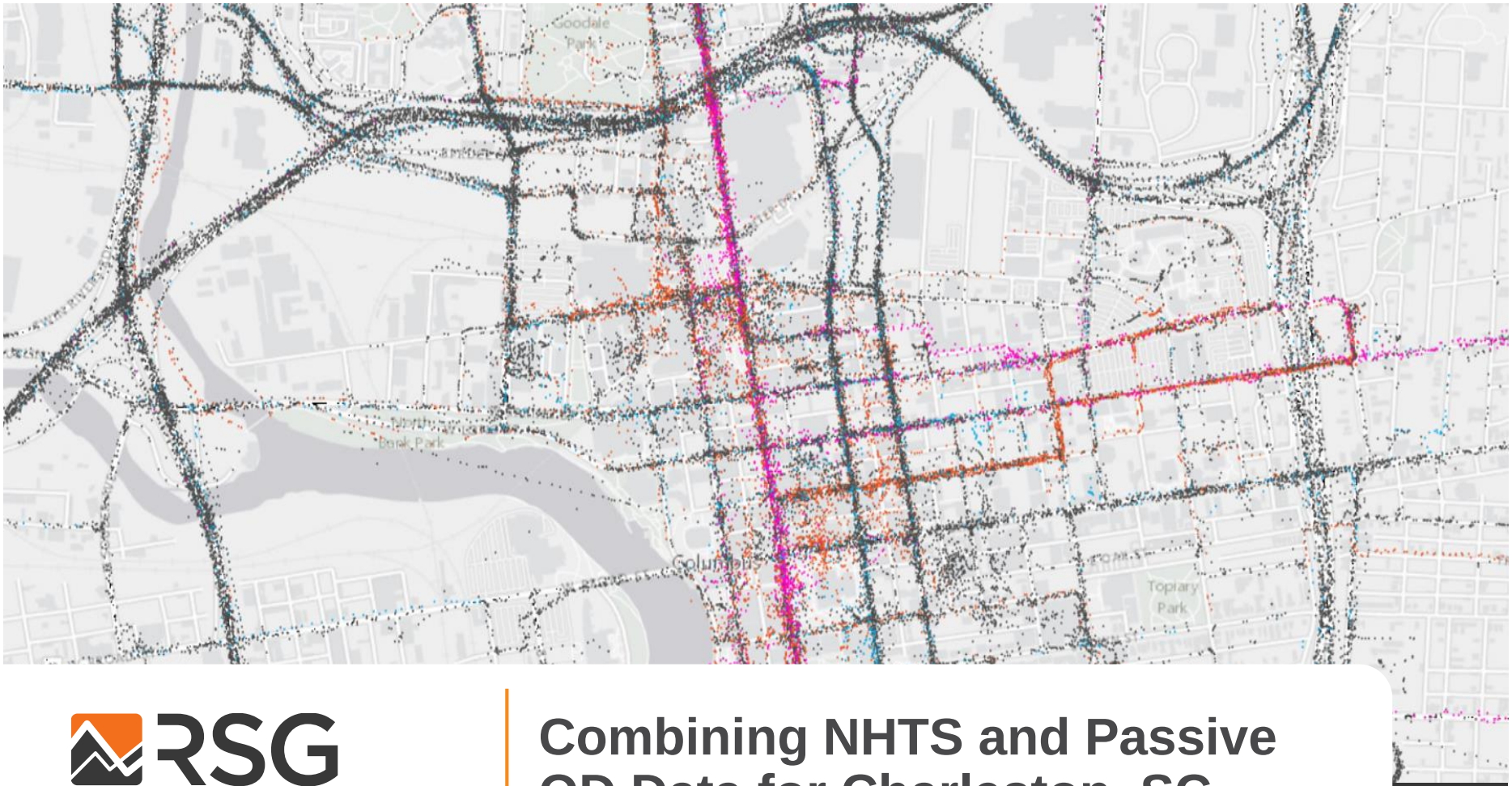


RSG
the science of insight

Combining NHTS and Passive OD Data for Charleston, SC

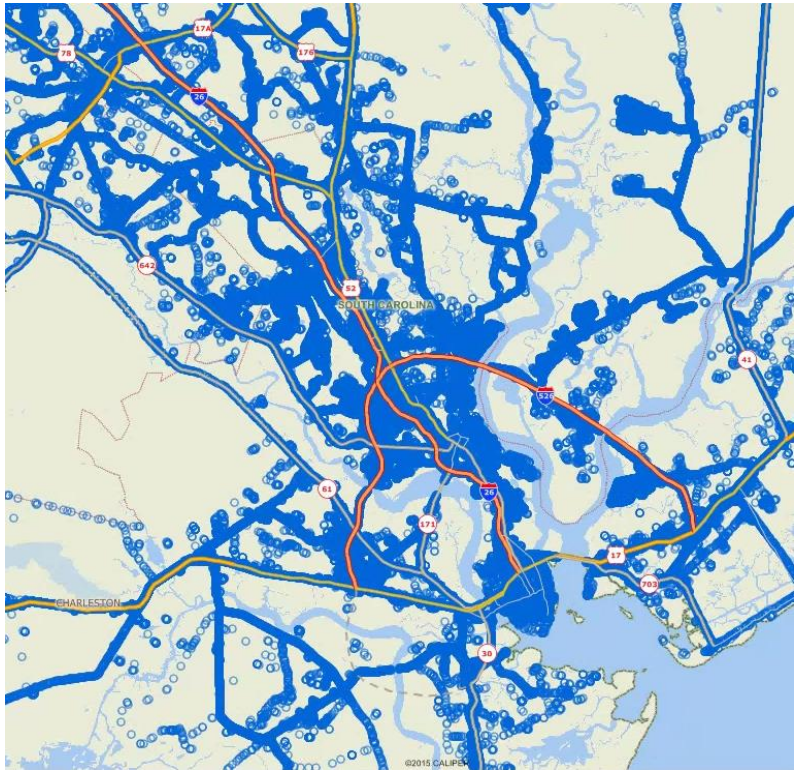
Vince Bernardin, PhD
Hadi Sadrsadat, PhD
Jason Chen, PhD

August 9, 2018

Charleston

- MPO: Berkeley-*Charleston*-Dorchester Council of Governments (BCDCOG)
- Population ~775k
- Among the top 20 fastest growing MSAs in US
- Model Update:
  - Hybrid w/ NHB-HB linkage
  - New Modes, Nesting
  - Destination Choice
  - Visitor Model

Charleston Data



2017 NHTS

Sample: 1,104 Households

PASSIVE DATA

AirSage

- 870 x 870 matrices
- By residents & visitors

ATRI

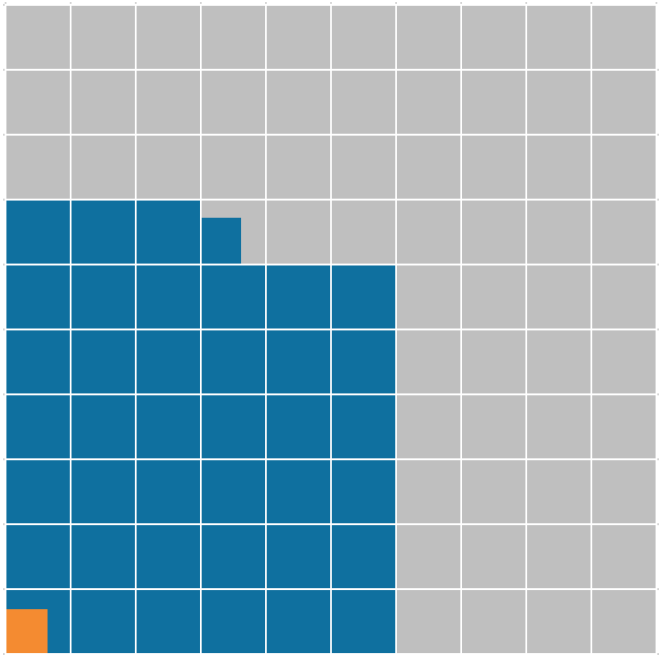
- Over 37,000 trucks
- Over 150k truck trips
- 30 days of data

Limitations of NHTS

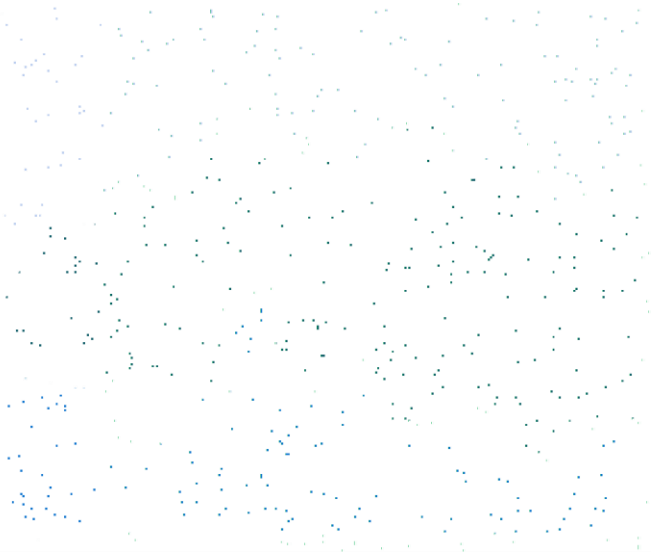
Why not just use NHTS?

- Never has been “just” NHTS
  - All surveys use control data for sampling
- Need for more frequent data to track changes – but no $
- Better location/spatial coverage from Passive OD data
  - OD Pairs

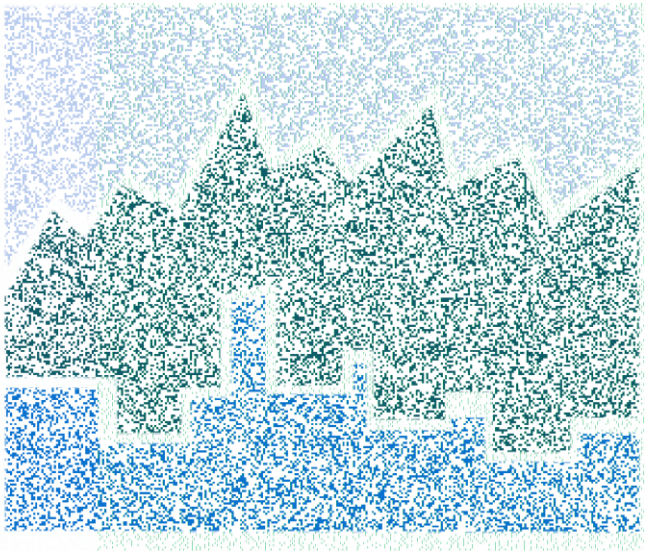
Total:	760,384	
Survey:	5,006	**0.7%**
AirSage:	253,304	**33.0%**



Can you recognize the pattern based on 0.7%?

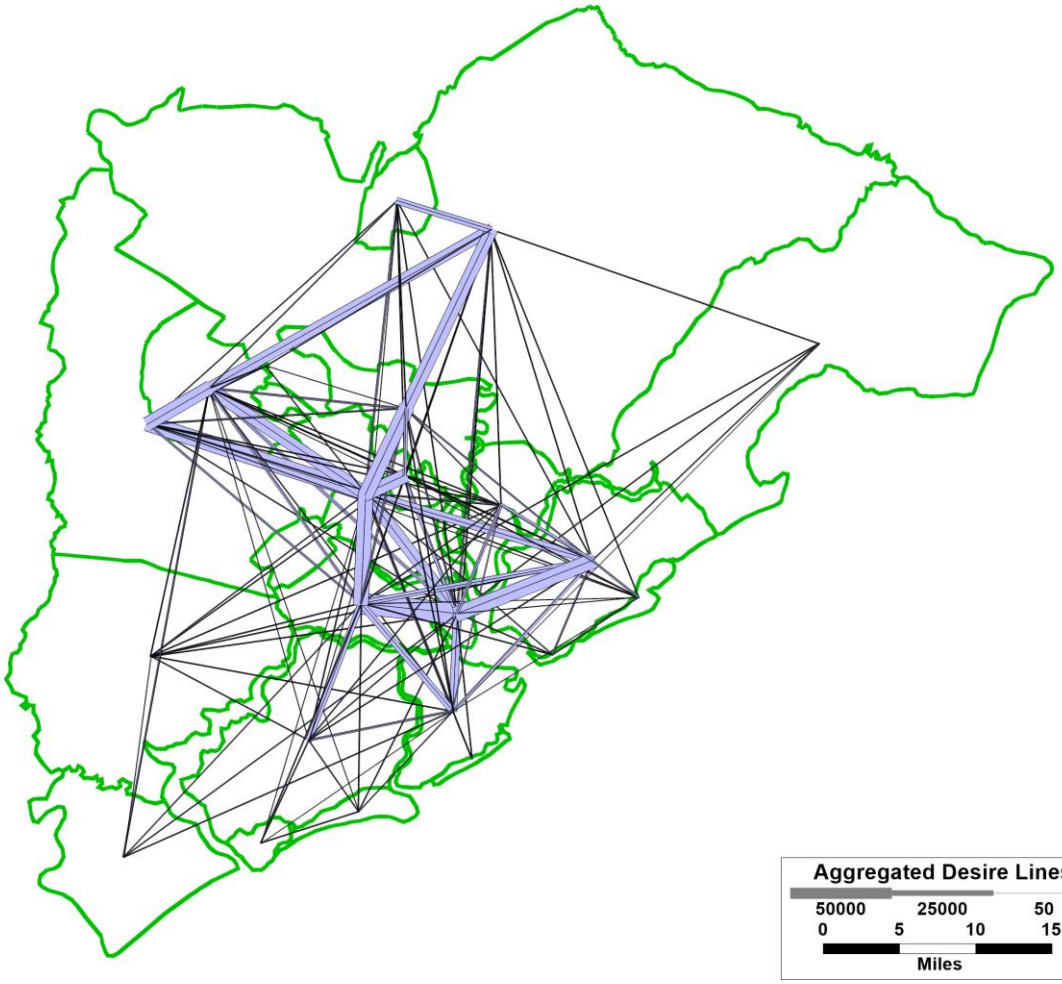


How about based on 33%?

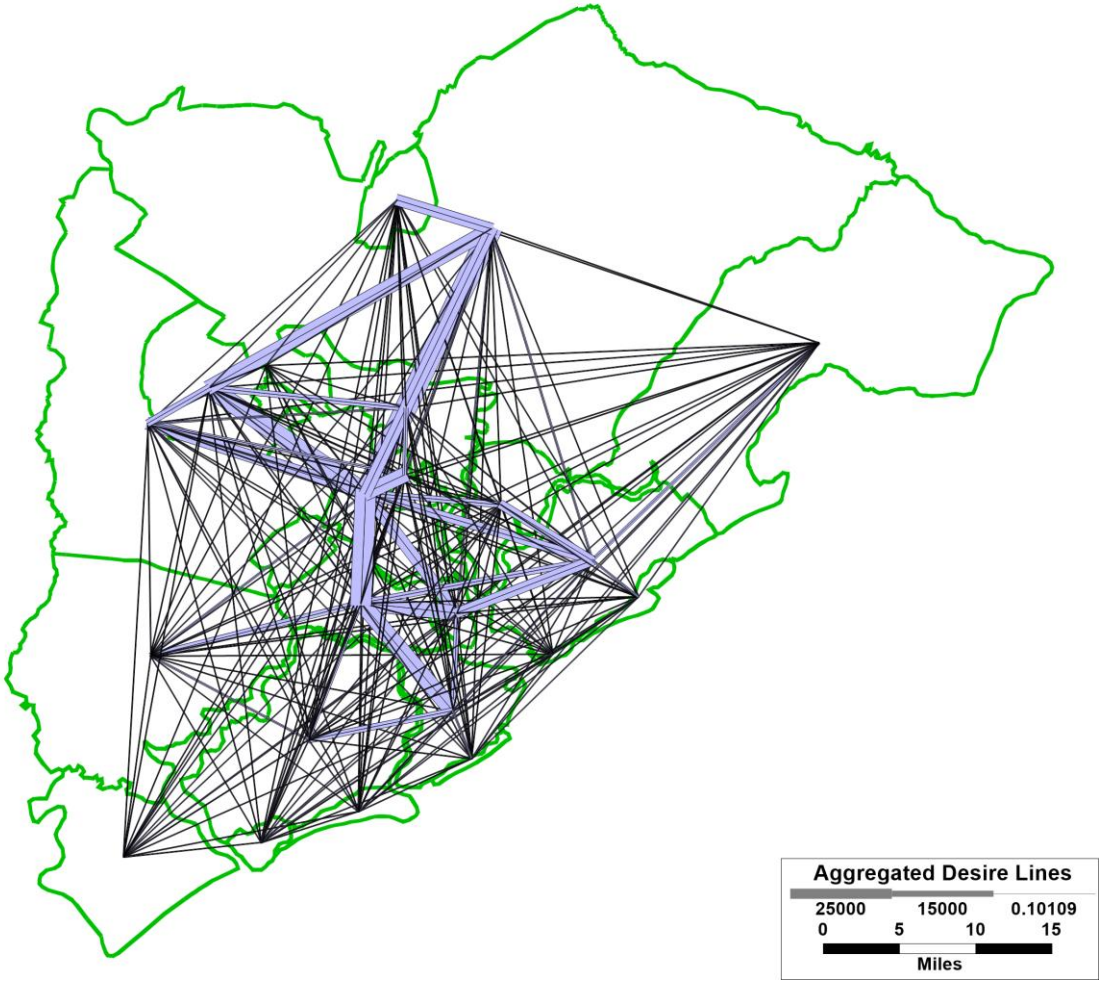


Big Data allows us to see the big picture.

NHTS desire lines (at the district level) provide data for only 58.1% of cells



Big Data desire lines (at the district level) provide data for 96.7% of cells



Limitations of Passive Data

What's missing?

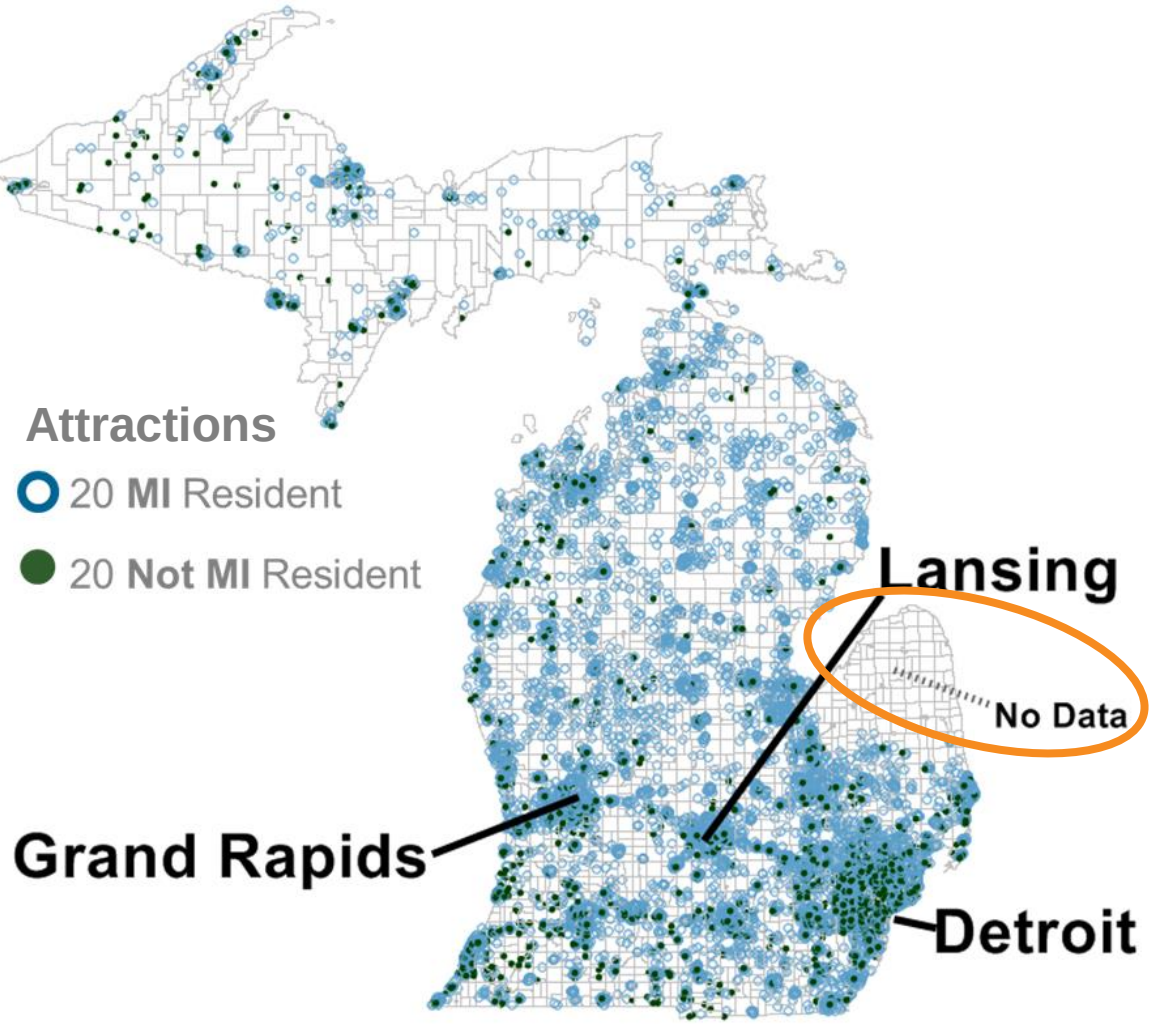
INFORMATION

- Travel mode
- Activity purpose
- Traveler characteristics

TRAVEL & TRAVELERS

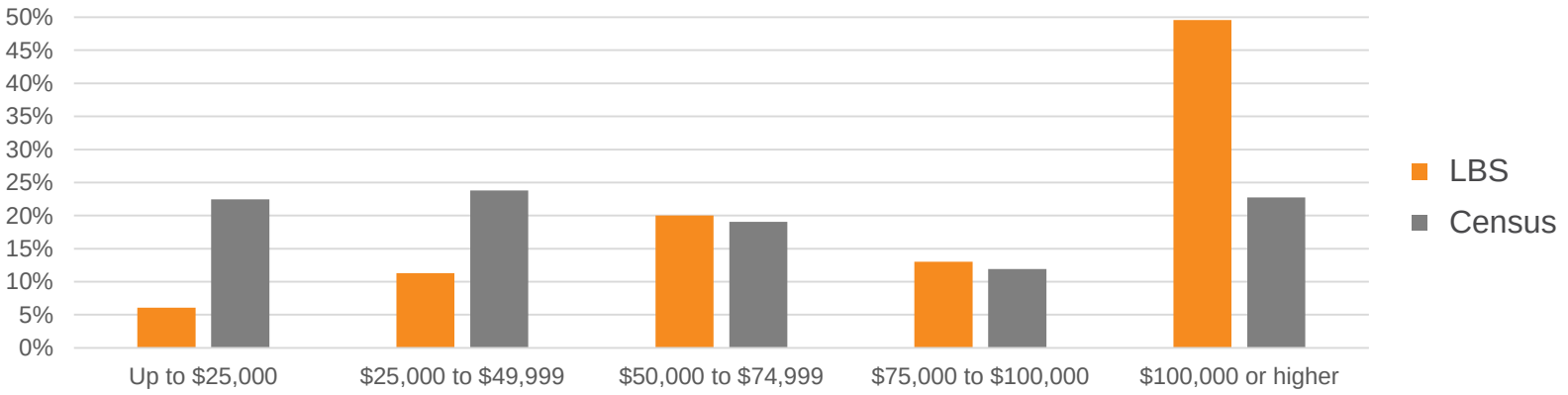
- Geographic coverage
- Seniors & low income populations
- Short activities & trips

Geographic Coverage Gaps

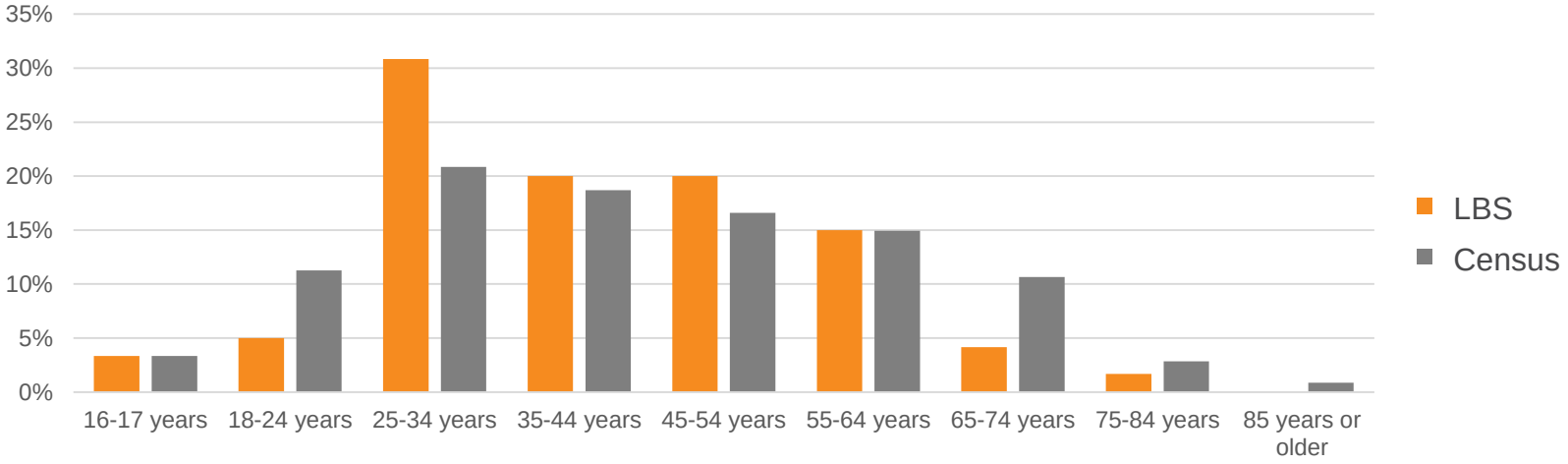


Demographic Biases

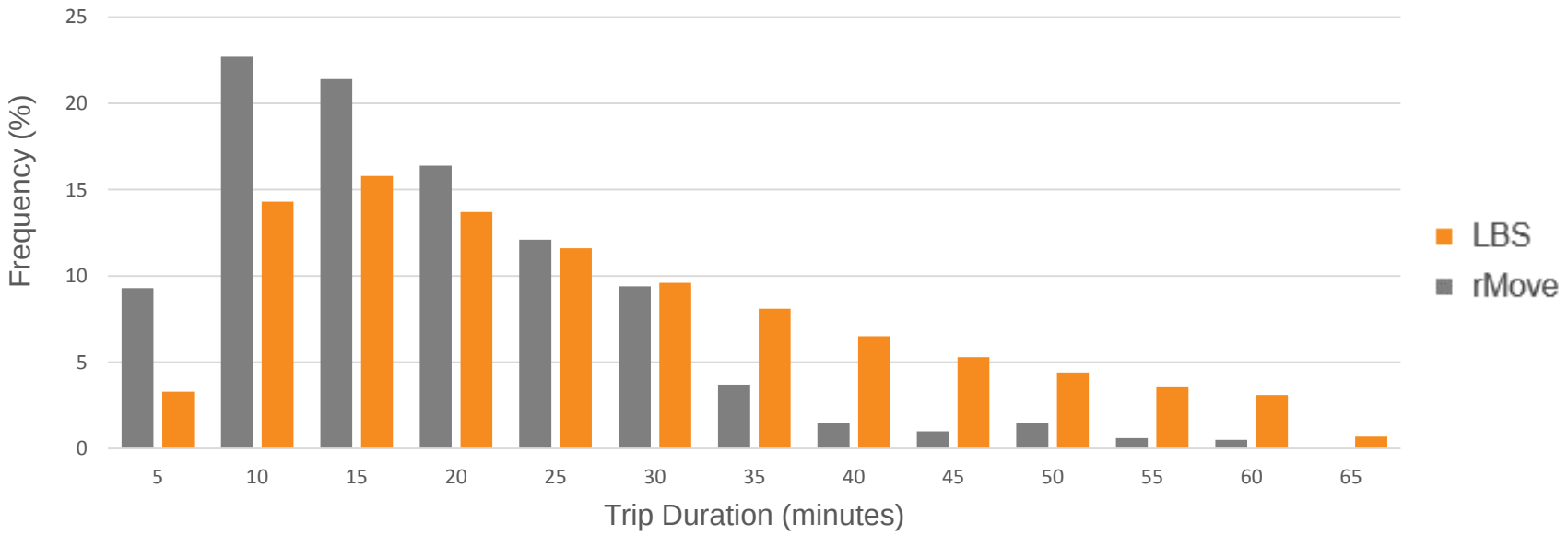
INCOME



AGE



Duration Bias



Understanding Data Fusion

Motivation and Elements

Other data sources – especially smartphone survey data – have what passive data lacks

		Surveys	Passive LBS	Counts	Census	Marketing
Demographics	(Who?)	✓			✓	✓
Locations	(Where?)	✓	✓	✓		
Purpose	(Why?)	✓				✓
Mode	(How?)	✓		✓		
Time	(When?)	✓	✓	✓		

The End Goal – Three Options for Privacy

1. **Aggregate OD Data by Market Segment**
2. **Anonymized Diary Data**
3. **Synthetic Diary Data**

Aggregate OD data eliminates most privacy issues but presents dimensionality limitations.

Aggregate OD Data by Market Segment

Anonymized Diary Data

Synthetic Diary Data

- Multidimensional matrix
- Easier for ODT only
- Substantial limitation on the dimensionality of the data
- Difficult to support ABMs
- Eliminates most – but not all – privacy issues and remaining issues are difficult

Anonymized diary data is easy to understand but difficult to guarantee privacy.

Aggregate OD Data by Market Segment

Anonymized Diary Data

Synthetic Diary Data

- Limited aggregation in multiple dimensions /addition of noise
- **Advantage**: real data, easy to understand/explain
- **Disadvantage**: difficult to guarantee 100% privacy

Synthetic diary data has no privacy concerns but is difficult to verify how well raw data is reproduced.

Aggregate OD Data by Market Segment

Anonymized Diary Data

Synthetic Diary Data

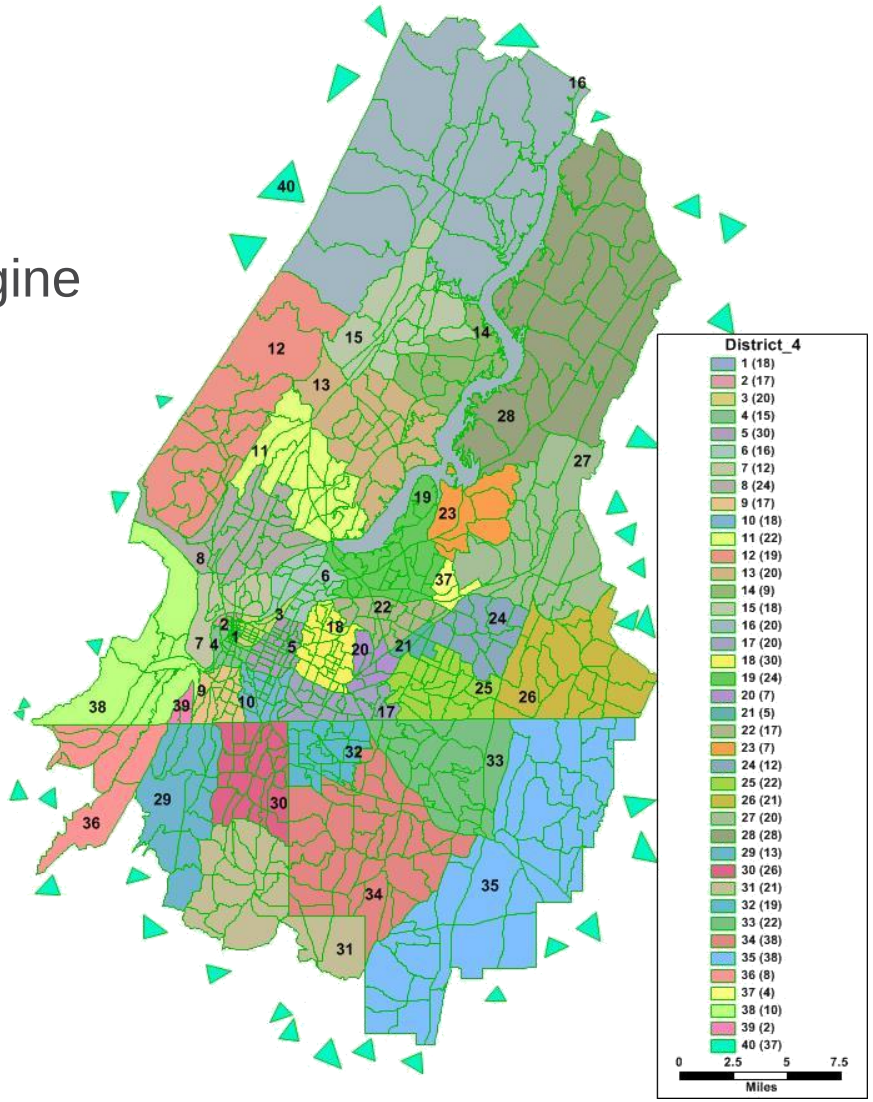
- Modeled data that reproduces certain aspects of raw data
- Model and data entangled
- **Advantage**: No privacy concerns
- **Disadvantage**: Important & difficult to verify how well raw data is reproduced

EXAMPLE APPLICATION

Chattanooga

Chattanooga Daysim

- Daysim ABM as a data fusion engine
- Shadow pricing: Used 40 district scheme with LEHD and AirSage data



DESTINATION DISTRICT O-D SHADOW PRICING CONVERGENCE SUMMARY

ITERATION	ABSOLUTE ERROR	MEAN ABSOLUTE % ERROR	WEIGHTED MEAN ABSOLUTE % ERROR	RMSE
1	516,595	23.3%	22.2%	37.1%
2	421,404	20.6%	19.1%	30.7%
…	…	…	…	…
24	59,962	11.8%	8.3%	**10.5%**

Synthetic Diaries/ODs Replicate Data

- Very good agreement – **10.5% RMSE**
- All cells within +/- 1%
- All residence/work Super Districts within +/-2.5%

TOTAL DAYSIM TRIP TABLE VS. AIRSAGE

Origin SuperDistrict	Destination Super District												Grand Total
	1	2	3	4	5	6	7	8	9	10	11	12	
1	0.5%	0.2%	-0.1%	0.0%	0.0%	-0.1%	-0.2%	-0.1%	0.0%	0.0%	-0.1%	-0.2%	0.0%
2	0.3%	0.0%	0.2%	0.0%	0.1%	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	-0.1%	0.7%
3	-0.1%	0.1%	0.0%	-0.1%	-0.2%	0.0%	0.1%	0.1%	0.0%	0.0%	0.0%	-0.1%	-0.1%
4	0.0%	0.1%	-0.1%	0.0%	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.4%
5	0.1%	0.1%	-0.1%	0.0%	0.2%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%
6	-0.1%	-0.1%	0.1%	-0.1%	0.1%	0.0%	0.1%	-0.1%	0.1%	0.0%	0.0%	0.0%	0.0%
7	0.0%	0.0%	0.2%	0.1%	0.1%	0.0%	0.2%	0.0%	0.1%	0.0%	0.0%	0.1%	0.7%
8	0.0%	0.1%	0.1%	0.1%	0.0%	-0.1%	0.1%	0.0%	-0.2%	0.0%	0.0%	0.0%	0.2%
9	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.3%	0.0%	0.0%	0.0%	0.2%
10	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.0%	0.3%
11	0.0%	0.0%	0.0%	-0.1%	0.0%	0.0%	-0.1%	0.0%	0.0%	0.1%	-0.1%	-0.3%	-0.5%
12	-0.2%	-0.3%	-0.1%	-0.2%	0.0%	-0.1%	-0.2%	-0.1%	-0.1%	0.0%	-0.3%	-0.7%	-2.4%
Grand Total	0.5%	0.2%	0.2%	-0.2%	0.4%	-0.3%	0.4%	0.1%	0.3%	0.3%	-0.5%	-1.3%	0.0%

Chattanooga Takeaways

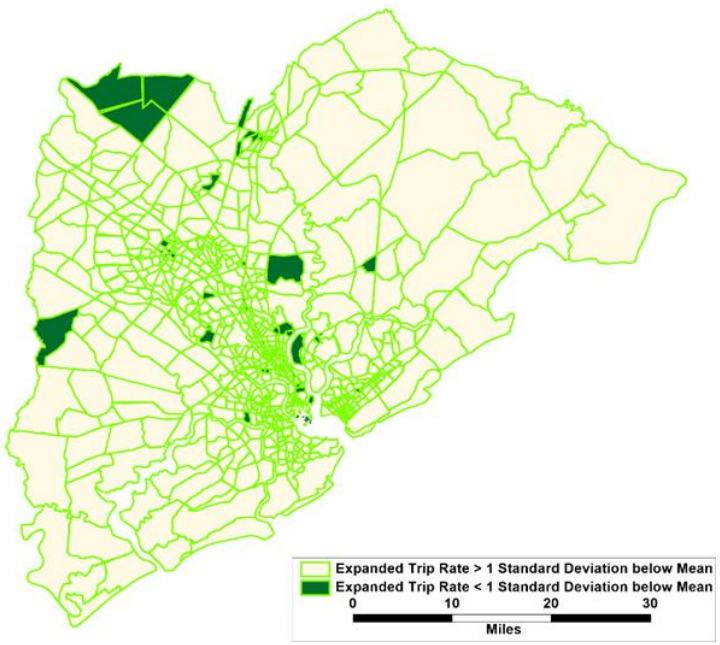
- Successfully created synthetic disaggregate data via fusion of AirSage and travel survey data in 2016
- Chattanooga Daysim produces synthetic trip-list
  - with OD patterns from AirSage
  - and traveler characteristics, travel modes, and activity purposes from survey
- First synthetic travel data from passive & survey data
- Great if limited to aggregate passive data (e.g., if GDRP in US)
- More/better is possible, especially with disaggregate methods

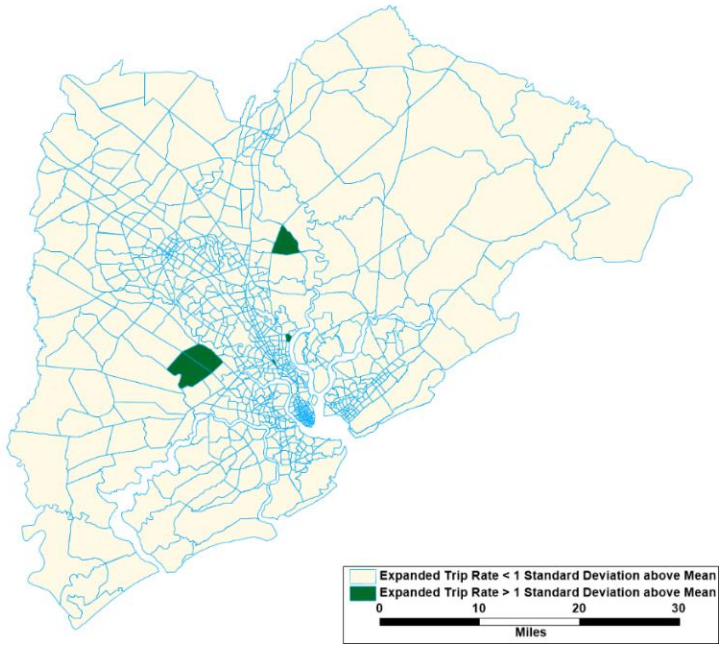
EXAMPLE APPLICATION

Charleston

Cross-Validation: Trip Rates

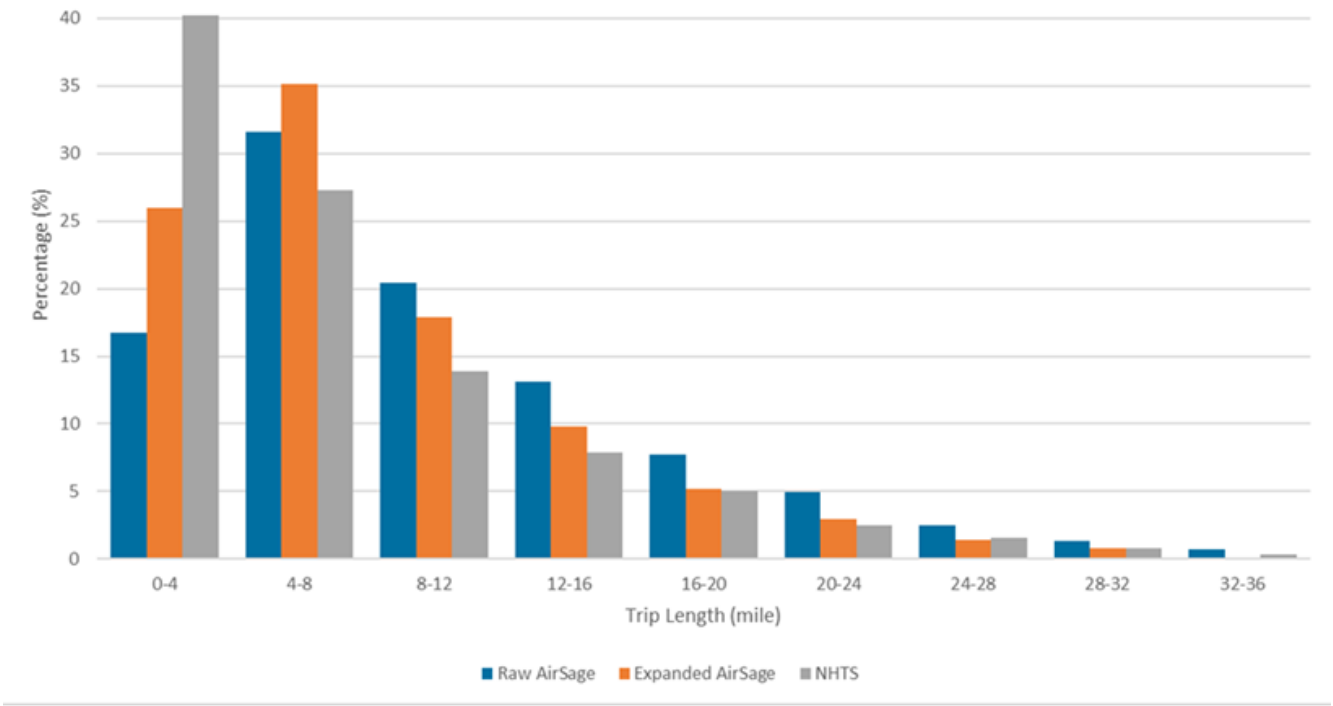
Internal Non-Truck Trips	Person Trip Rate
NHTS Add-on	8.8
Passive Data (Initial Data)	17.1
Passive Data (Expanded Data)	9.5





Cross-Validation: Trip Lengths

Data	Average Trip Length	
	w/ Intrazonals	**w/o Intrazonals**
NHTS	6.6	**7.3**
Initial AirSage	7.1	**10.3**
Expanded AirSage	5.7	**8.4**



Data Expansion (Passive OD – Count Fusion)

- Expansion to traffic counts (519 stations)
- Expansion by vehicle class
  (Auto, Single-Unit Truck, Multi-Unit Truck)
- Expansion process steps
  - Fratar at external stations
  - Iterative Screenline Fitting (ISF)
    - only for all classes together
  - Constrained ODME
    - Factors constrained 0.5 to 3.0

Destination Choice (Survey – Passive OD Fusion)

- Estimated destination choice models
  - with and without constants in the utility function

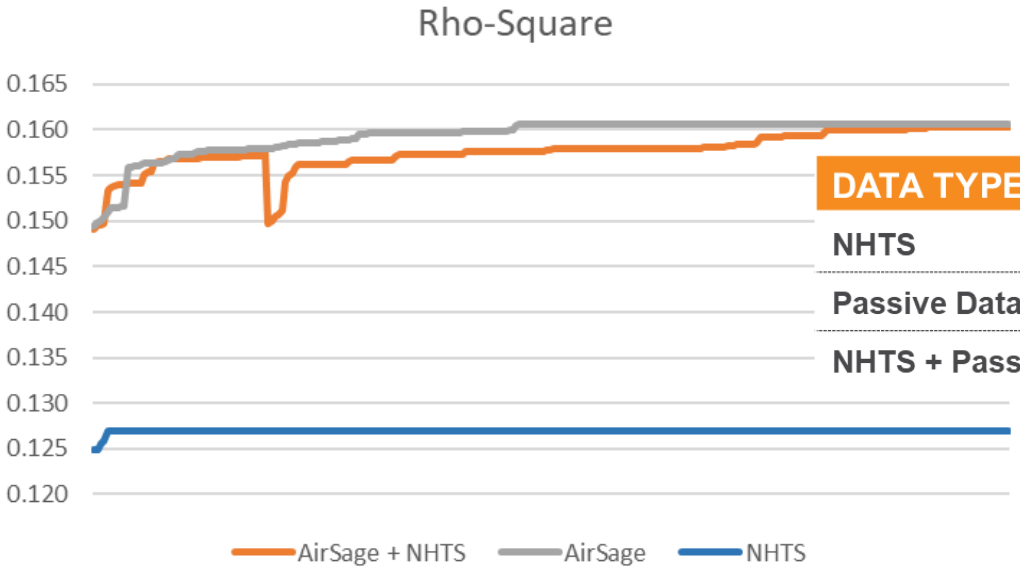
$$U = \sum \alpha + \sum \beta x + \varepsilon$$

  - from NHTS, Passive Data, and both (**simultaneously**)

- Composite MLE with genetic algorithm for simultaneous estimation (with embedded shadow-pricing for constants)
  - Destination choice models with constraints / agglomeration effects are not GEV or guaranteed to be globally convex

Destination Choice (Survey – Passive OD Fusion)

- Models without constants estimated
  - NHTS given equal weight as AirSage (generous)
  - Models using AirSage yield better fit, different sensitivity
- Embedded shadow-pricing causing oscillation, working on algorithm to estimate models with constants



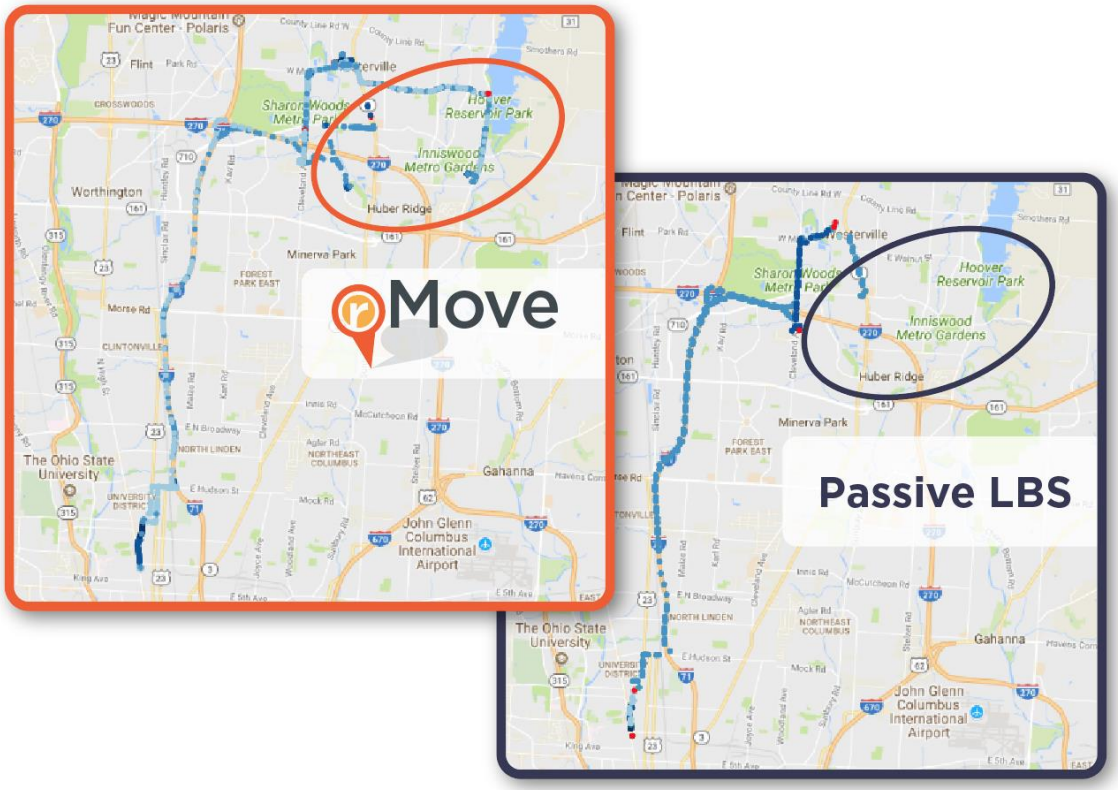
DATA TYPE	NO CONSTANT	W/ CONSTANTS
NHTS	0.1270	n/a
Passive Data	0.1607	tbd
NHTS + Passive Data	0.1603	tbd

Charleston Takeaways

- NHTS trip rates may be slightly low
- NHTS helpful in identifying coverage issues in AirSage
- AirSage helpful in identifying special generators
- Count-based expansion of AirSage may not fully correct duration (short trip) bias
- Successfully estimated spatial models **simultaneously** from NHTS and AirSage data using maximum composite likelihood estimation
- Use of AirSage data significantly improves spatial models
- Important to investigate sensitivities for forecasting

Final Thoughts

rMove has what Passive Data lacks



- rMove trace data as labeled training set for AI
- Imputation of
  - Missing trips
  - Mode
  - Purpose

Questions on Data Fusion for NextGen NHTS

PRIVACY

- How will privacy be protected?
- What will be the data product(s) produced by data fusion?

FORECASTING / CONSISTENCY

- How will data fusion and forecasting models be related?
- Will there be consistent assumptions?

TRANSPARENCY

- What methods will be used?
- Will they be documented?

VALIDATION

- How will the fusion be validated?
- What error statistics or other measures will be reported?

Summary

- Data fusion is attractive given the limitations of both survey and passive data
- There are different privacy protection strategies and data fusion methodologies
- Aggregate passive data has been successfully fused with disaggregate survey data to produce disaggregate synthetic data
- Data fusion produces better forecasting models
- Fusion using disaggregate passive data may be even more promising

Contact

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

www.rsginc.com