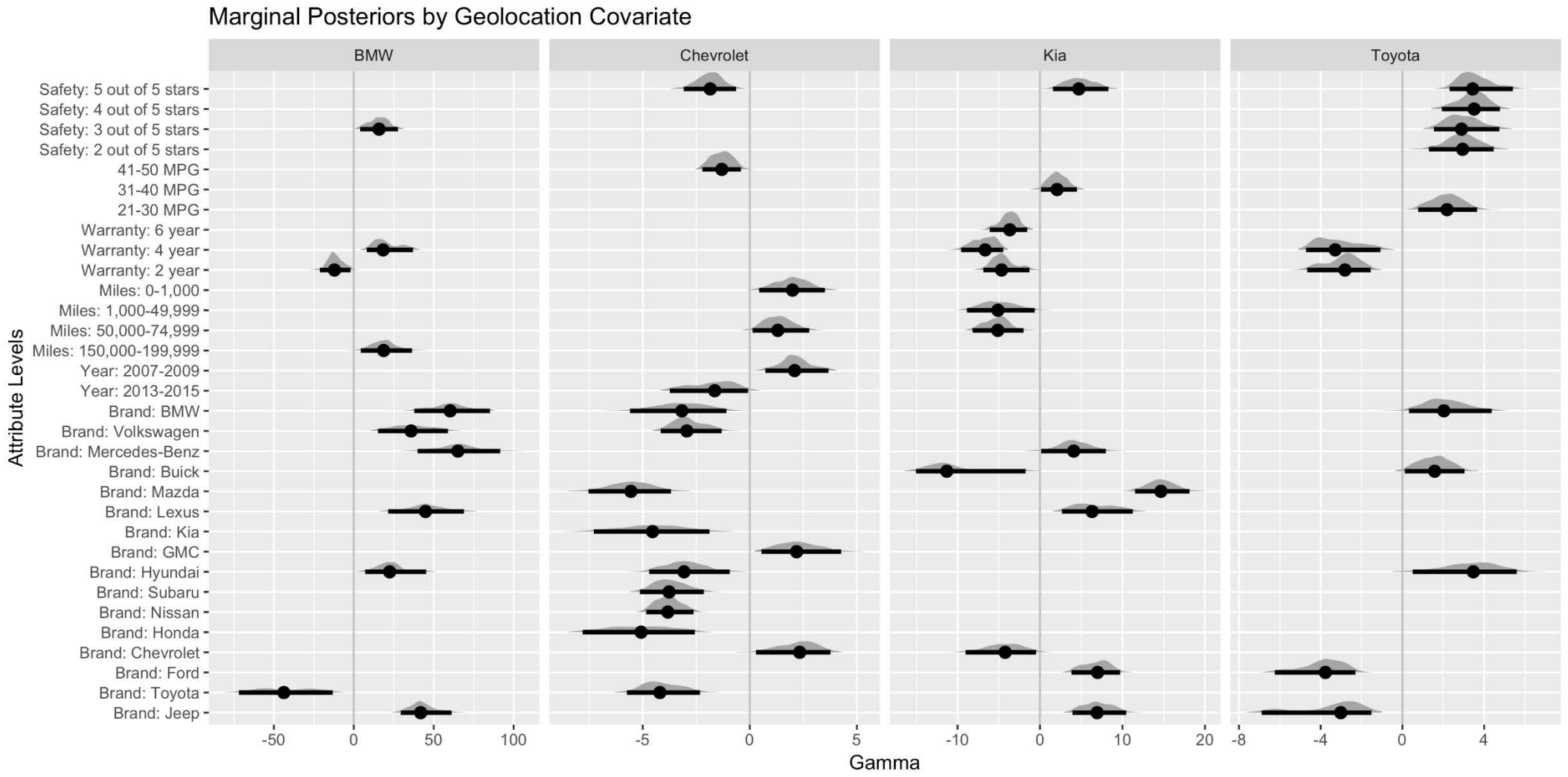
# Uniform Improvement in Predictive Fit



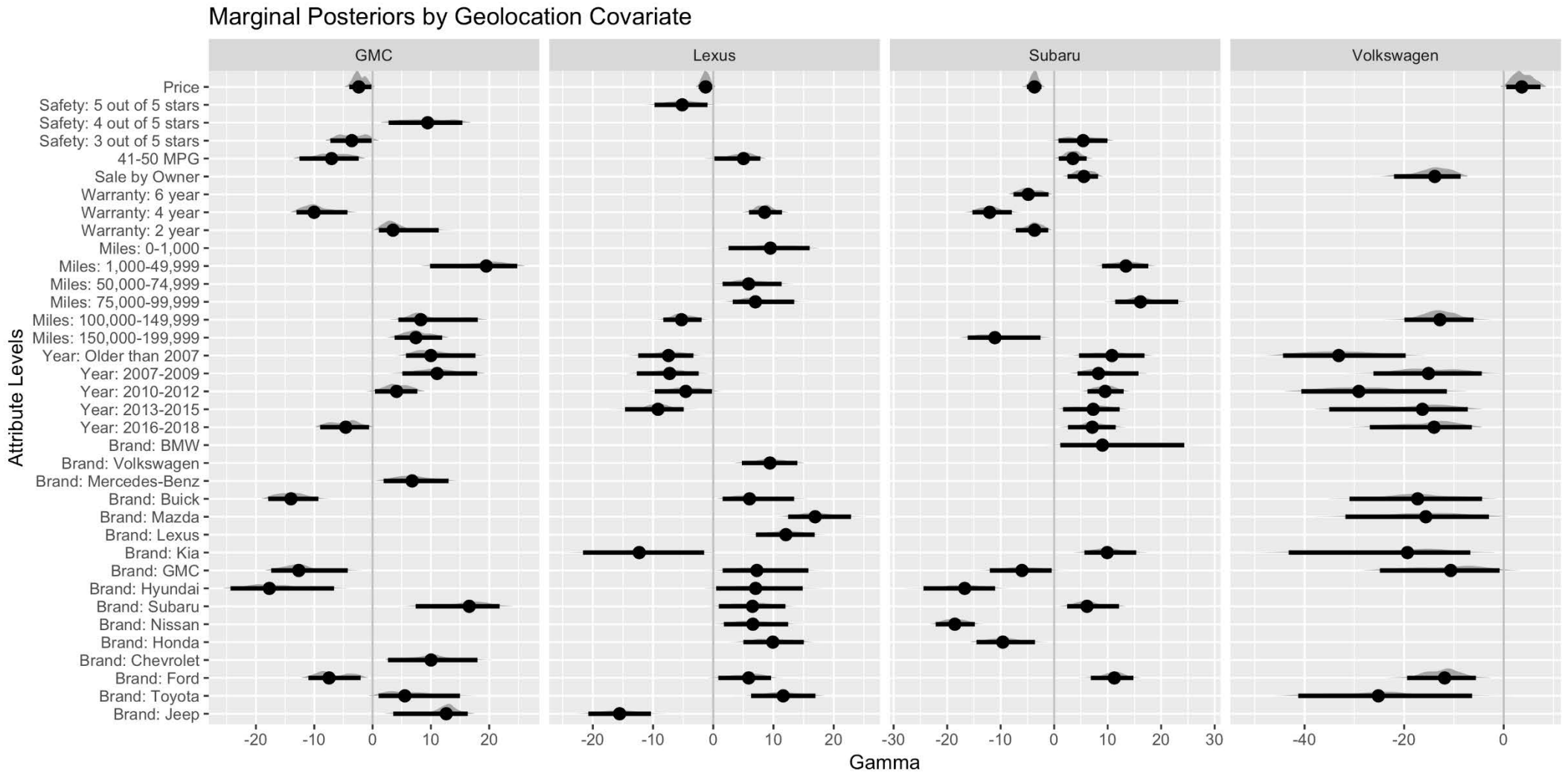
# Where Do the Covariates Matter?



# Geolocation Covariates and Attribute Levels (Part 1)



# Geolocation Covariates and Attribute Levels (Part 2)



# Next Steps

## Validation and Modeling Heterogeneity

- Instead of a hold-out sample, we'll pull actual data on actual purchases in six months.
- In addition to covariates, there's also a question of the heterogeneity model.
- Currently exploring various approaches which reduce coefficient matrix dimensionality:
  - mixed membership (Dotson, Buschken, Allenby 2019)
  - sparsity-inducing priors
  - tree models