RSG
the science of insight

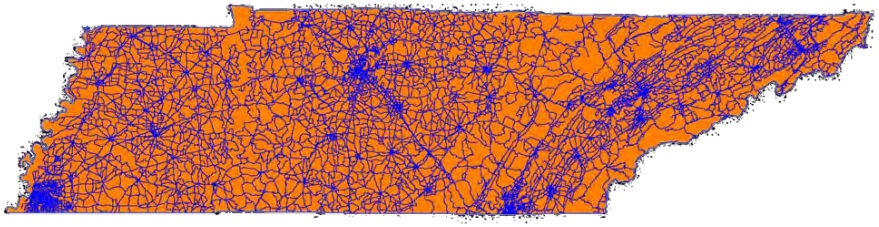
Incorporating Big Data in Statewide and MPO Travel Demand Models in Tennessee

Vince Bernardin, PhD, RSG

February 7, 2017

What does Big Data mean for modeling?

Recent Experiences



- Tennessee Statewide Model
  - First statewide model calibrated with AirSage
- Chattanooga MPO Model
  - First activity-based model incorporating AirSage

rMOVE by RSG

- Other experiences
  - California, New York, Michigan, Indiana, Ohio, Virginia, Maryland, North Carolina, Florida, Iowa, South Dakota, Utah, Idaho, Oregon, Washington, Alaska…
  - rMove, Bluetooth, toll transponders

Types of Passive OD Data

Types of Big OD Data

	Cell-Tower Signaling	LBS	GPS	Bluetooth
Source	Cell phone – tower communications	Smartphone / tablet apps	On-board navigation devices / apps	Bluetooth devices
Universe	All travel	All travel	Heavy trucks, medium from some providers, private from some providers	All travel
OD Demand Types	Aggregate trip ODs	Aggregate trip ODs	Aggregate trip ODs; sometimes disaggregate traces also available but with restricted use	Disaggregate trip ODs
OD Travel Time Data (Including Reliability)	Not available	Not available	Available with varying degrees of processing effort depending on provider	Generally produced as part of the processing of trips

- Different types of data are different
  - Important to know what can and cannot be done with each type

Precision, Sample Size & Coverage

	Cell-Tower Signaling	LBS	GPS	Bluetooth
Locational Precision	>100 m often ~200–2000 m	10–100 m often ~30 m	1–10 m	10–100 m
Sample Penetration	6–10%	5–8%	9–12% truck; ~0.5% private	4–9%
Data Collection Time Period	Typically 1 month	Currently a few months	1 month to 2 years depending on provider and pricing	Typically <1 month
Coverage Issues	Poor coverage in some (mostly rural) areas	--	--	Coverage limited—requires mounting detector devices

- This is what you are paying for
  - Sample penetration and/or sample penetration x time period is how much information you are getting
  - Precision limits what you can do with it
- CAUTION: sample penetration may vary by region & over time

Representativeness & Expansion

	Cell-Tower Signaling	LBS	GPS	Bluetooth
Trip-Length / Duration Bias	Confirmed	Suspected	Confirmed	Not suspected
Included / Default Expansion	Residence market share-based; generally requires adjusted to counts	None/single count-based factor, believed to require adjustment for biases	None/single count-based factor, generally requires adjustment for biases	Typically expanded to counts

- This is the big ‘gotcha’ that you will have to fix
- Realize you have to budget for data expansion

Segmentation & Applications

	Cell-Tower Signaling	LBS	GPS	Bluetooth
Number of Zones	Limited by pricing and locational precision	Depends on pricing scheme	Relatively unlimited in most pricing schemes	Limited by number of detector devices
Select Link / Corridor Analysis	Generally indirect only	Indirect only currently but a subset may support direct in the future	Limited or unlimited direct depending on provider, or indirect	Direct only if detector placement allows; indirect
Filtering of Intermediate Stops on Long Trips	Premium option	Not currently available	Depending on provider may be possible as a post-process	Possible as a post-process
Residency Information	Premium options for regional residents vs. nonresidents or home block groups	Not currently available but LBS could support residence class data	Not available due to ID persistence limitations	Generally not possible
Purpose	Premium option for imputed purposes	Premium option for imputed purposes	Not available due to ID persistence limitations	Generally not possible
Vehicle Class	Not available	Not currently available	From some providers *Heavy and medium trucks, private vehicles*	Generally not possible

- Do you want/need direct or indirect corridor level info?
- Are long-distance or visitor trips important?

Cost and Effort

Description	Cell-Tower Signaling	LBS	GPS	Bluetooth
Data Cost	Intermediate	Expensive	Inexpensive to Expensive *depending on provider, amount/length of data period, and amount of processing included*	Expensive
Additional Processing Required	Intermediate	Limited to Intermediate	Substantial to Limited *depending on provider*	Usually included in price
Vendors	AirSage	StreetLight, Cuebiq	ATRI, StreetLight, INRIX, TomTom, HERE	TTI, RSG, others

- Consider full cost, including processing, not just data
- Costs will vary obtain multiple bids / quotes when possible
- Buy what you need, not more, not less

Summary Comparison of Data Types

	Cell Tower Signaling	LBS	GPS	Bluetooth
Description				
Universe	All Travel	All Travel	Heavy Trucks, Medium from some providers, Private from some providers	All Travel
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	Generally customizable down to individual hours of the day; effort to get multiple time periods may vary significantly by provider
OD Demand Types	Aggregate Trip ODs	Aggregate Trip ODs	Aggregate Trip ODs; sometimes disaggregate traces also available but with restricted use	Disaggregate Trip ODs
OD Travel Time Data (including reliability)	Not available	Not available	Available with varying degrees of processing effort depending on provider	Generally produced as part of the processing of trips
Precision and Coverage				
Locational Precision	> 100 m often ~ 200 - 2000 m	10-100 m often ~ 30 m	1 - 10 m	10-100 m
Sample Penetration	6-10%	5-8%	9-12% truck; ~0.5% private	4-9%
Data Collection Time Period	Typically 1 month	Too new to know	1 month - 2 years depending on provider & pricing	Typically < 1 month
Coverage Issues	Poor coverage in some (mostly rural) areas			Coverage limited - requires mounting detector devices
Representativeness & Expansion				
Trip Length / Duration Bias	Confirmed	Suspected	Confirmed	Not suspected
Included / Default Expansion	Residence market share based; generally requires adjusted to counts	None / Single count-based factor; believed to require adjusment for biases	None / Single count-based factor; generally requires adjusment for biases	Typically expanded to counts
Segmentation & Applications				
Number of Zones	Limited by pricing and locational precision	Depends on pricing scheme	Relatively unlimited in most pricing schemes	Limited by number of detector devices
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	Direct only if detector placement allows; indirect
Filtering of Intermediate Stops on Long Trips	Premium option	Not currently available	Depending on provider may be possible as a post-process	Possible as a post-process
Residency Information	Premium options for Regional Residents vs. Non-Residents or Home Block Groups	Not currently available but LBS could support residence class data	Not available due to ID persistence limitations	Generally not possible
Purpose	Premium option for imputed purposes	Premium option for imputed purposes	Not available due to ID persistence limitations	Generally not possible
Vehicle Class	Not available	Not currently available	From some providers Heavy & Medium trucks, private vehicles	Generally not possible
Resource Requirements & Availability				
Data Cost	Intermediate	Expensive	Inexpensive to Expensive depending on provider, amount/length of data period, and amount of processing included	Expensive
Additional Processing Required	Intermediate	Limited to Intermediate	Substantial to Limited depending on provider	Usually included in price
Vendors	AirSage	Streetlight, Cuebiq	ATRI, Streetlight, Inrix, TomTom, HERE	TTI, RSG, others

This table presents generalizations at the time of presentation which may or may not obtain for particular regions or in the future

The Power of Big Data

The Power of Big Data

TN STATEWIDE DATA

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

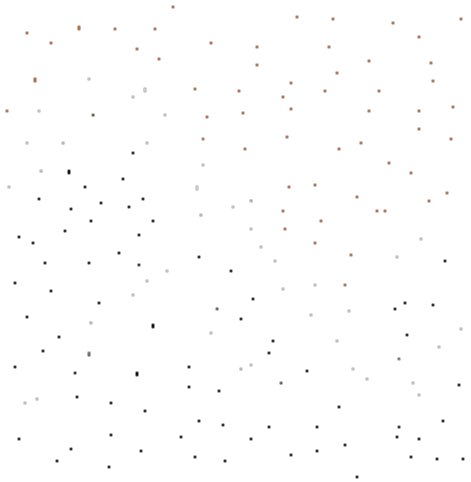
- Trip Table (OD pairs)
  - Total: 12,744,900
  - Survey: 39,782 0.3%
  - AirSage: 3,355,539 26.3%

CHATTANOOGA DATA

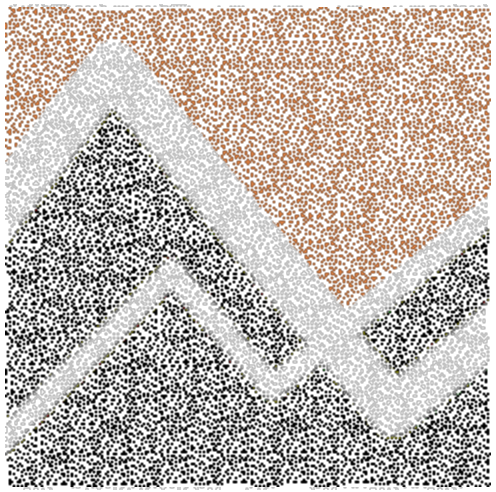
- 2010 household survey
  - 1,502 households
- AirSage and ATRI datasets

- Trip Table (OD pairs)
  - Total: 529,984
  - Survey: 8,350 2.0%
  - AirSage: 182,742 34.5%

Can you recognize the pattern based on <2%?



How about based on >25%?



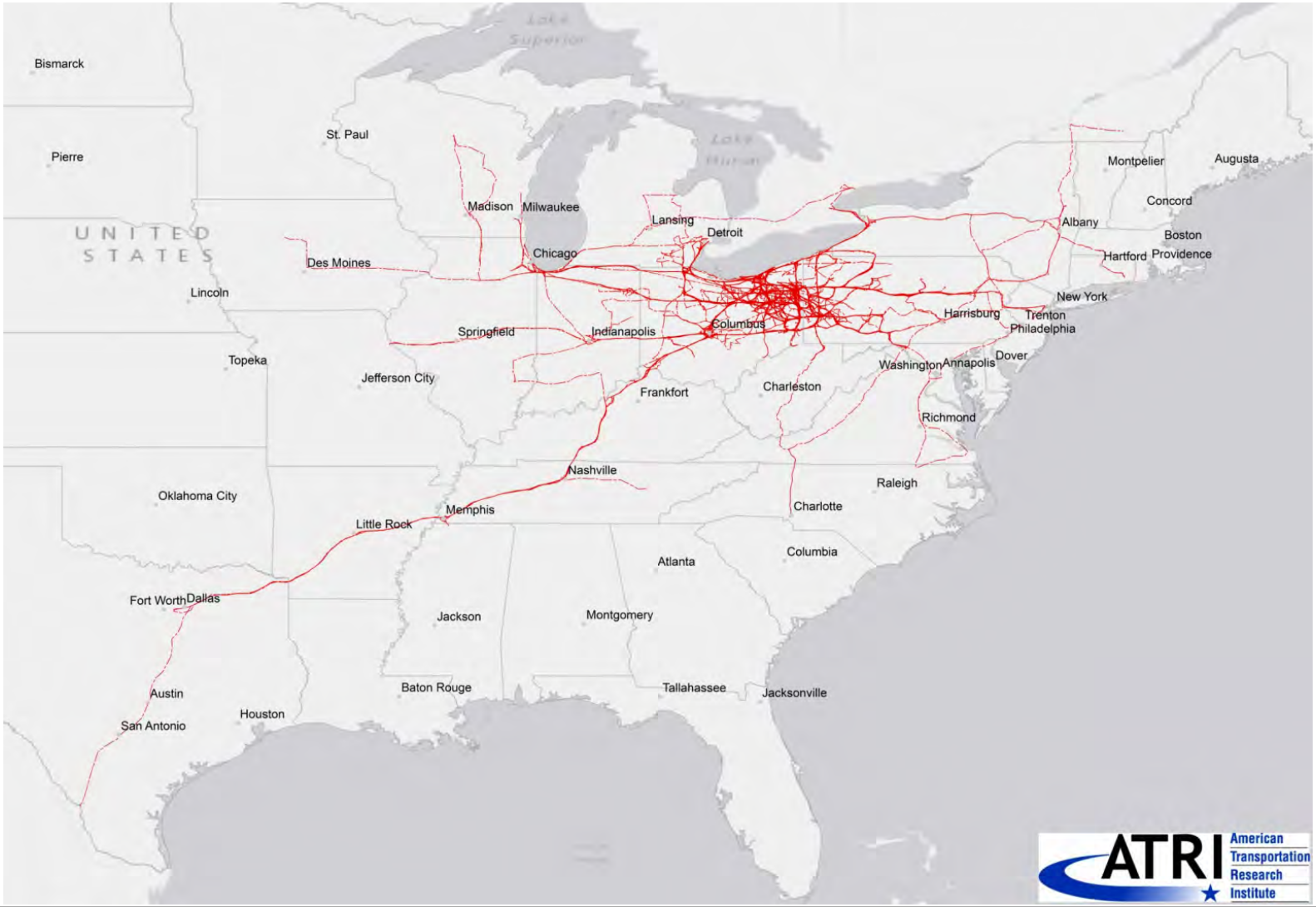
Big Data allows us to see the Big Picture

US 30 Study Area

Trucks using the US 30 corridor – after 1 Day

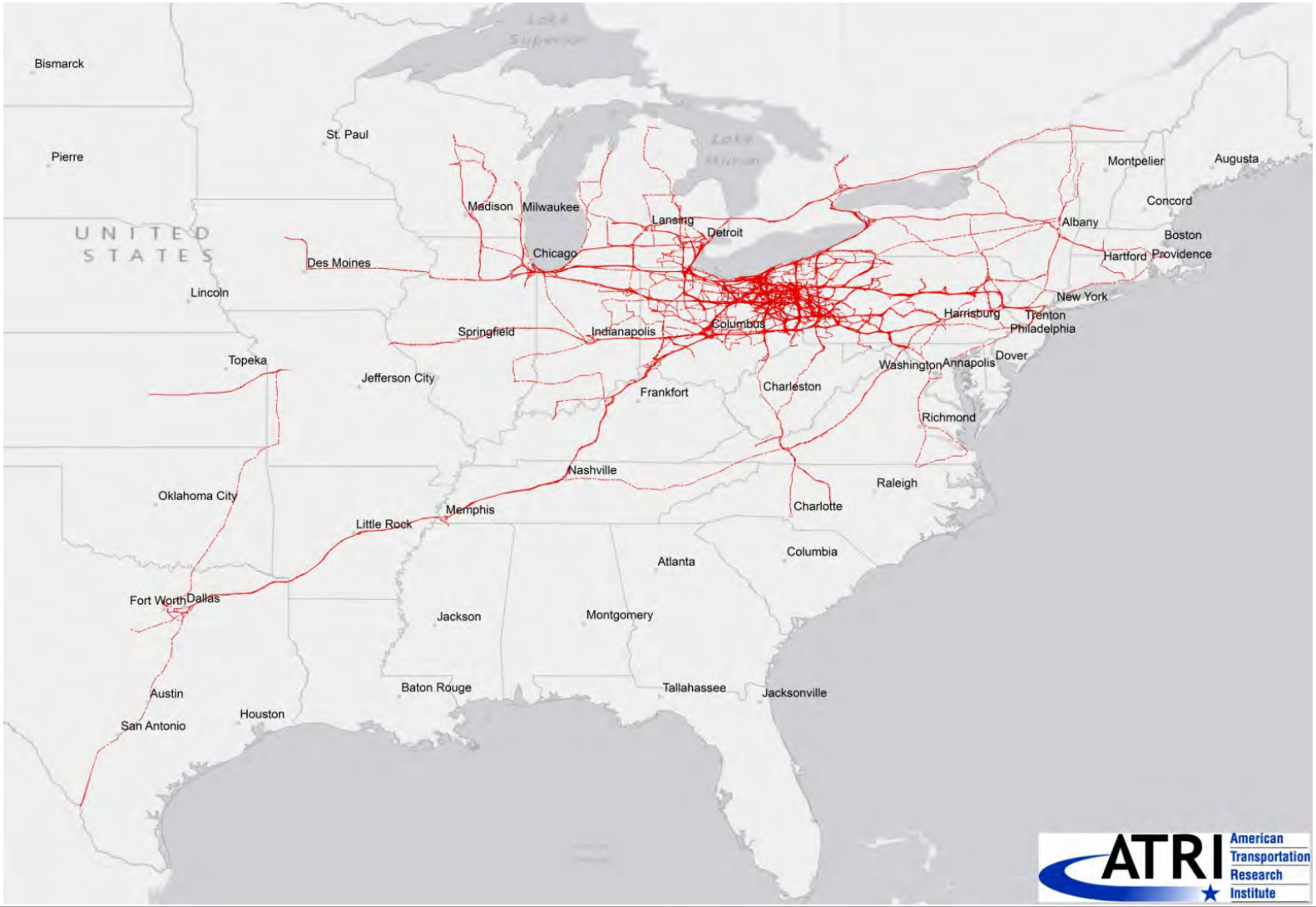
Trucks using the US 30 corridor – after 2 Days

Trucks using the US 30 corridor – after 3 Days



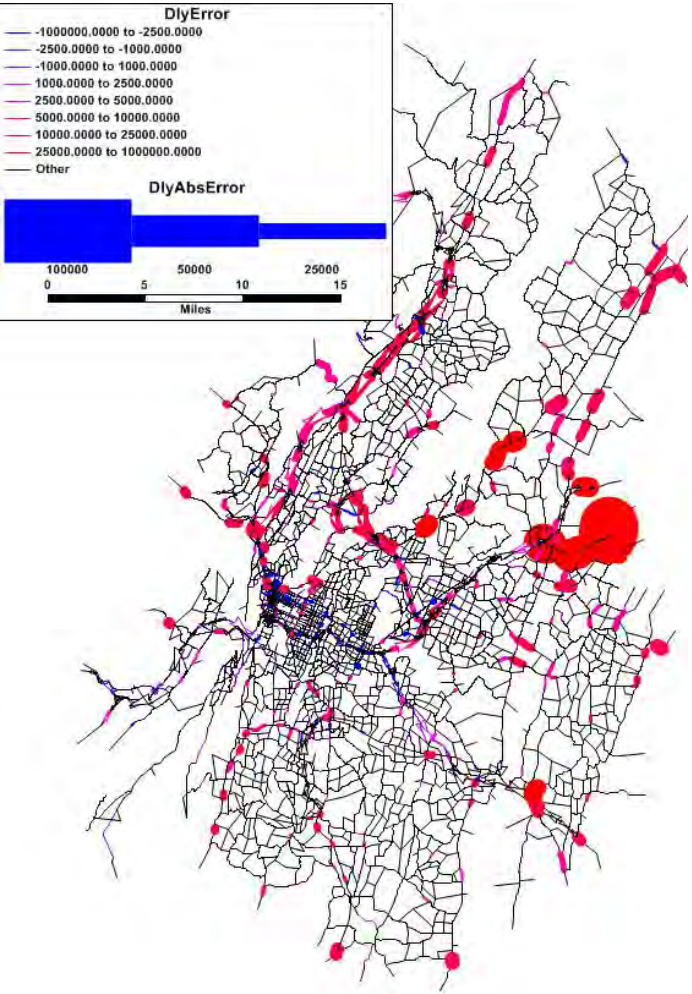
Trucks using the US 30 corridor – after 5 Days

Trucks using the US 30 corridor – after 7 Days

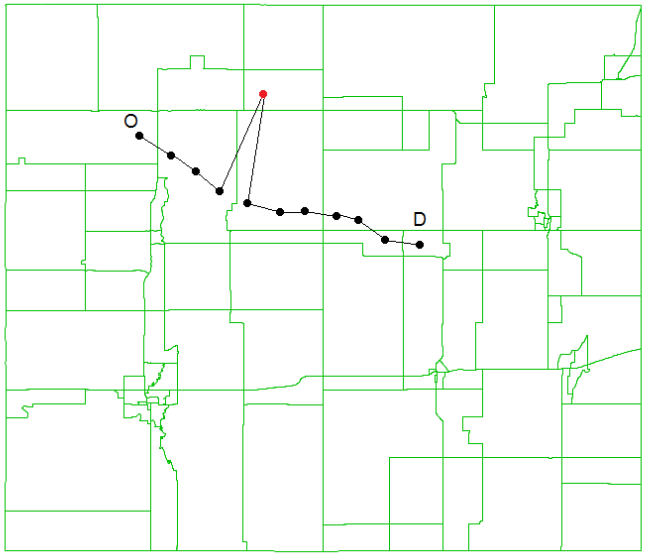


The Limitations of Big Data

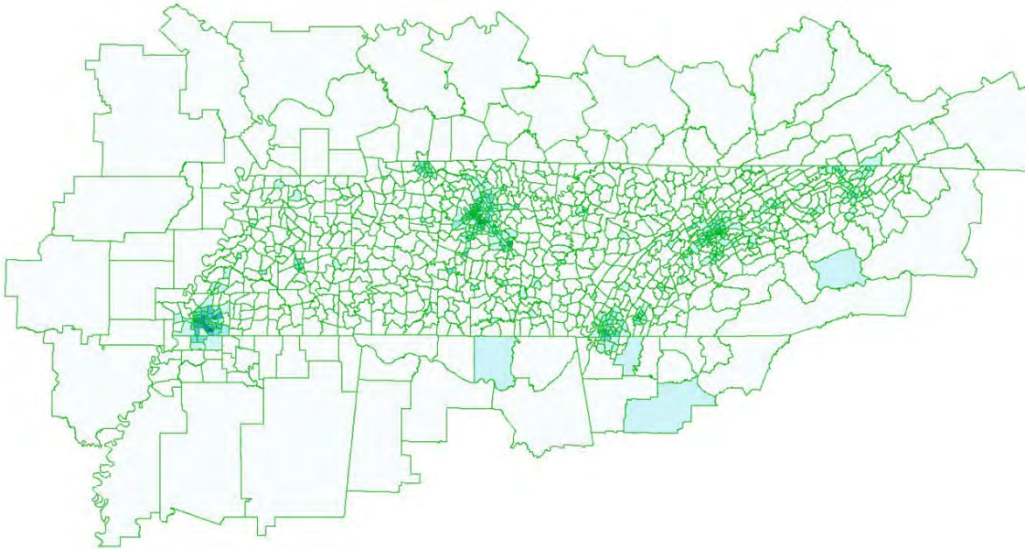
Cleaning Required

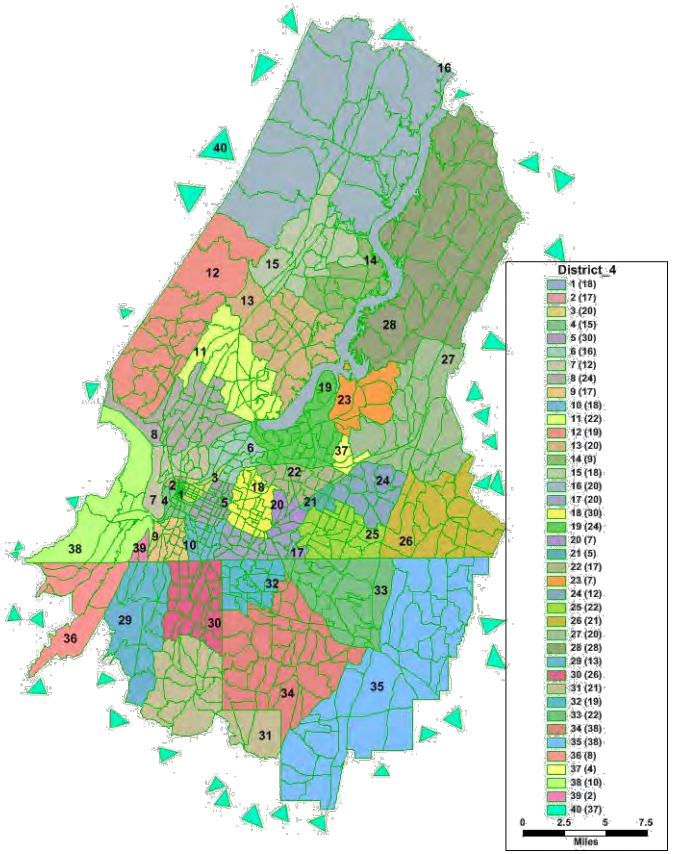


- Filtering / cleaning
  - Needs vary by data source – but all need it
  - GPS jumps/blips and equivalent
  - Missing data



Limited Spatial Resolution

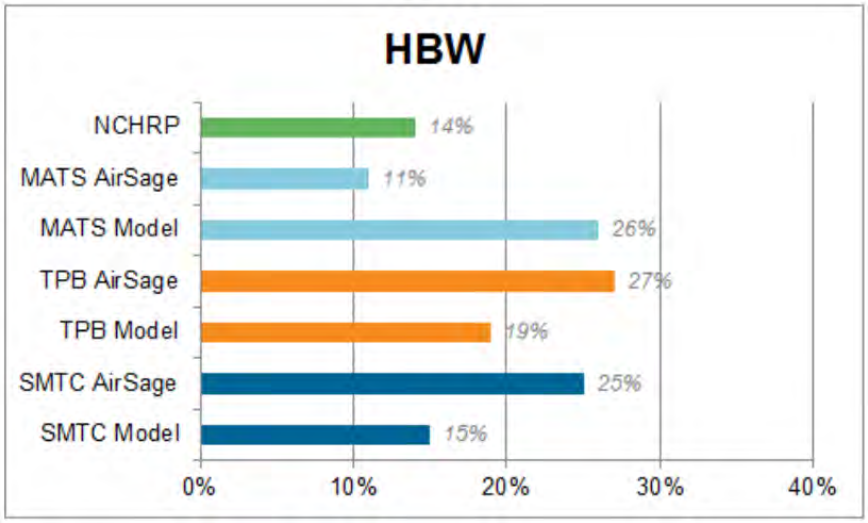


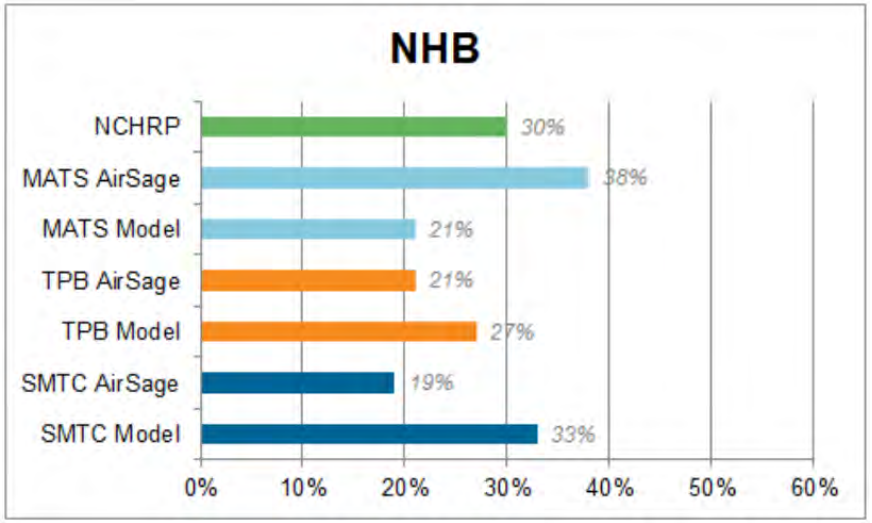


- Accuracy and Cost
  - Cell-based has limited precision
  - Pricing based on number of zones / districts

No Purpose or Mode

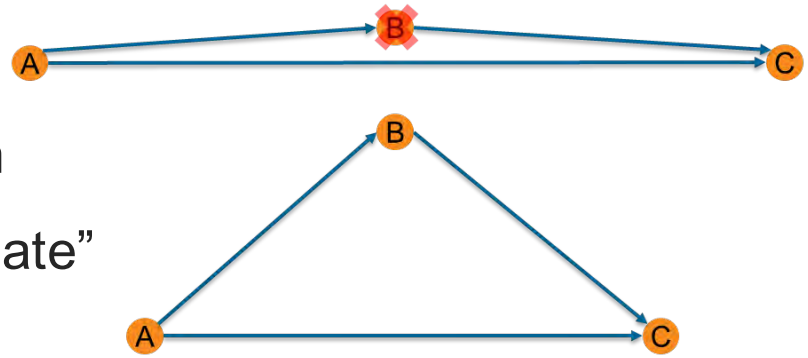
- Just ODs – not a survey substitute
  - Imputation can't reproduce surveys (so far)
  - GPS ID persistence prohibitive
  - Better to supplement with CTPP / LEHD

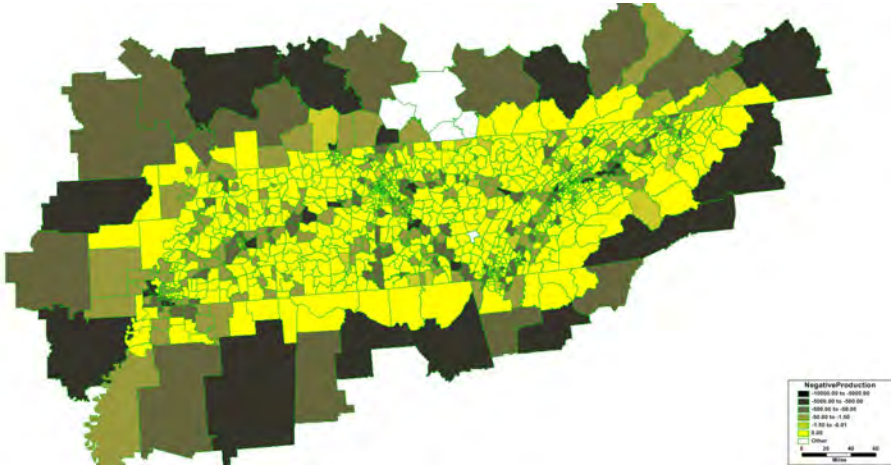


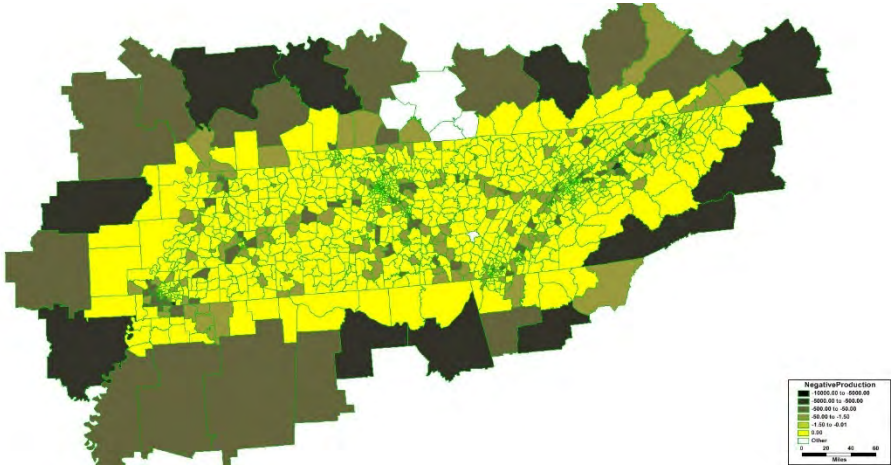


Trip Definitions

- Combining AirSage and ATRI
  - Need consistent trip/stop definition
  - Whether or not to count “intermediate” stops and break up long trips



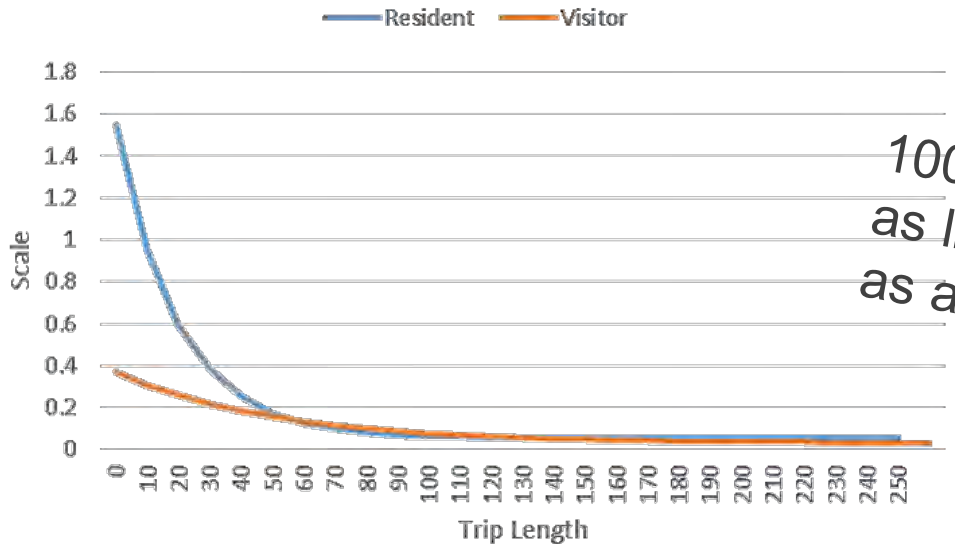




- Filtering short stops eliminated 87% of inconsistencies
  - From 11% of cells & 0.20% of trips to 1% of cells and 0.09% trips

Not Representative

- Big Sample NOT Random Sample
  - Locational biases, holes
  - Trip length / duration biases
  - Not corrected by penetration-based expansion



100 mi trip is 12 times as likely to be detected as a 10 mi trip

Data Expansion

Types of Expansion Methods

- Multiple methods commonly used together
- Seven methods (**bold**) in known use:
  - *Non-Traffic Count-based Sample Penetration Methods*
    1. **Market Penetration-based**
    2. **Trip Generation-based**
  - *Traffic Count-based Methods*
    3. **Simple Single-factor Scaling**
    - *Multi-factor Scaling*
      4. **Matrix Partitioning / Iterative Screenline Fitting**
      - *Network Assignment-based*
        5. **Parametric Scaling**
        - *Non-Parametric*
          6. **Direct ODME**
          7. **Indirect ODME**

TDOT used 1, 3, 5, & 6

Chattanooga used 1, 2, 3, 4

Pros and Cons of Expansion Methods

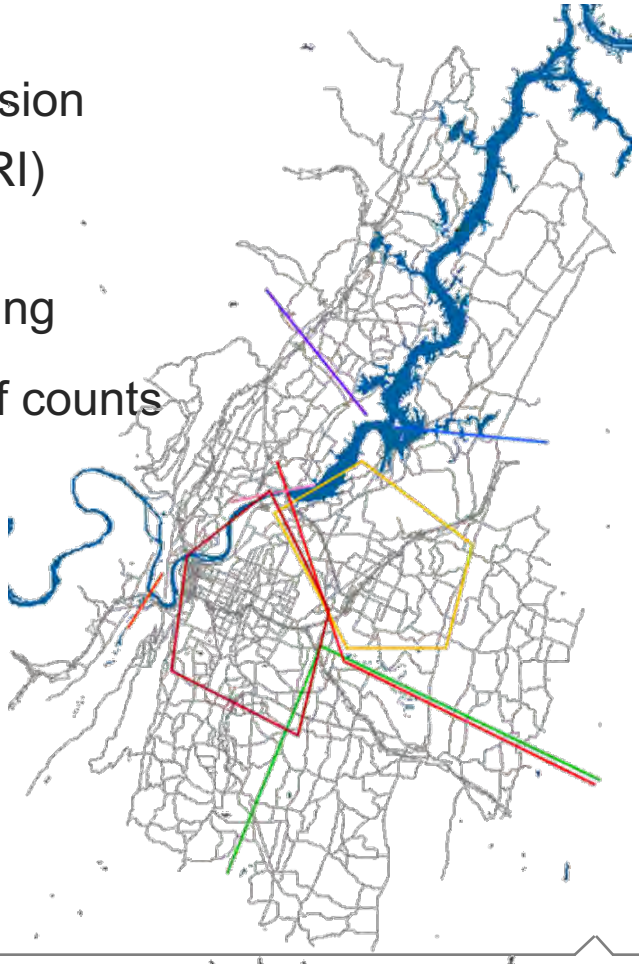
- Methods often combined to address multiple issues
- Level of effort and significance of biases vary

		Fix Trip Length Bias	Fix Coverage Problems		Independent of Network	Ease of Application	Holdout Count Sample	Transparency
1	Market Penetatraion-based	✗	✓		✓	✓	✓	-
2	Trip-Generation-based	✗	✓		✓	-	✓	✓
3	Single-factor Scaling	✗	✗		✓	✓	-	✓
4	Matrix Partitioning / Iterative Screenlines	✓	✓		✓	-	✓	✓
5	Parametric Scaling	✓✓	✗		✗	✗	-	-
6	Direct ODME	✓	✓		✗	✓	-	✗
7	Indirect ODME	✓	✓		✗	✗	-	-

Chattanooga Expansion Adjustments

FOUR-STEP ADJUSTMENT

- How best to expand to traffic counts?
  1. AirSage's Market Penetration-based Expansion
  2. Trip-Generation-based filling of "holes" (ATRI)
  3. Single-factor Scaling
  4. Matrix Partitioning / Iterative Screenline Fitting
- No reliance on network model; holdout sample of counts



TDOT Expansion Adjustments

ANOTHER FOUR-STEP ADJUSTMENT

- How best to expand to traffic counts?
  1. AirSage's Market Penetration-based Expansion
  2. Single-factor Scaling
  3. Parametric Scaling – fit distance-based adjustment factor curves for residents and non-residents
  4. Non-parametric – used ODME for residual adjustments
- Avoid massive ODME adjustments, provide explanation/understanding of bias and correction

Parametric Adjustment

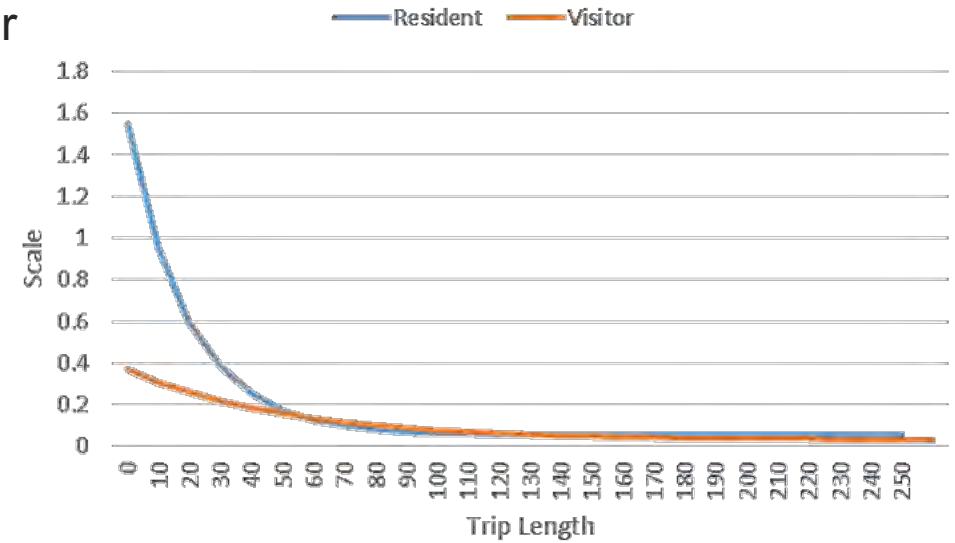
RESIDENT TRIPS

- Scale = 0.0612 + 1.6404*Exp(-0.05071*Length)
- Implication: **100 mi trip is 12 times as likely to be detected as a 10 mi trip**

Statistic	Value
Overall Percentage of Error	-0.5%
Urban Percentage of Error	-2.2%
Rural Percentage of Error	5.2%

VISITOR TRIPS

- Scale = 0.02920 + 0.3376*Exp(-0.01951*Length)
- AirSage now updating default Visitor expansion method to be consistent with Resident method
- Visitors are already long distance travelers – may be more likely to have cell phones / higher auto occupancy



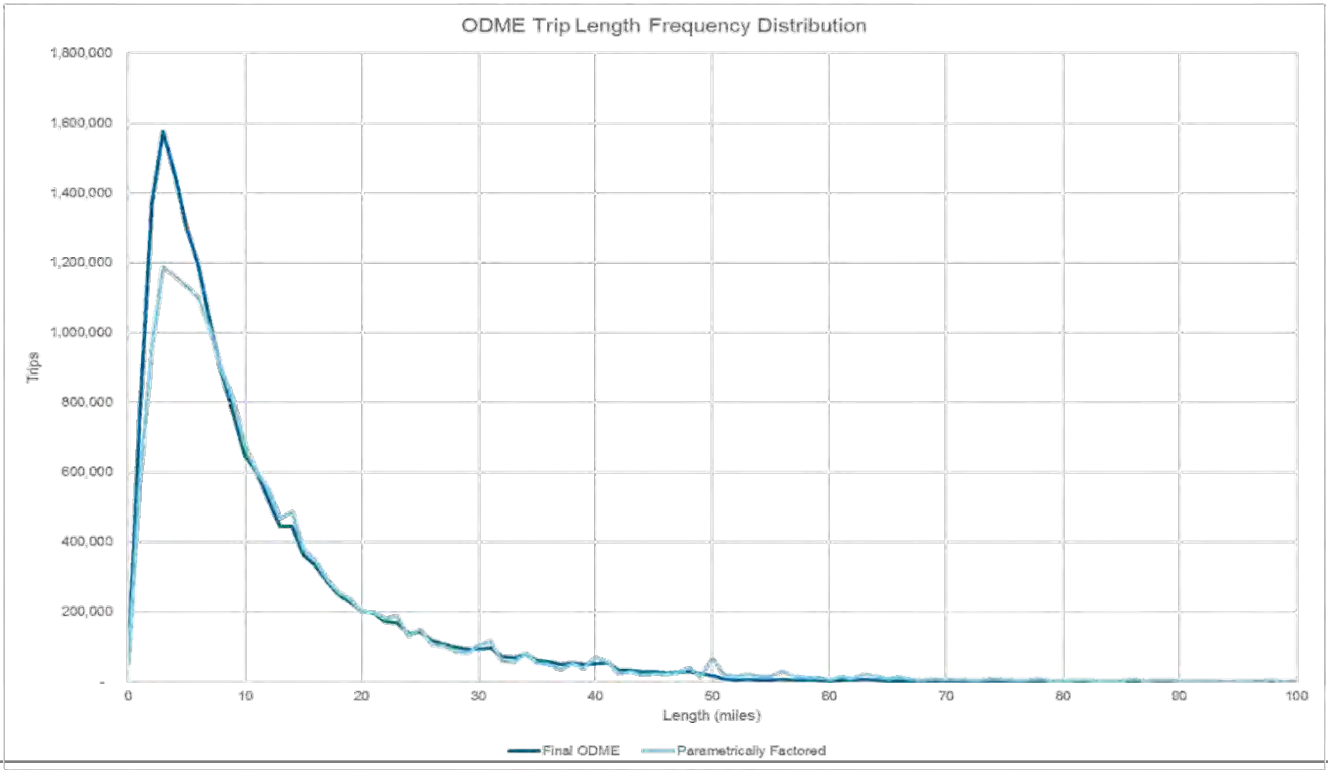
Non-Parametric Adjustment (ODME)

CONTROLS

- Minimum factor 0.5
- Maximum factor 5.0
- Only 10 iterations

RESULTS

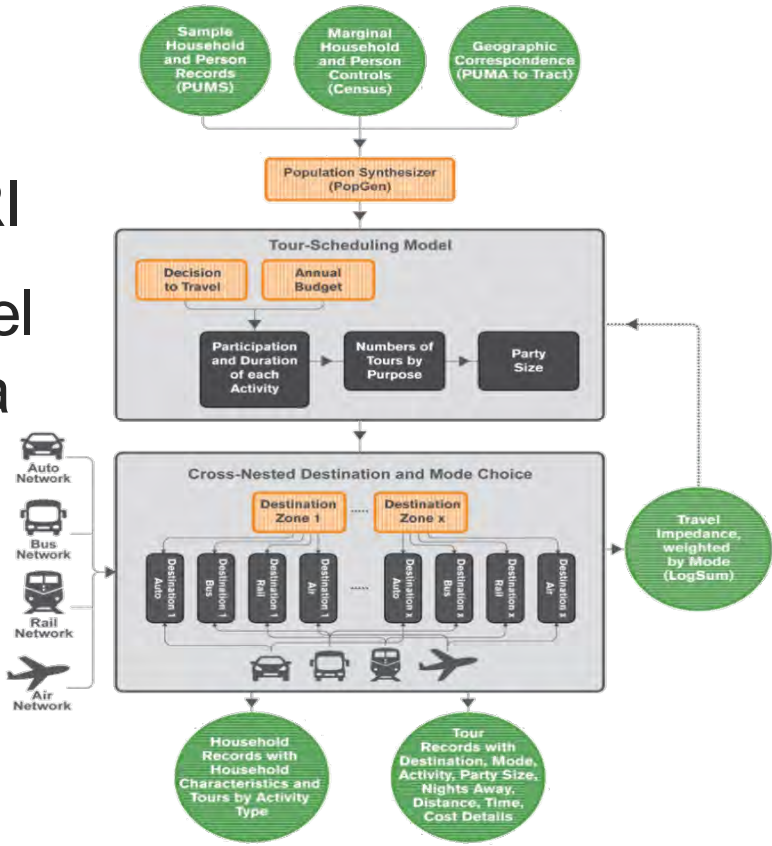
- Matrix MAPE 4.5%
- RMSE vs. counts from 55.4% to 36.1%
- Some additional increase in short trips

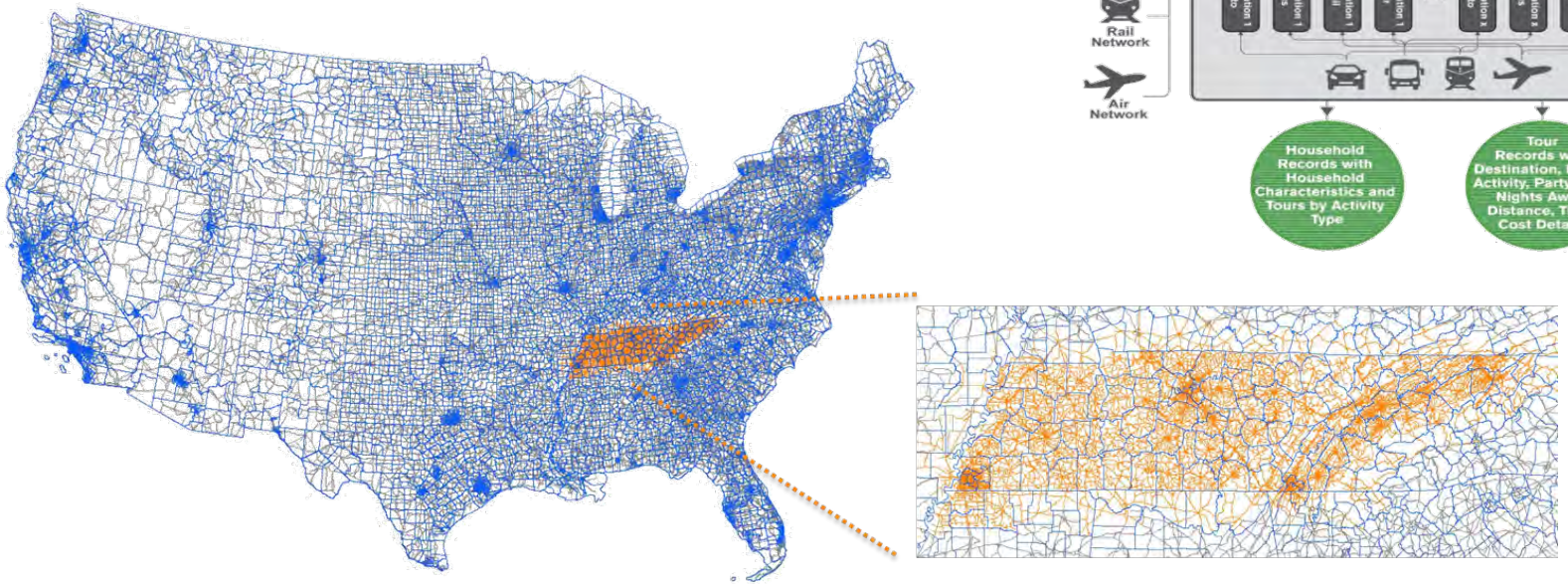


Model Results with Big Data

TN Long Distance Trips

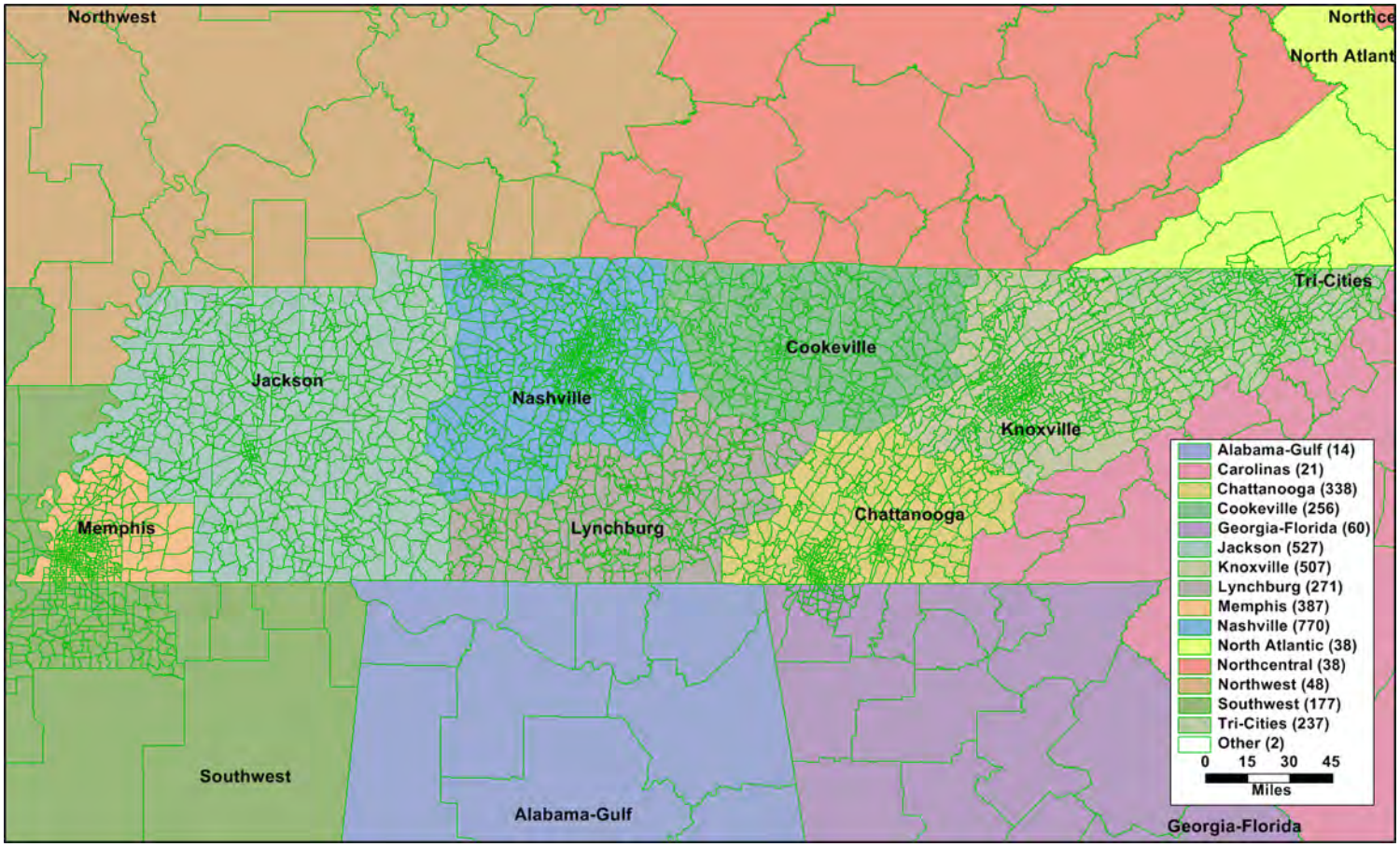
- Modeled Trips pivot off AirSage/ATRI
- FHWA National Long Distance Model Calibrated to Regional AirSage Data
  - Psychological bias against state border crossings





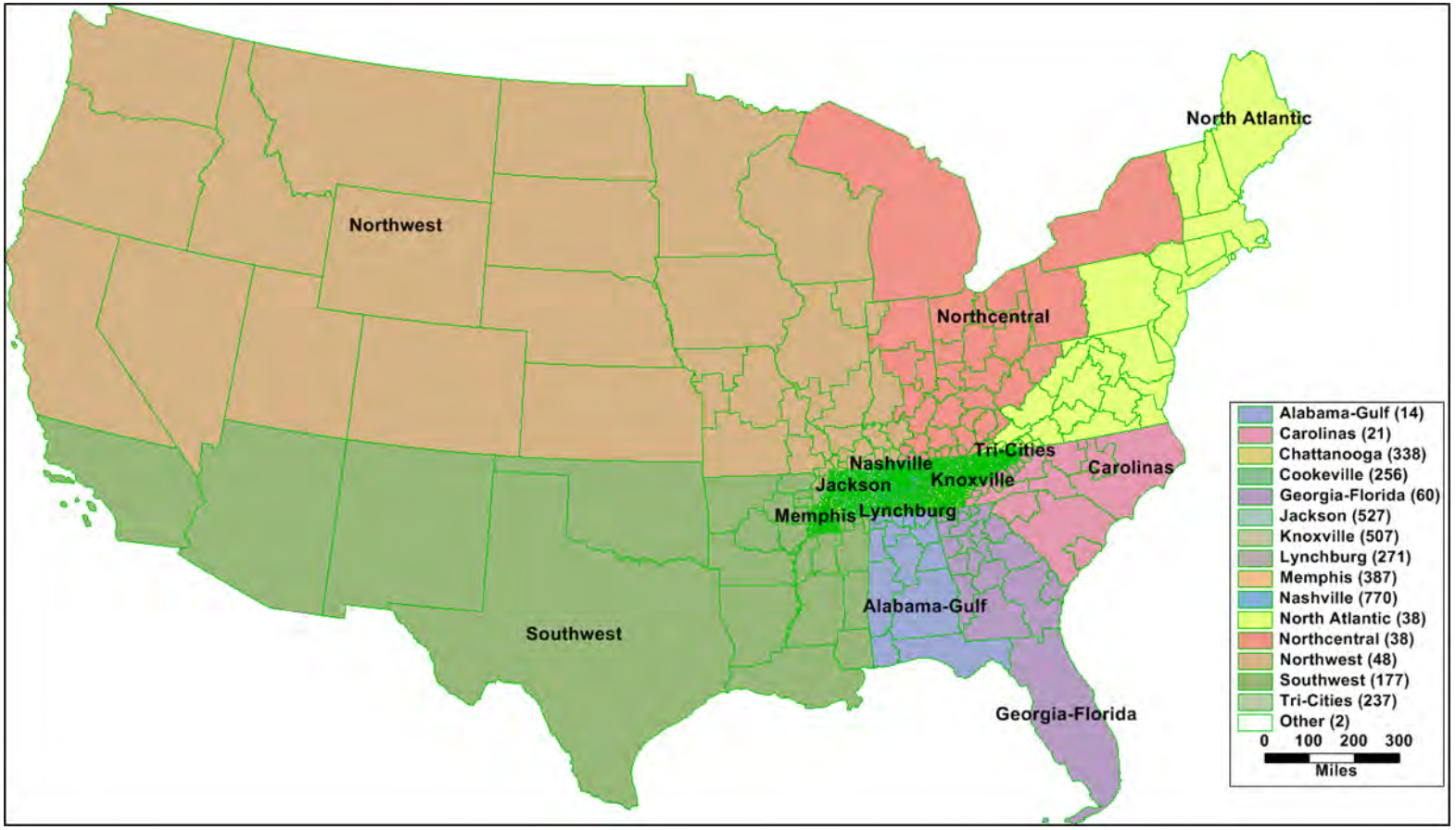
TN Internal Districts

DISTRICT SCHEME (INTERNALS)



TN External Districts

DISTRICT SCHEME (EXTERNALS)



Within TN Trip Distribution

DISTRICT-TO-DISTRICT COMPARISON

- Importance of comparing actual OD pattern, not just TLFD
- Very good agreement, overall
- District level origins & destinations all within 1.5%
- District level ODs all within 2% except within Nashville district

Relative Percentage Difference (Model Version 3 vs AirSage) I-I Trips

Origin districts	Destination districts								
	Tri-Cities	Knoxville	Chattanooga	Cookeville	Lynchburg	Nashville	Jackson	Memphis	***Total***
Tri-Cities	0.8%	0.3%	0.0%	0.0%	0.0%	-0.3%	0.0%	0.0%	0.7%
Knoxville	0.6%	2.0%	0.4%	-0.1%	0.0%	-1.5%	-0.3%	-0.3%	0.9%
Chattanooga	0.0%	0.1%	0.7%	0.3%	0.2%	-0.5%	-0.2%	-0.2%	0.4%
Cookeville	0.0%	-0.1%	0.4%	0.1%	0.0%	-0.6%	-0.1%	-0.1%	-0.3%
Lynchburg	0.0%	-0.1%	0.2%	0.0%	0.1%	-1.3%	0.1%	0.0%	-0.9%
Nashville	-0.3%	-1.4%	-0.2%	-0.9%	-1.5%	6.1%	-0.3%	-1.2%	0.4%
Jackson	0.0%	-0.3%	-0.1%	-0.1%	0.1%	-0.3%	-0.1%	0.8%	0.0%
Memphis	0.0%	-0.3%	-0.2%	-0.1%	0.0%	-1.1%	0.3%	0.1%	-1.2%
Total: All	1.0%	0.3%	1.1%	-0.8%	-1.0%	0.5%	-0.5%	-0.7%	0.0%

To/From TN Trip Distribution

DISTRICT-TO-DISTRICT COMPARISON

- Generally good agreement
- District level origins & destinations all within 10%, most within 3%
  - Smoky Mtns not attracting enough to/from Knoxville
- District level ODs all within 4% except within Nashville - Northcentral

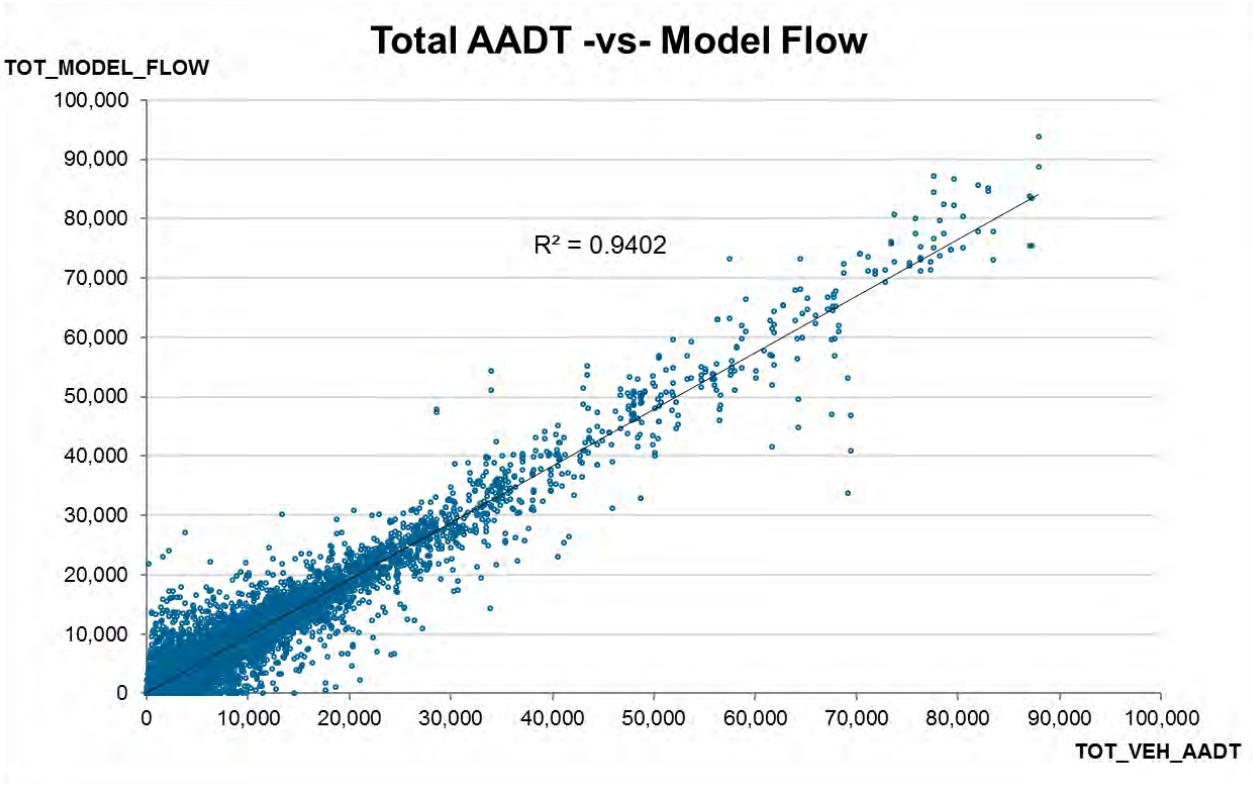
Relative Percentage Difference (Model Version 3 vs AirSage) I-E & E-I Trips

Internal districts	External districts							
	Northwest	North Atlantic	Northcentral	Carolinas	Alabama-Gulf	Southwest	Georgia-Florida	***Total***
Tri-Cities	0.4%	0.1%	0.8%	3.6%	0.0%	0.2%	0.3%	5.3%
Knoxville	0.5%	-2.6%	-1.2%	-1.7%	-0.7%	0.3%	-2.0%	-7.3%
Chattanooga	0.0%	-0.1%	-0.5%	-0.4%	-1.1%	0.1%	2.7%	0.8%
Cookeville	0.0%	-0.2%	0.9%	-0.3%	-0.1%	-0.1%	-0.2%	0.0%
Lynchburg	-0.4%	0.1%	0.4%	0.0%	0.7%	-0.1%	-0.4%	0.2%
Nashville	-0.7%	-0.3%	6.6%	-0.8%	-3.6%	-2.3%	-2.0%	-3.1%
Jackson	0.0%	0.1%	0.6%	0.0%	0.0%	-1.9%	0.0%	-1.2%
Memphis	0.5%	0.3%	0.8%	0.1%	-0.1%	3.4%	0.3%	5.2%
Total	0.3%	-2.6%	8.4%	0.5%	-4.9%	-0.4%	-1.3%	0.0%

Assignment Validation

- Great fit - One of best statewide models in the country
- Used ODME – with constraints, (some other statewide models do to)

VOLUME RANGE	RMSE	TDOT TARGET
< 5,000	102.1%	101.4%
5,000 to 10,000	35.6%	56.3%
10,000 to 20,000	22.0%	51.4%
20,000 to 30,000	16.4%	35.7%
30,000 to 40,000	14.8%	32.0%
> 40,000	11.1%	12.2%
All	36.6%	60.0%

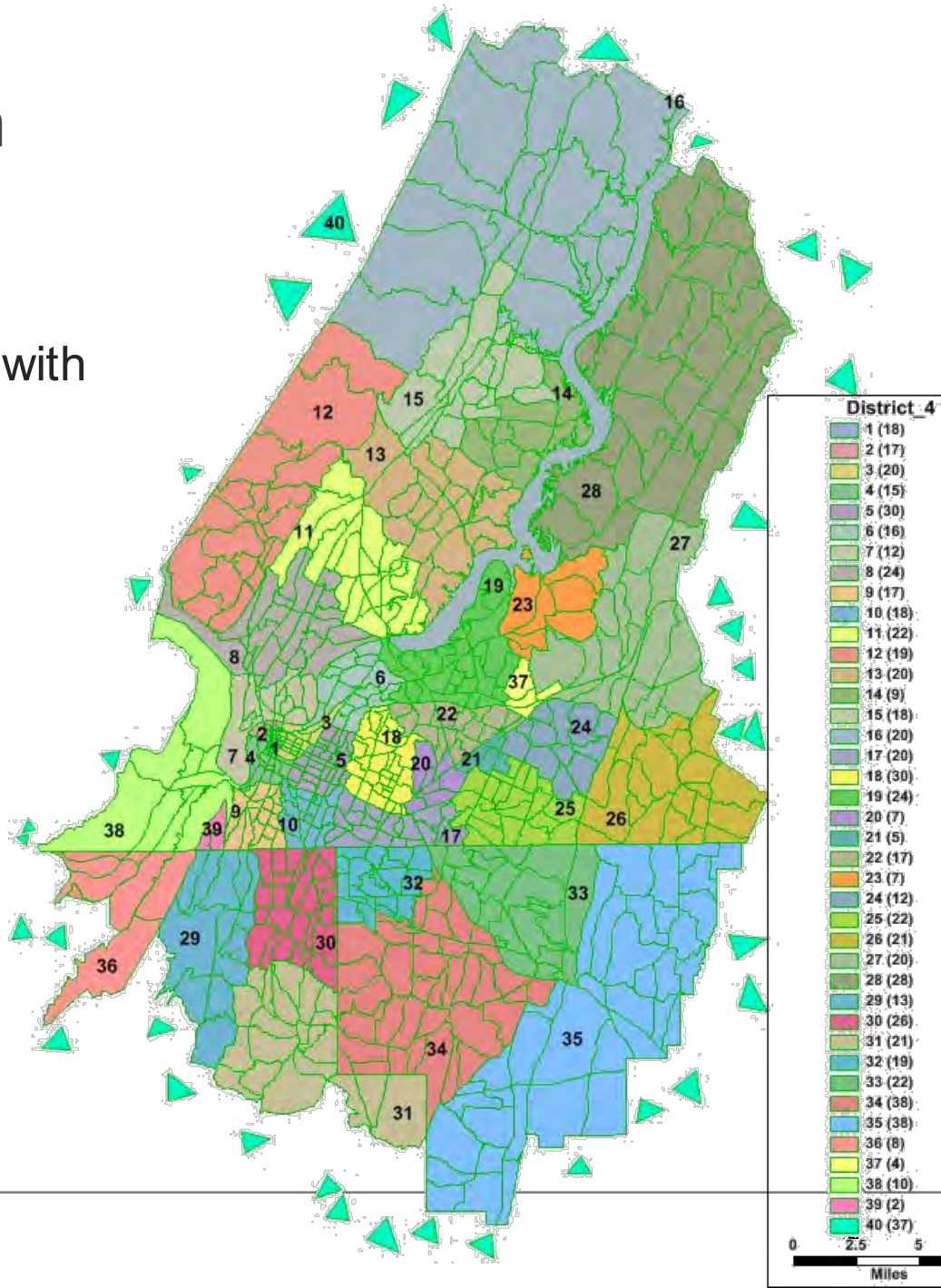


Chattanooga Daysim

- 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%



Total Daysim Trip Table vs. AirSage

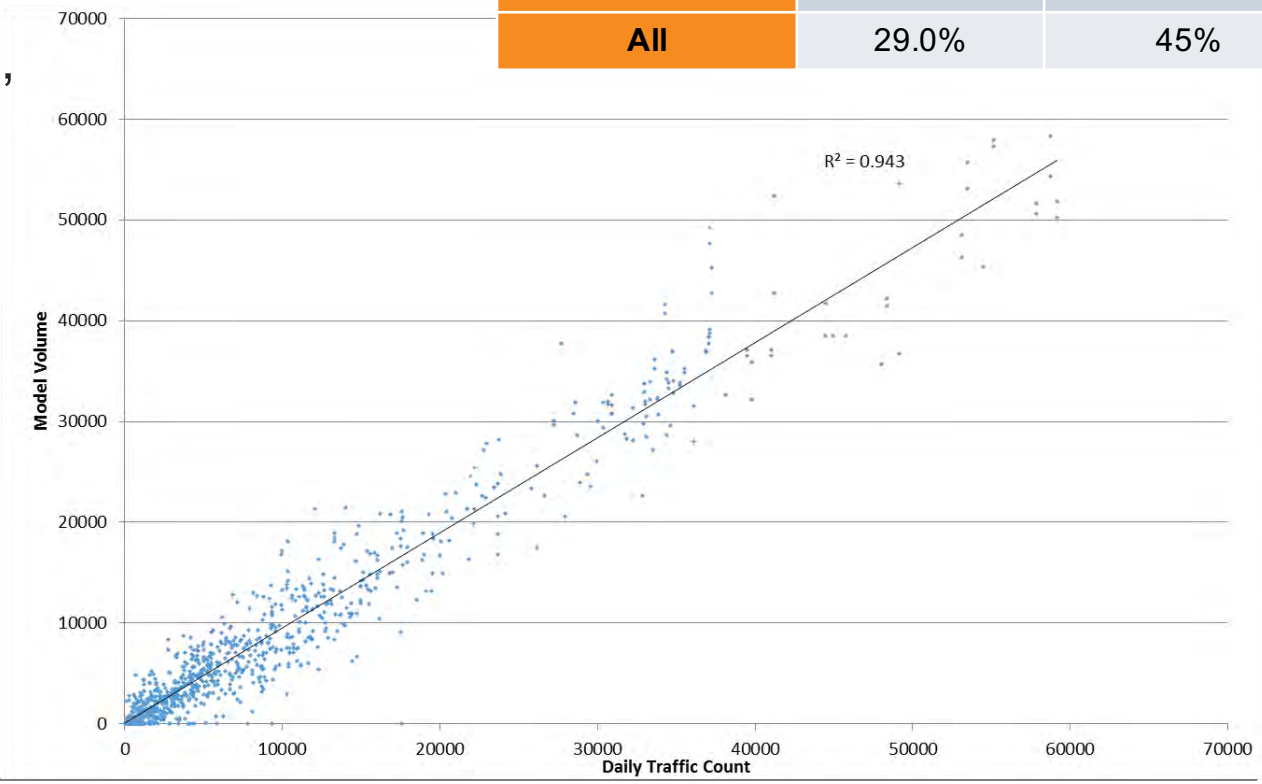
- Daysim vs. AirSage
  - Very good agreement – **10.5% RMSE**
  - All cells within +/- 1%
  - All residence/work Super Districts within +/-2.5%

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%

Assignment Validation

- Great fit!
  - Better than old model
  - Far exceeds TDOT standards
  - No ODME, only screenline factoring

VOLUME RANGE	RMSE	TDOT MAXIMUM
< 5,000	62.1%	100%
5,000 to 10,000	37.9%	45%
10,000 to 15,000	28.0%	35%
15,000 to 20,000	22.7%	30%
20,000 to 30,000	15.7%	27%
30,000 to 50,000	14.1%	25%
50,000 to 60,000	9.9%	20%
All	29.0%	45%



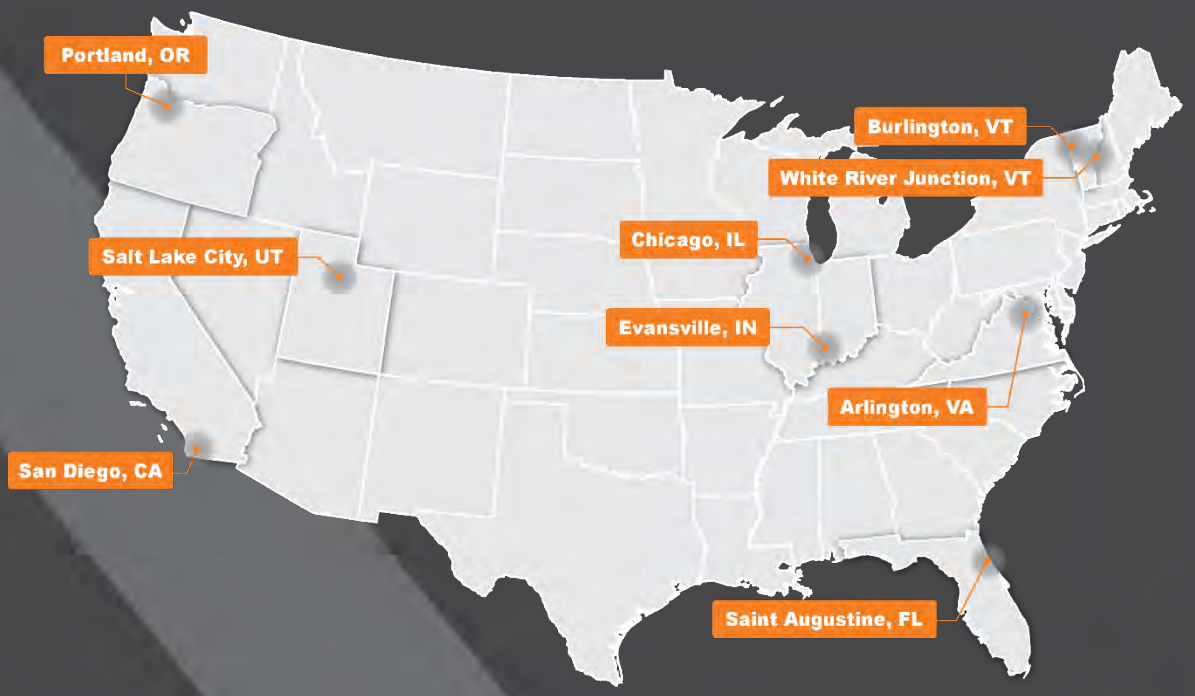
Final Thoughts

What's Next?

- Data Driven Forecasting
  - Pivoting, destination choice models with constants
  - Better accuracy, analog to STOPS
- Evolving Data & Methods
  - Big data may provide frequent info during transformational changes
  - New data sources entering the market
  - Data fusion: surveys & big data

- For more detail on much of the information in this presentation see forthcoming TMIP guide

TMIP
Better Methods. Better Outcomes.



Contact

www.rsginc.com

Vince Bernardin, Jr, PhD
DIRECTOR OF TRAVEL FORECASTING

Vince.Bernardin@rsginc.com
812.200.2351