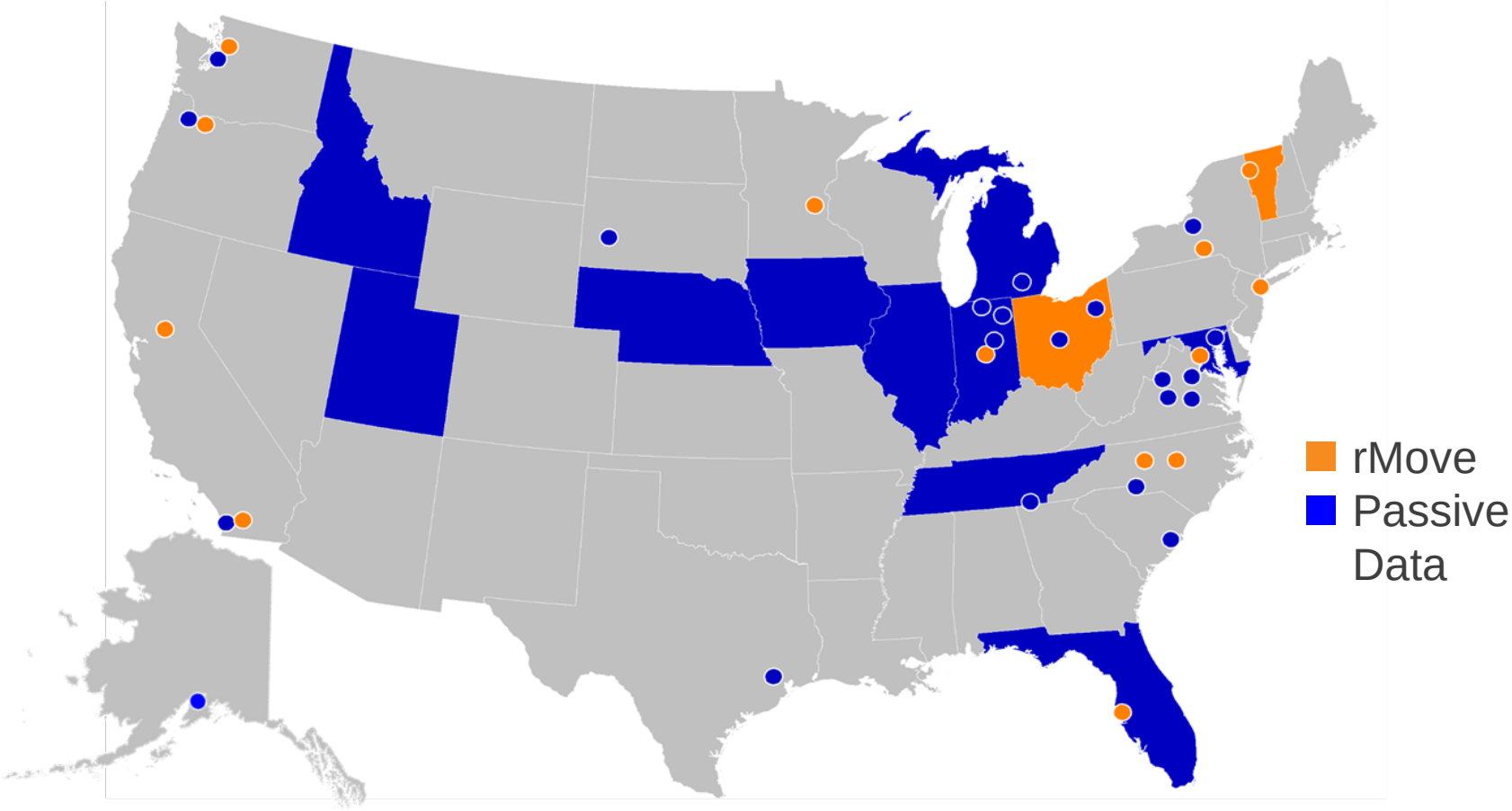
RSG
the science of insight

Overview of Methods for Validation and Expansion of Passive Origin-Destination Data

Vince Bernardin, PhD

June 26, 2018

RSG's Mobile Data Experience



Over 40 projects in more than 20 states

Types of Passive OD Data

Cellular tower signaling

LBS
(Location-Based Services)

GPS
(Global Positioning Systems)

WiFi beacons

Summary Comparison of Data Types in Q4 2017

	CELLULAR	LBS	GPS	
Description				
Universe	All Travel	All Travel	Trucks	Private Autos
Time Periods	Average Weekday or Average Weekend or Individual Day of Week; multi-hour periods within the day	Average Weekday or Average Weekend or Individual Day of Week; multi-hour periods within the day	Generally customizable down to individual hours of the day; effort to get multiple time periods may vary significantly by vendor	
OD Demand Types	Aggregate Trip ODs	Aggregate Trip ODs or disaggregate traces with restrictions	Aggregate Trip ODs; sometimes disaggregate traces also available with restricted use	
Precision and Coverage				
Locational Precision	> 100 m often ~ 200 - 2000 m	10-100 m often ~ 50 m	1 - 10 m	1 - 10 m
Sample Penetration	6-10%	5-8%	9-12%	~0.5%
Data Collection Time Period	Typically 1 month	One or more months	1 month - 2 years depending on provider & pricing	
Coverage Issues	Poor coverage in some (mostly rural) areas		"Urban canyon" effects	
Representativeness & Expansion				
Trip Length / Duration Bias	Confirmed	Confirmed	Confirmed	Confirmed
Included / Default Expansion	Residence market share based; generally requires adjusted to counts	None / residence based; generally requires adjusment for biases	None / Single count-based factor; generally requires adjusment for biases	
Additional Processing Required	Intermediate	Substantial to Limited depending on provider	Substantial to Limited depending on provider	
Segmentation & Applications				
Number of Zones	Limited by pricing and locational precision	Depends on pricing scheme	Relatively unlimited in most pricing schemes	
Select Link / Corridor Analysis	Generally indirect only	Indirect only currently but a subset may support direct in the future	Limited or Unlimted direct depending on provider, or indirect	
Filtering of Intermediate Stops on Long Trips	Premium option	Premium option	Depending on provider may be possible as a post-process	
Residency Information	Premium option	Premium option	Not available due to ID persistence limitations	
Trip Purpose	Premium option for imputed purposes	Premium option for imputed purposes	Not available due to ID persistence limitations	

The Power of Passive Data

The Power of Big Data

TN STATEWIDE DATA

- Combined household survey
  - NHTS + 4 MPOs
  - 10,344 households
- AirSage and ATRI datasets

- Trip Table (OD pairs)

– Total:	12,744,900	
– Survey:	39,782	0.3%
– AirSage:	3,355,539	26.3%

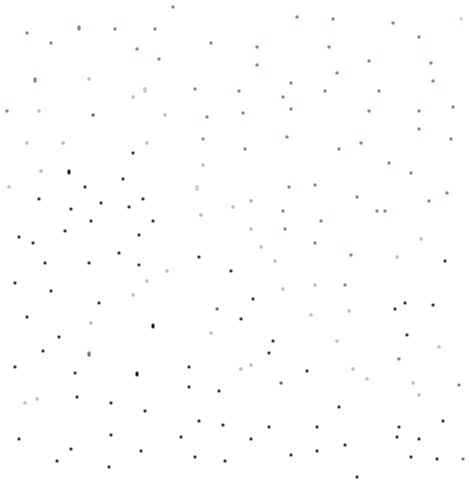
CHARLESTON, SC DATA

- 2016 NHTS
  - 1,014 households
- AirSage and ATRI datasets

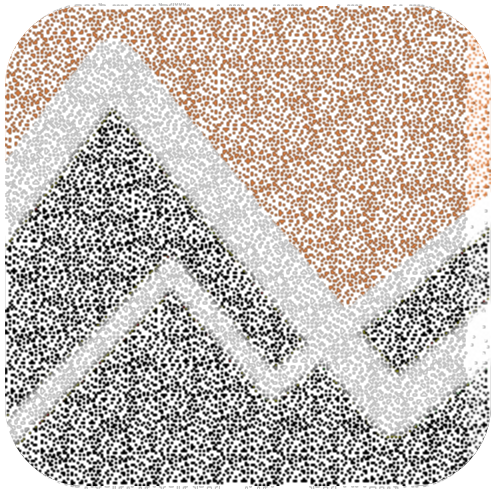
- Trip Table (OD pairs)

– Total:	760,384	
– Survey:	5,006	0.7%
– AirSage:	253,304	33.0%

Can you recognize the pattern based on <2%?



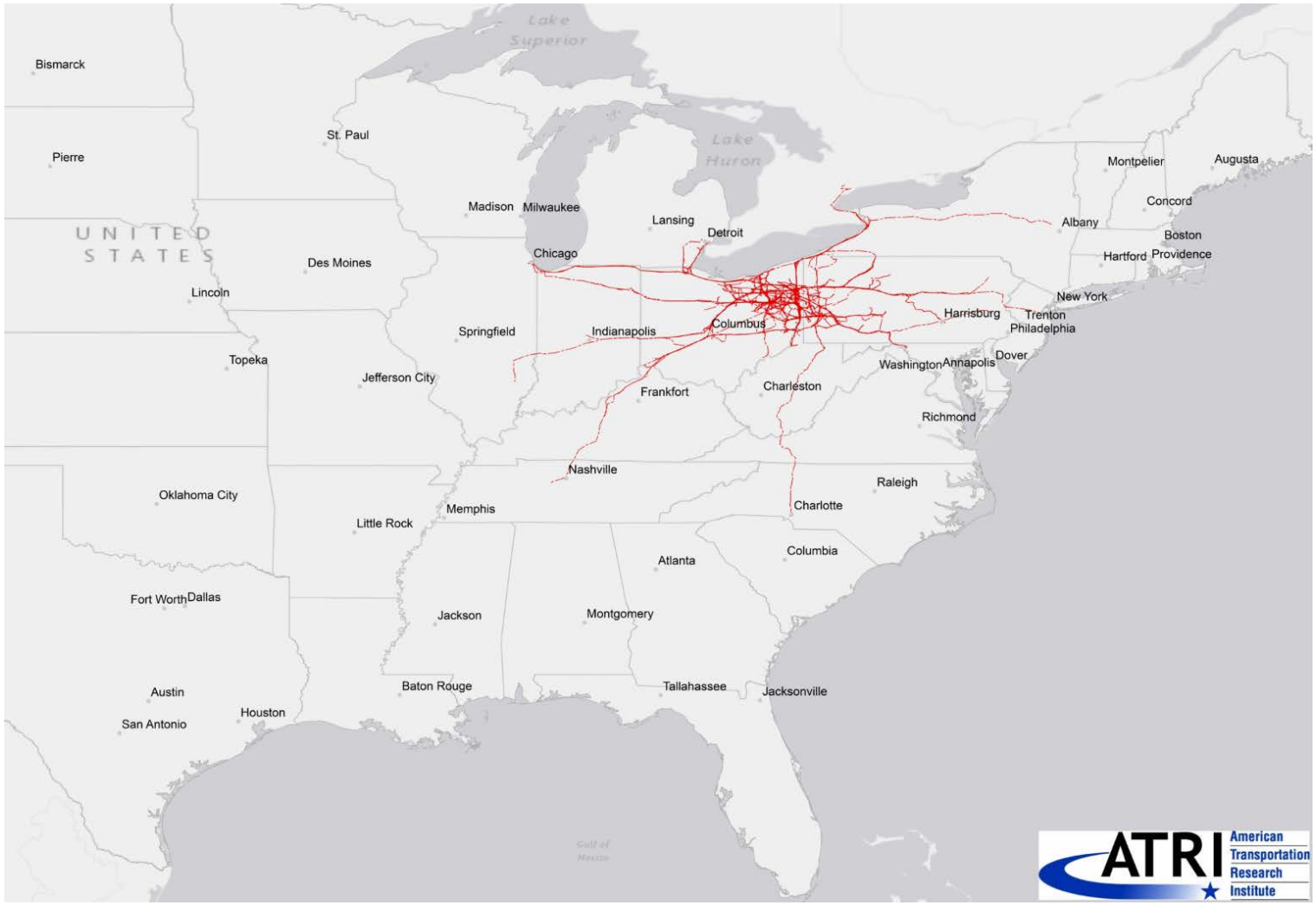
How about based on >25%?



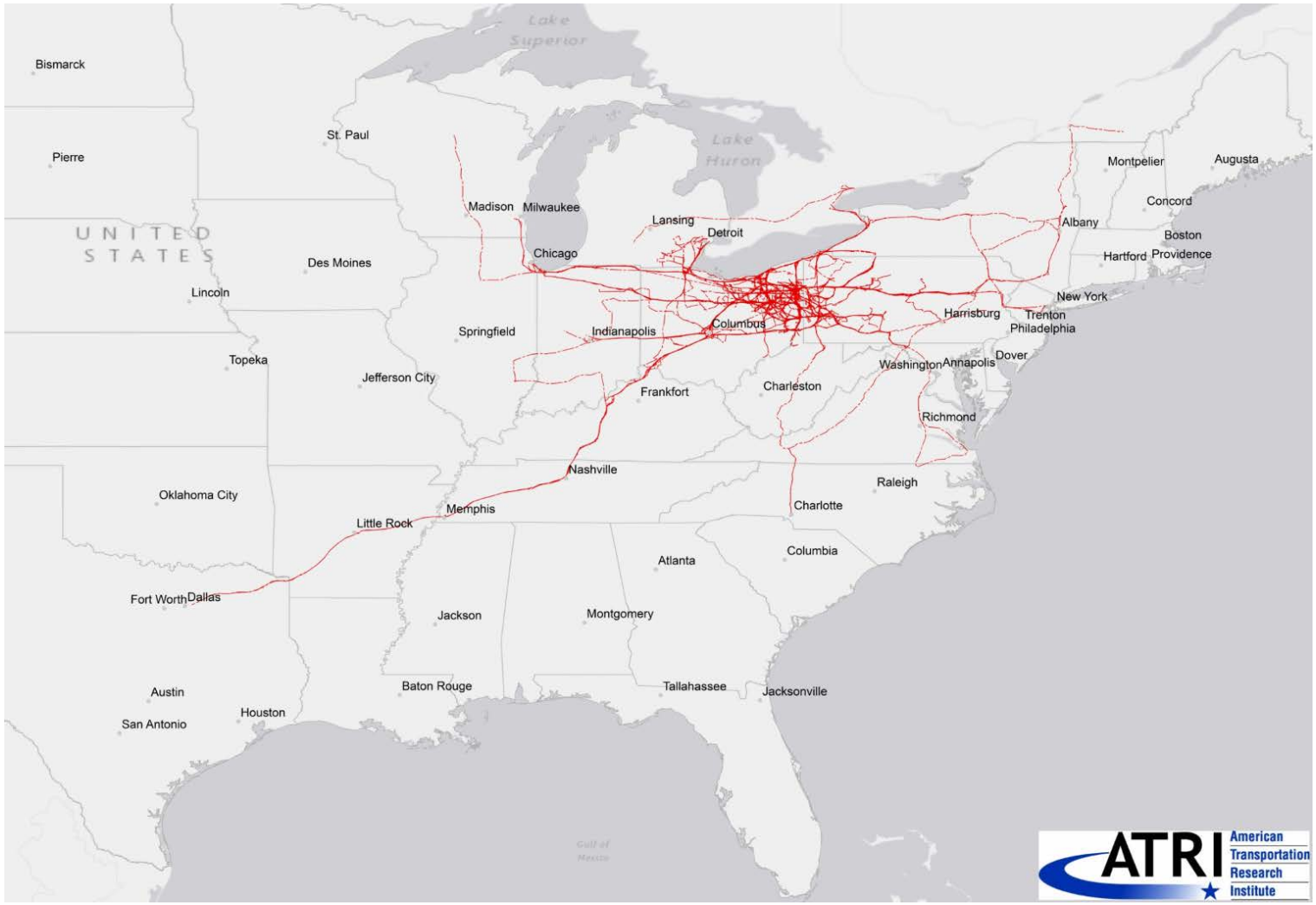
Big Data allows us to see the big picture.

US 30 Study Area

Trucks Using the US 30 Corridor – After 1 Day



Trucks Using the US 30 Corridor – After 2 Days



Trucks Using the US 30 Corridor – After 5 Days

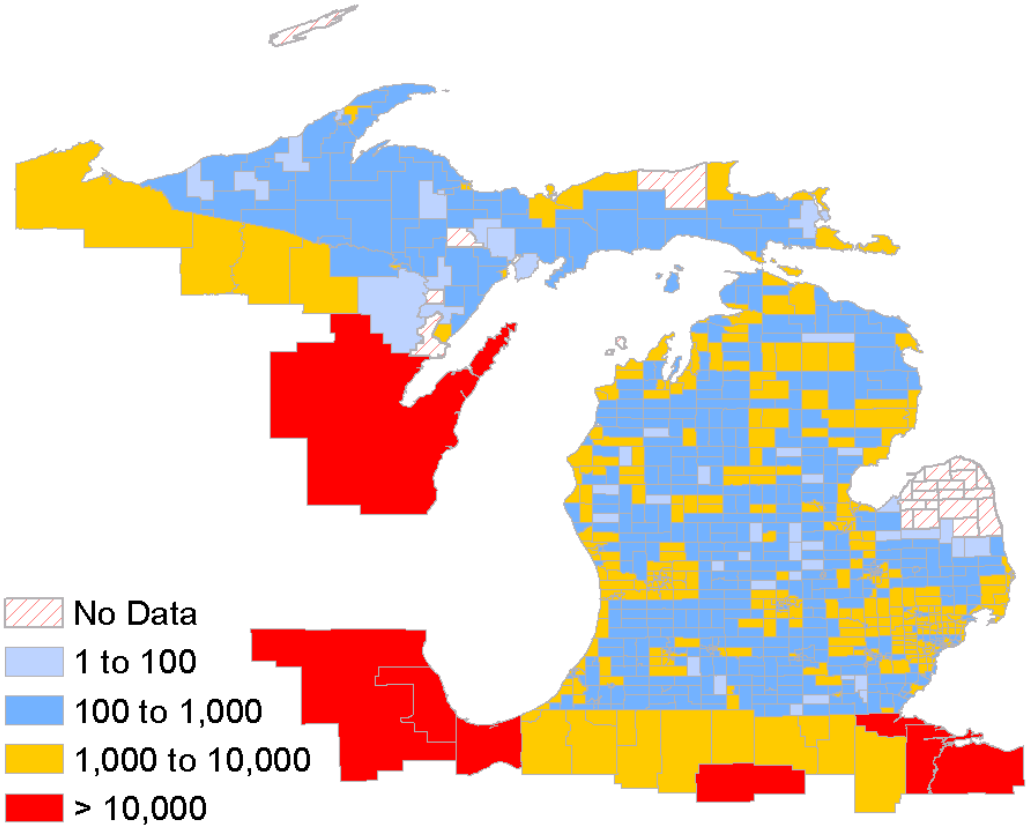
The Problem with Passive Data

What’s Missing?

- **Information**
  - Travel mode
  - Activity purpose
  - Traveler characteristics
- **Travel & Travelers**
  - Geographic coverage
  - Seniors & low income populations
  - Short activities & trips

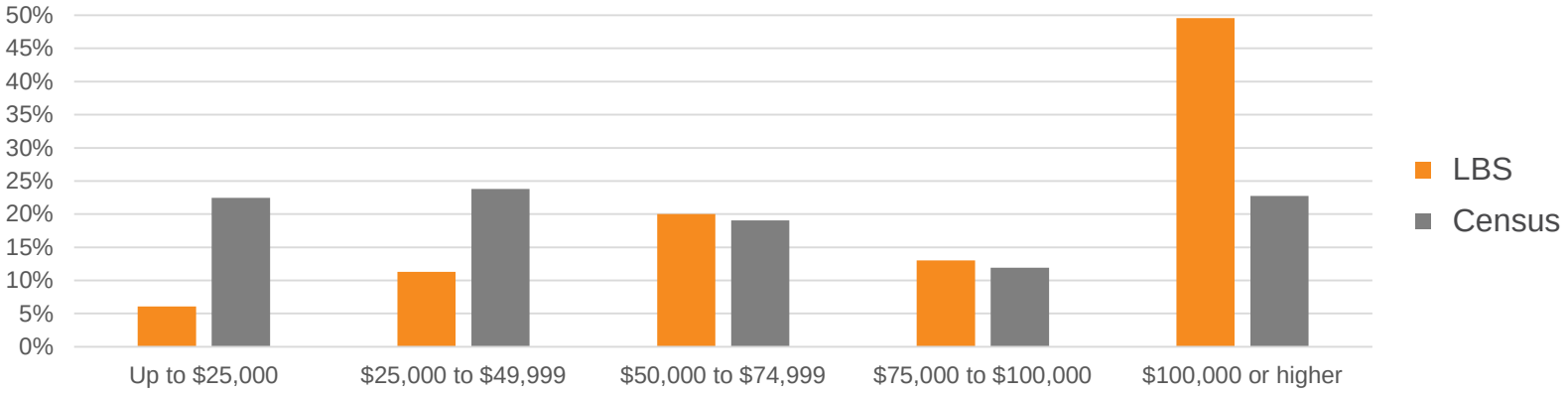
Geographic Coverage Gaps

RESIDENT FALL SHORT-DISTANCE ORIGINS

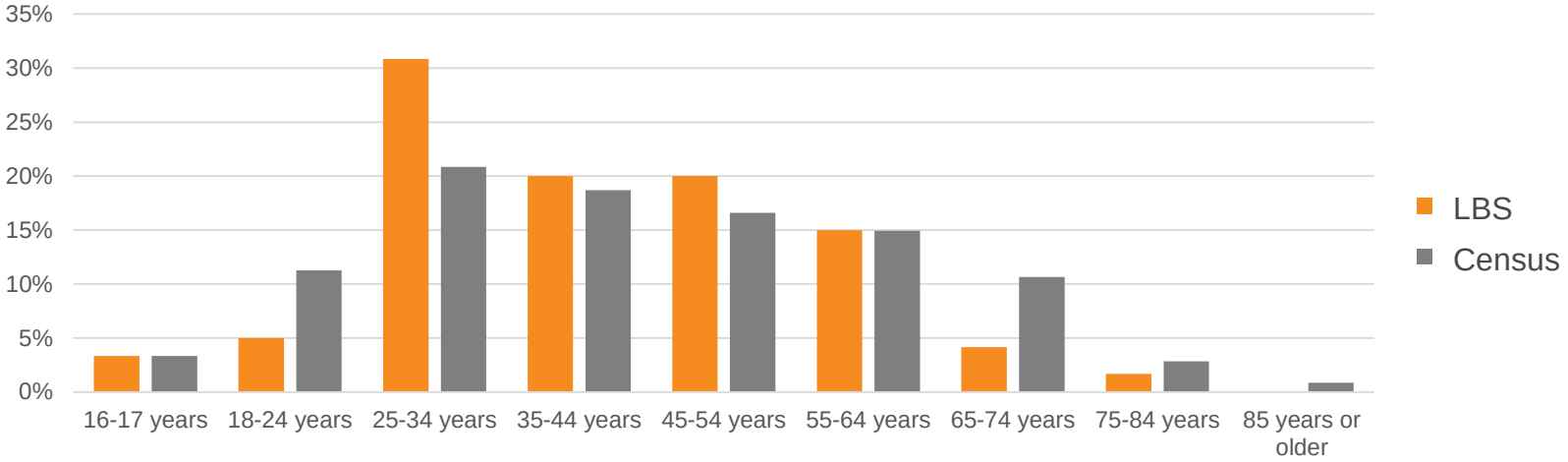


Demographic Biases

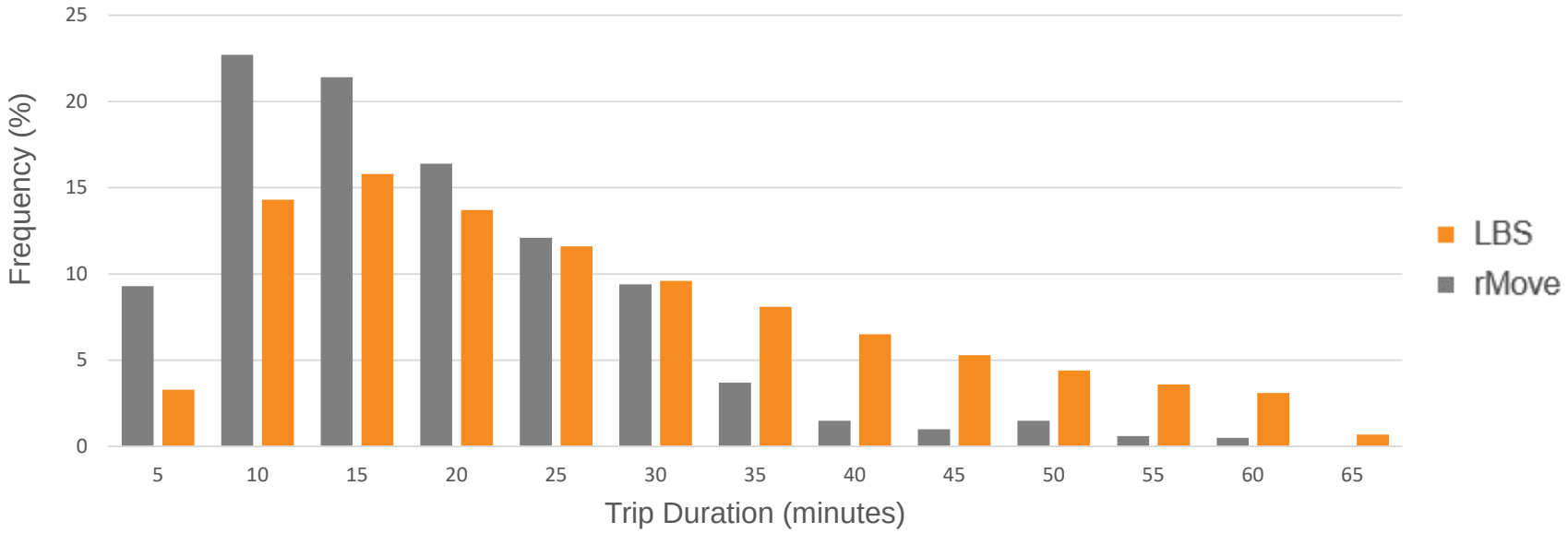
INCOME



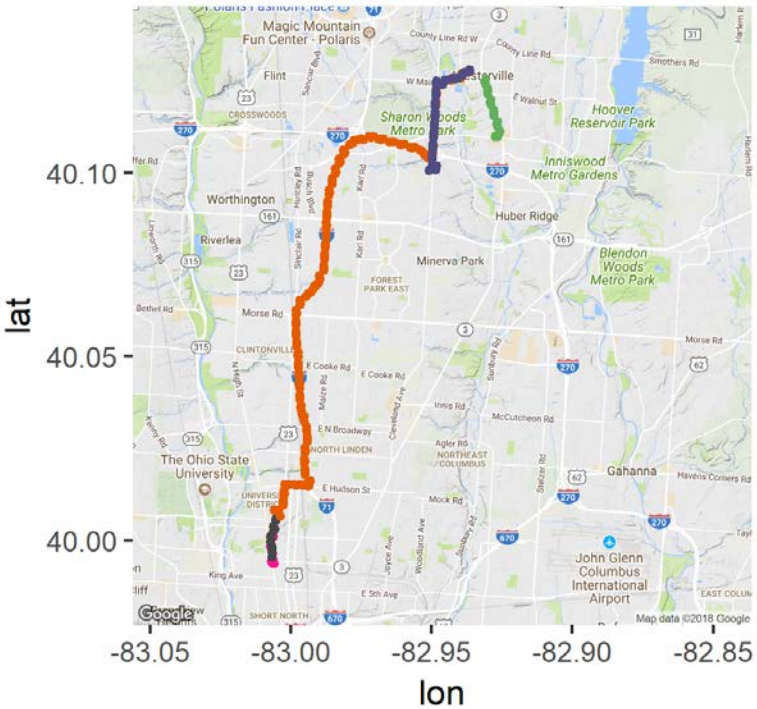
AGE



Duration Bias

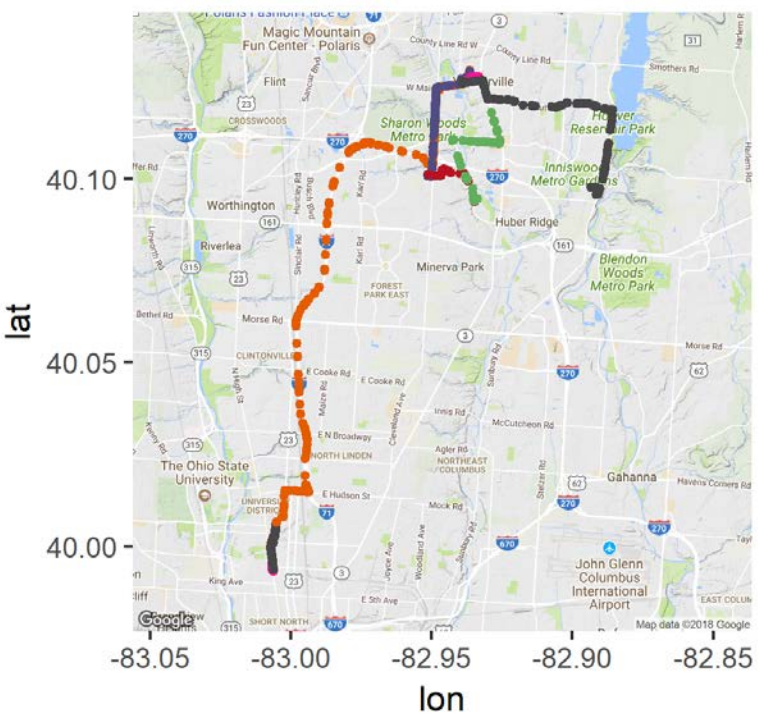


Missing Short Trips Due to Temporal Sparsity



LBS

10:48
15:32
15:57
17:36
18:04
18:07
19:24
19:27



rMove

Fixing the Problems

If we can measure bias, we can correct for it.

- Aware of 8 expansion methods currently in use, and new methods being actively being researched
- Most robust expansion schemes combine several methods
  - SE data based and simple scaling to counts are among most commonly used and most commonly used alone
  - But these cannot correct for trip/activity duration biases
- Group methods first by type of control data used
  - Then subdivide count-based methods based on single/multiple factors, network-based/not, parametric/non-parametric

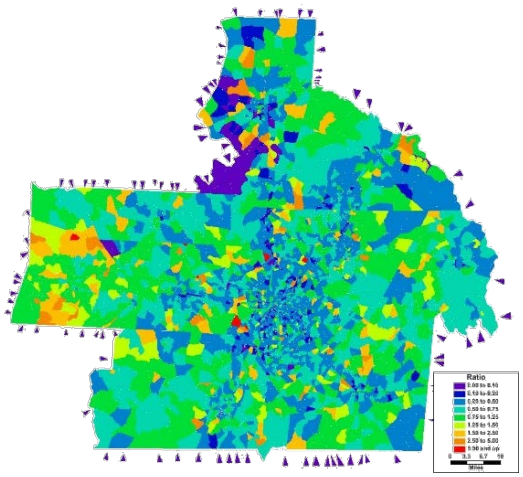
Taxonomy of Expansion Methods

- Demographic Data Methods
  - **1. Market Penetration (Residence-Based)**
  - **2. Trip Generation-Based**
- Traffic Count Methods
  - **3. Simple Scaling to Counts**
  - Multi-factor Scaling
    - Non-Assignment-Based
      - **4. Iterative Proportional Fitting to Counts (Frataring)**
      - **5. Iterative Screenline Fitting (ISF) / Matrix Partitioning**
    - Network Assignment-Based
      - Nonparametric (ODME)
        - » **6. Direct ODME**
        - » **7. Indirect ODME**
      - **8. Parametric Scaling to Counts**
- *Trace Data Methods*

Demographic Data Methods

1. Market Penetration-Based

- Requires device ID persistence to impute residence location
  - Not currently viable for GPS datasets
- Compare resident devices per area to population to compute expansion factors by device residence areas
- Good for addressing demographic biases, not for duration bias



2. Trip Generation-Based

- Does not require residence imputation/ID persistence
- Compares trips to/from zone to estimated trips to/from zone to estimate expansion factor
- May be better for data validation than data expansion

Simple Count-Based Methods

3. Simple Scaling to Counts

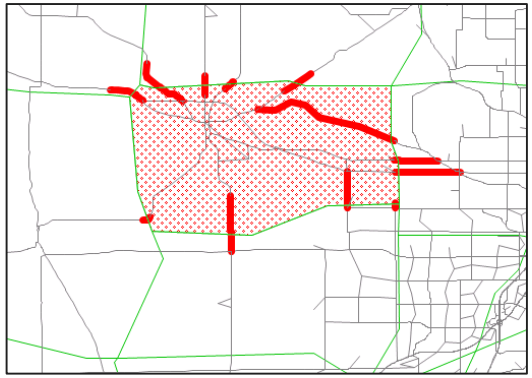
- Use a single expansion factor to minimize average loading error
  - Usually done via assignment but can be done with map-matching for data with sufficient locational precision (GPS, some LBS)
- Almost always used as part of/in combination with other more complex count-based methods
- Sometimes explained in terms of vehicle occupancy but this is only one of several effects that can be captured/reflected

4. Iterative Proportional Fitting to Counts (Fratar)

- Requires counts into/out of zone
- Commonly used for expanding external stations
- Also sometimes for airports and other special generator zones

5. Iterative Screenline Fitting (ISF)

- Loop over screenlines
  - Uses screenlines which partition region into two sets of zones – which partition the OD matrix into quadrants
  - Diagonal quadrants receive factor of 1
  - Off-diagonal quadrants receive factor based on ratio of weighted total counts to aggregated OD trips
    - Weight based on number of screenlines each count is on, etc.
  - Average new factors from this screenline with prior expansion factors



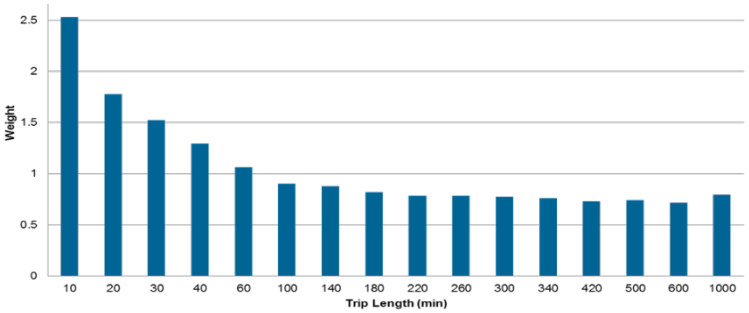
Non-Parametric Assignment-Based Methods

6. Direct ODME

- OD/cell-specific expansion factors (lots)
- Beware of over-fitting to counts!
  - Many different ODME methods, **important** to use one that either minimizes error with respect to both counts and the original ODs or that minimizes error with respect to counts but only within certain constraints (e.g., -50% and +200%) – easier if ODME done after other methods
  - Should measure difference/distance from original to output OD flows (e.g., MAE, MAPE), not just compare TLFDs
- Relatively easy to do but difficult to interpret / understand

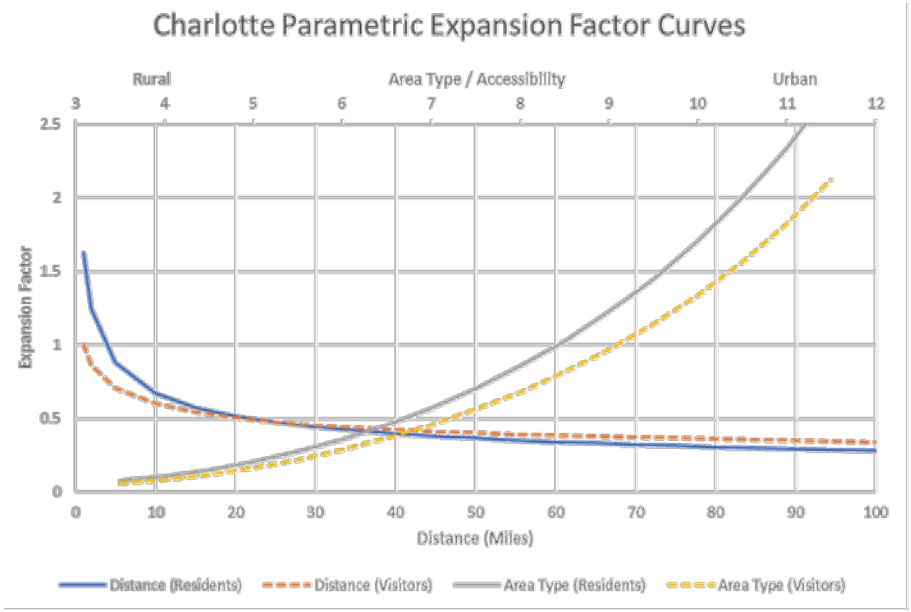
7. Indirect ODME

- Analyze results of ODME to create simpler set of expansion factors based on distance, regions, etc.

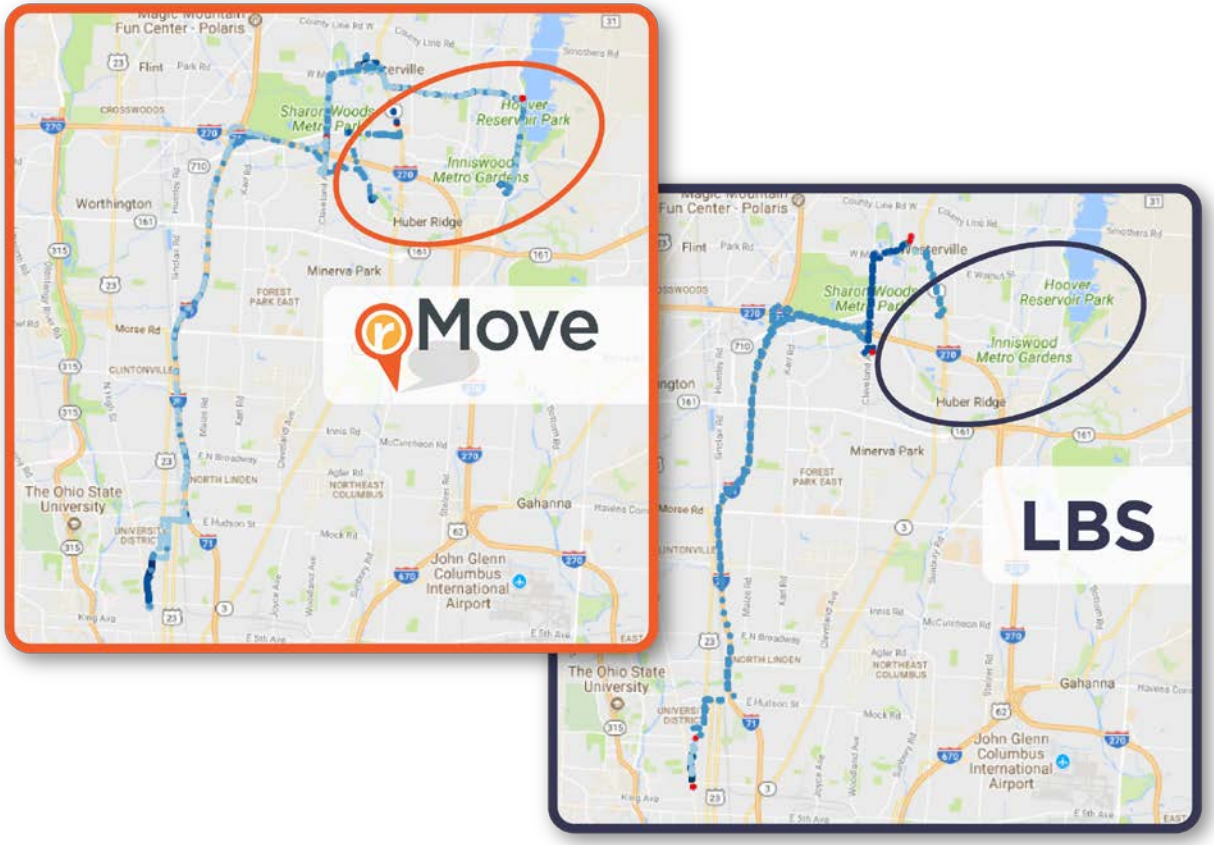


8. Parametric Scaling to Counts

- Uses assignment within a larger framework to estimate/ calibrate parameters for an expansion factor function
- Terms often include:
  - Distance
  - Area type or accessibility
  - Intradistrict/intrazonal
  - Adjacency
- Estimation is NP-Hard
  - Mixed success with genetic algorithm
  - Mixed success with regression on ODME
  - Manual calibration



9. Disaggregate Trace Auditing



Example of matched traces with short trips missing in LBS

Comparison of Expansion Methods

	Fix Trip Length Bias	Fix Coverage Problems	Fix Demographic Bias	Independent of Network	Ease of Application	Holdout Count Sample	Transparency
1 MARKET PENETRATION-BASED	✗	✓	✓✓	✓	✓	✓	-
2 TRIP GENERATION-BASED	✗	✓	✓✓	✓	-	✓	✓
3 SINGLE-FACTOR SCALING	✗	✗	✗	✓	✓	-	✓
4 FRATARING	✗	✓	✗	✓	✓	✗	✓
5 ITERATIVE SCREENLINES	✓	✓	✓	✓	-	✓	✓
6 DIRECT ODME	✓	✓	✓	✗	✓	-	✗
7 INDIRECT ODME	✓	✓	✓	✗	✗	-	-
8 PARAMETRIC SCALING	✓✓	✓	✓	✗	✗	-	-
9 DISAGGREGATE TRACE AUDITING	✓✓	✓	✓✓	✓	✗	✓✓	✓

Final Thoughts

Summary

- Many types/sources of passive OD data
- All suffer from systematic biases
- Biases can be corrected through analysis together with other data sources
- Ensemble expansion methods are best for now
- Count-based methods are necessary for now
- Smartphone travel survey data is especially promising in correcting passive data at the disaggregate level

Contact

www.rsginc.com

Vince Bernardin, Jr, PhD
DIRECTOR OF FORECASTING
Vince.Bernardin@rsginc.com
812.459.3500