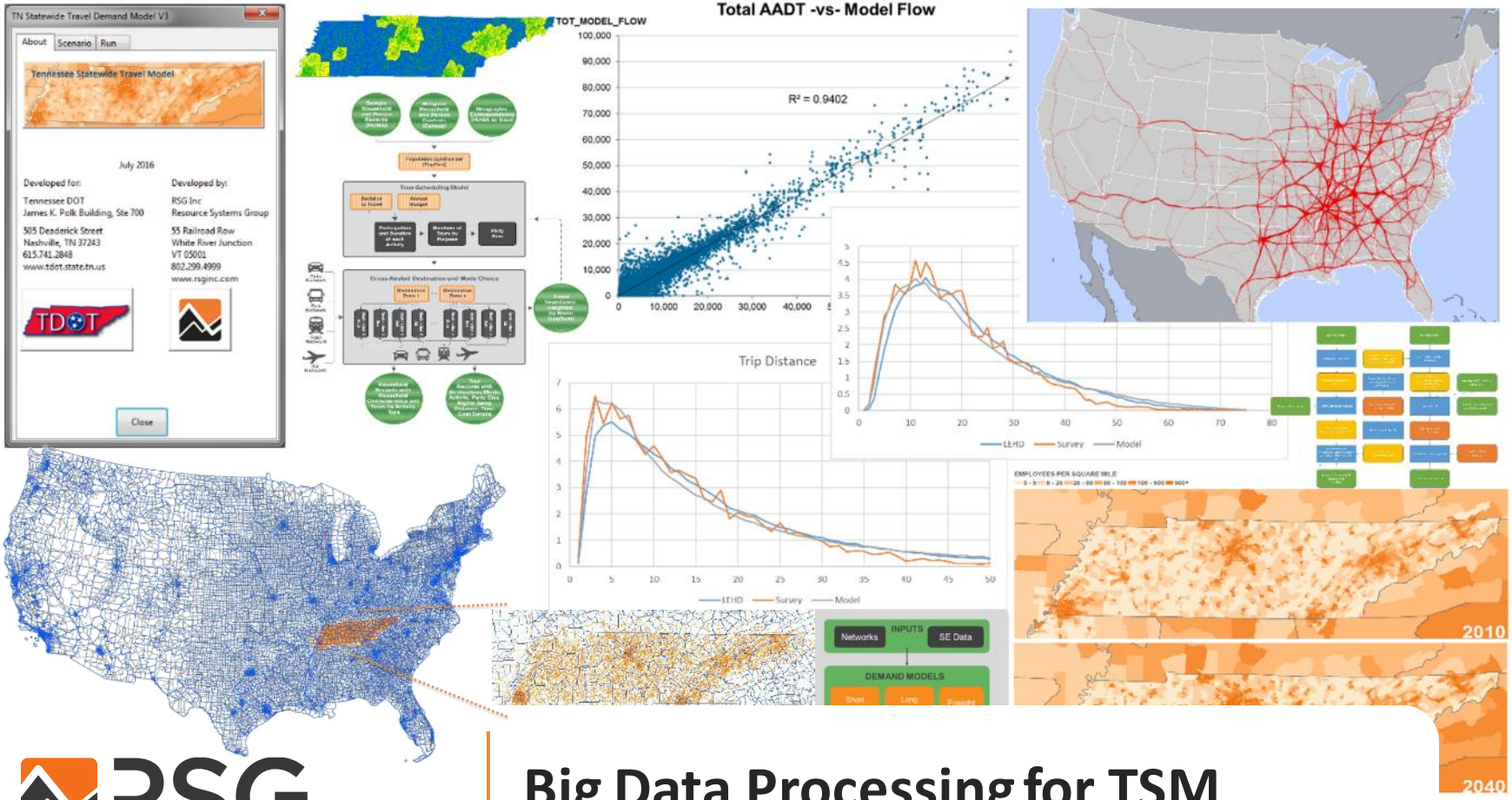


RSG
the science of insight

Big Data Processing for TSM

TNMUG Meeting

Steven Trevino & Hadi Sadrsadat, PhD

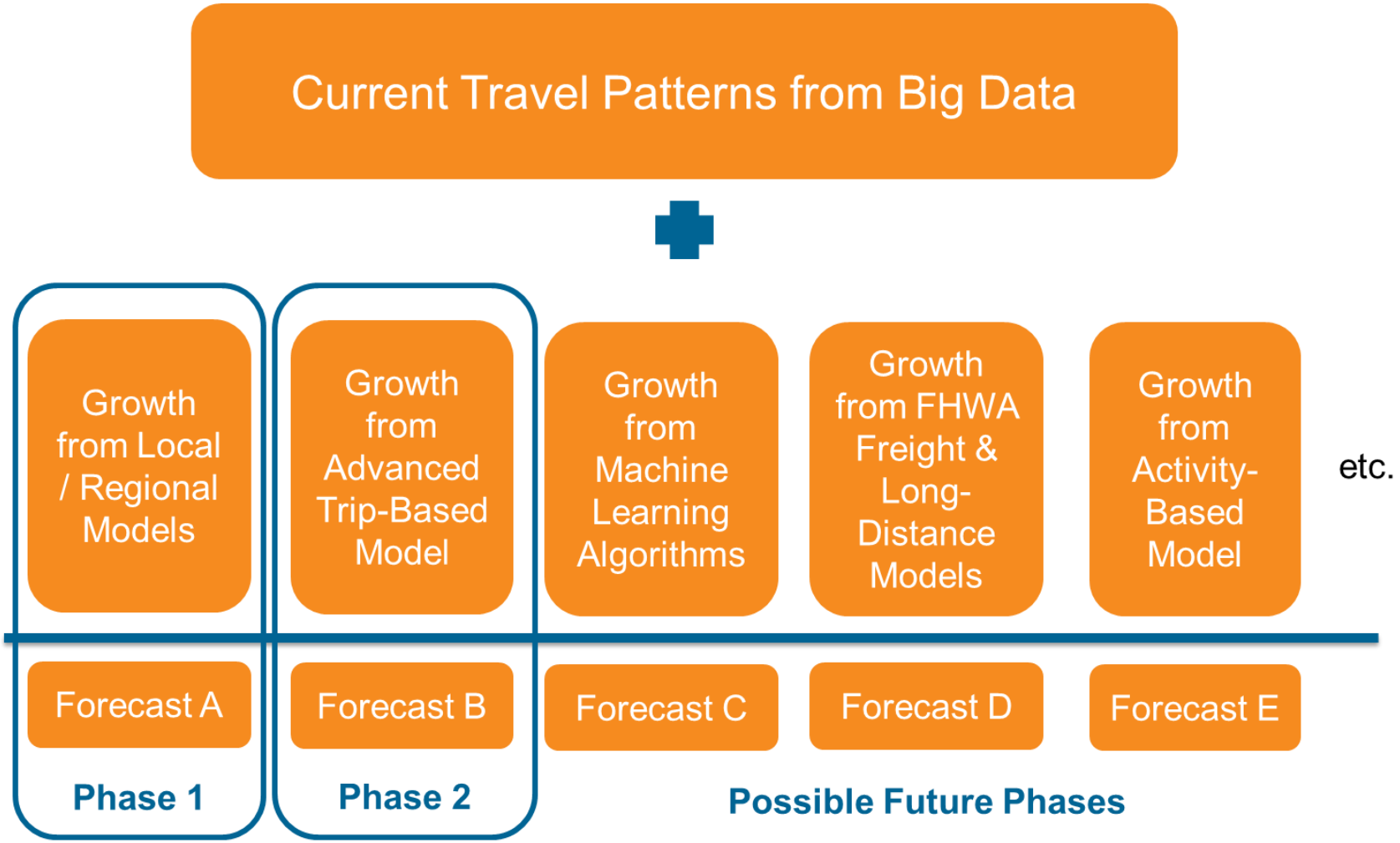
September 22, 2020

TSM Update Objectives

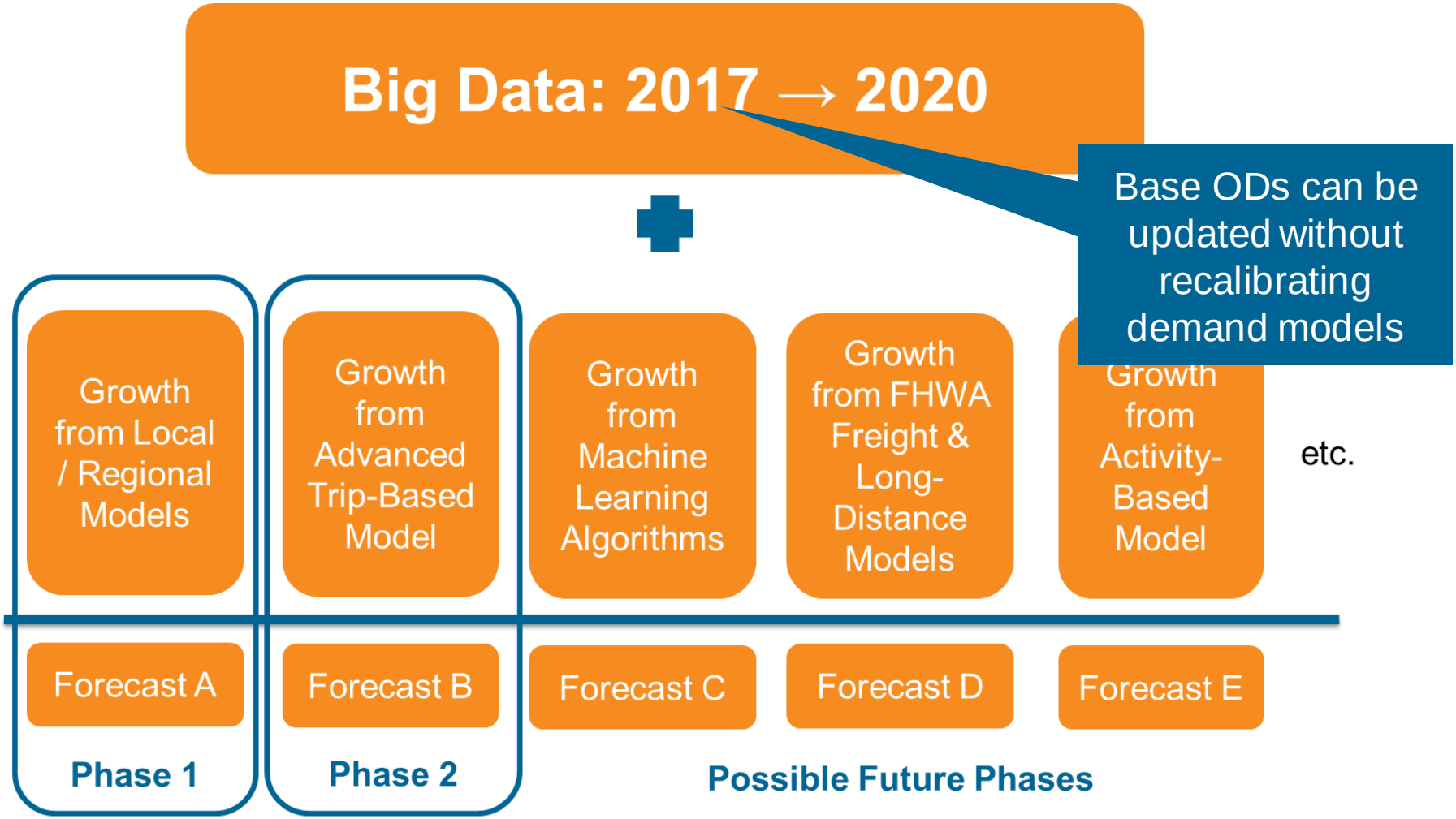
- Socioeconomic Data
- Network Updates
- General Model Enhancements
- 2018 Big Data
- Alternative Long-Distance Passenger Model
- Model Calibration & Validation
- CAV Framework

Background

ILSTDM's Extensible Modular Framework



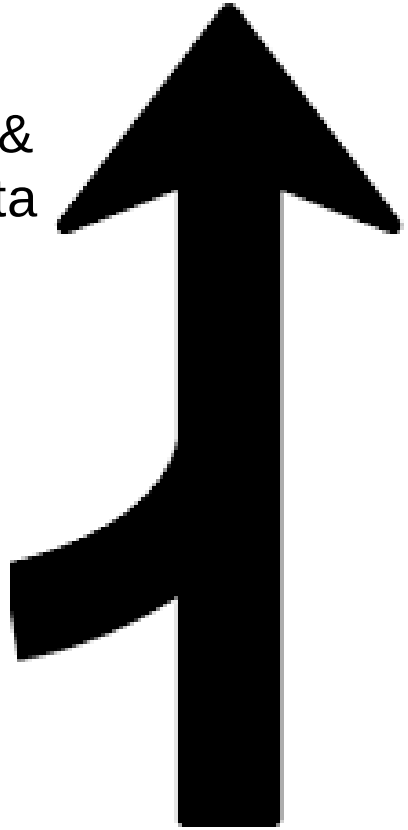
Updating Base without Recalibration



rMerge & LBS Data

What is rMerge?

rMerge is high-quality passive LBS data products & services enriched and validated with traditional data and grounded in RSG's deep expertise in travel behavior



RSG
the science of insight

How is rMerge Applied?

LBS data is reconciled, expanded, and validated against traditional data sources

Census

Traffic Counts

Travel Surveys

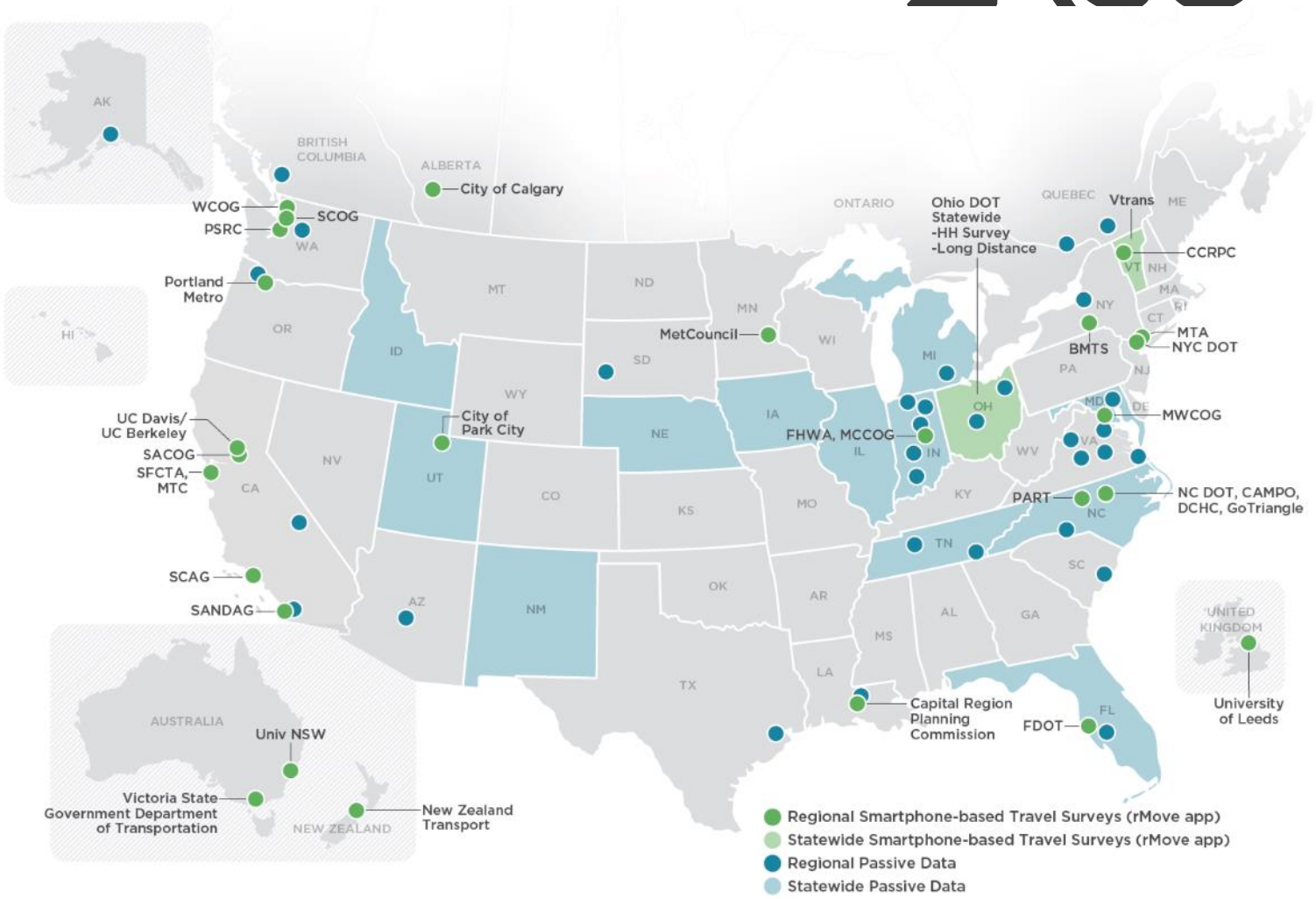
LBS data

Big data from smartphone apps is the primary raw data source from which rMerge is derived

Mobile Data Experience

RSG

nsight



Over 50 projects in over 25 states

How Big is this Big Data?

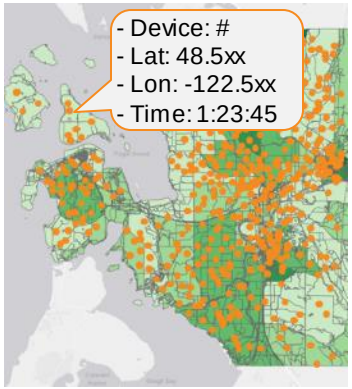
- 10-15% population on any given day (DAU)
- 50% of population over a month (MAU)
- ~ 3.0 million devices for TN during April 2019
- Larger sample than surveys or pure navigational GPS

How is Privacy Protected?

- Raw LBS data
  - Only identifying information is “ad-id”, which RSG replaces before processing
- Home & Work Locations
  - Necessary for:
    - Differentiating residents & visitors
    - Identifying trip purpose (e.g., home-based work)
    - Checking and correcting for demographic bias
  - RSG never reports info below the zone
  - RSG suppress/perturbs info for small zones
  - OD Aggregation prevents reassociation of data to individuals

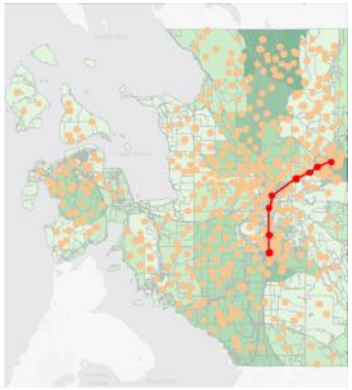
RSG's 4-step process for passive OD tables

1 PREPARE INPUT DATA



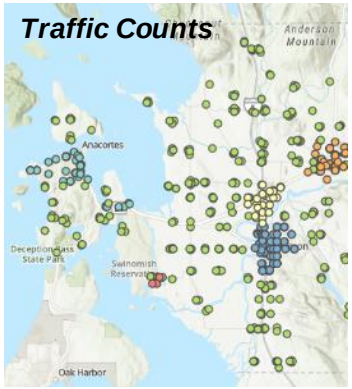
Billions of individual device location points from commercial LBS data are extracted, evaluated for basic metrics & cleaned*

2 IDENTIFY TRIPS



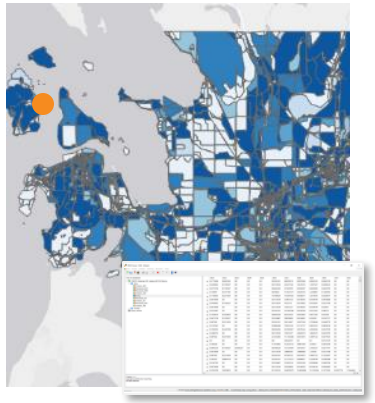
Points are clustered to identify stop locations, locations are classified (home, work, other) and linked to create trips

3 EXPAND TO REGION



Trips are expanded to region based on Census and traffic count data, surveys and other sources to provide representative O-D flows

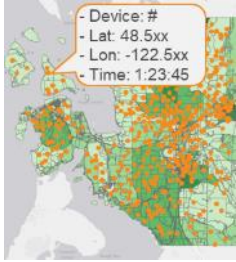
4 AGGREGATE & VISUALIZE



Trip data aggregated to OD matrices, with key dimensions (such as time period, visitor / resident) broken out

* Typically represents 10-15% of population per day, or 50%+ for one month of observations

Raw LBS input data collected & cleaned



1 PREPARE INPUT DATA | 2 IDENTIFY TRIPS | 3 EXPAND TO REGION | 4 AGGREGATE

a EXTRACT PINGS

- Select study period and geography (shapefile)
- Intersect shapefile with national LBS data to extract devices with at least one ping* in region

b COMPUTE METRICS

- Compute device level metrics on the extracted pings:
  - Total and average daily distance and number of pings
  - Number of unique coordinates, hours, and days
  - Time between pings and maximum speed

c QUALITY FILTERS

- Remove bad devices & pings with noise reduction filters
- Refine thresholds based on deep exploratory analysis

* ping is a latitude/longitude coordinate with a timestamp registered by a device

Tennessee LBS Data Summary

TN LBS Data: April 2019	
Sightings	2,798,272,340
Total Devices	2,968,506
Good Devices	1,042,720
Locations	13,052,093
Trips	56,319,378

- LBS data represents a sample of 6.3% of TN residents

Trips identified based on “stop” locations

1 PREPARE INPUT DATA | 2 IDENTIFY TRIPS | 3 EXPAND TO REGION | 4 AGGREGATE

a **CLUSTER PINGS**
- Remove pings with poor horizonal accuracy (>100 meters)
- Cluster pings using density-based algorithm
- Tag clusters as stopped vs. moving based on rolling window speed

b **CLASSIFY STOPS**
- Classify stopped clusters as “home”, “work”, or “other”
  - Based on recurring activity patterns, page-rank / node centrality metric, hours spent, and days seen at each cluster

c **BUILD TRIPS & QA/QC**
- Create trips by connecting successive dwells (visits to a cluster)
- Tag time periods
- Create plots, maps, and checks to validate trip output quality

Expansion process matches regional counts and Census data

1 PREPARE INPUT DATA | 2 IDENTIFY TRIPS | 3 EXPAND TO REGION | 4 AGGREGATE

BIG DATA EXPANSION?

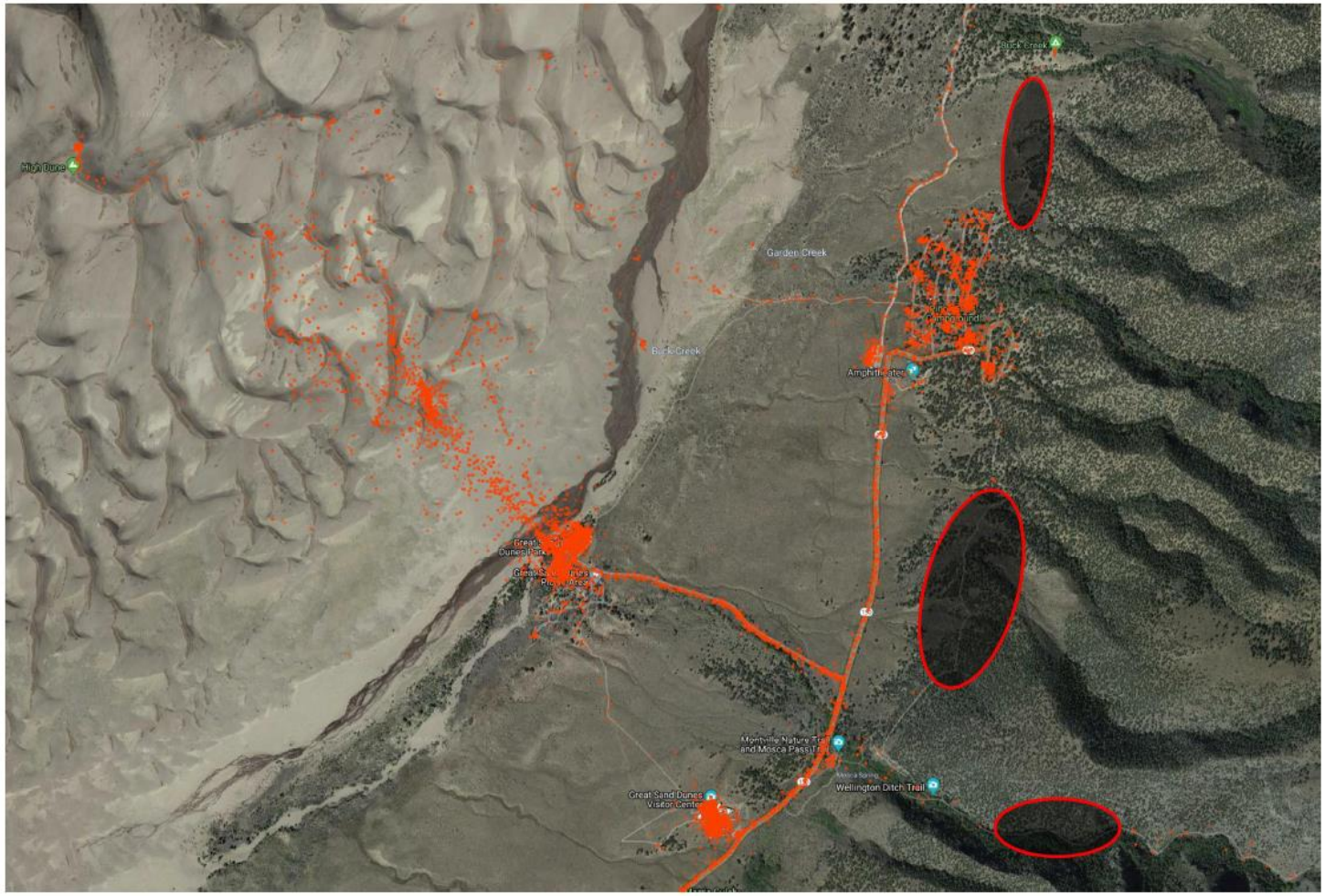
- Big data are large scale observations.
- But they are still only a sample of all travel.
- And they are NOT a random sample.
- Big data are known to have systematic **biases**.
- But if we can **measure** bias, we can **correct** for it.

What's Missing in Big Data?

- **Travelers**
  - Seniors & low income populations
- **Travel**
  - Geographic coverage
  - Short activities & trips
  - Other unknowns?

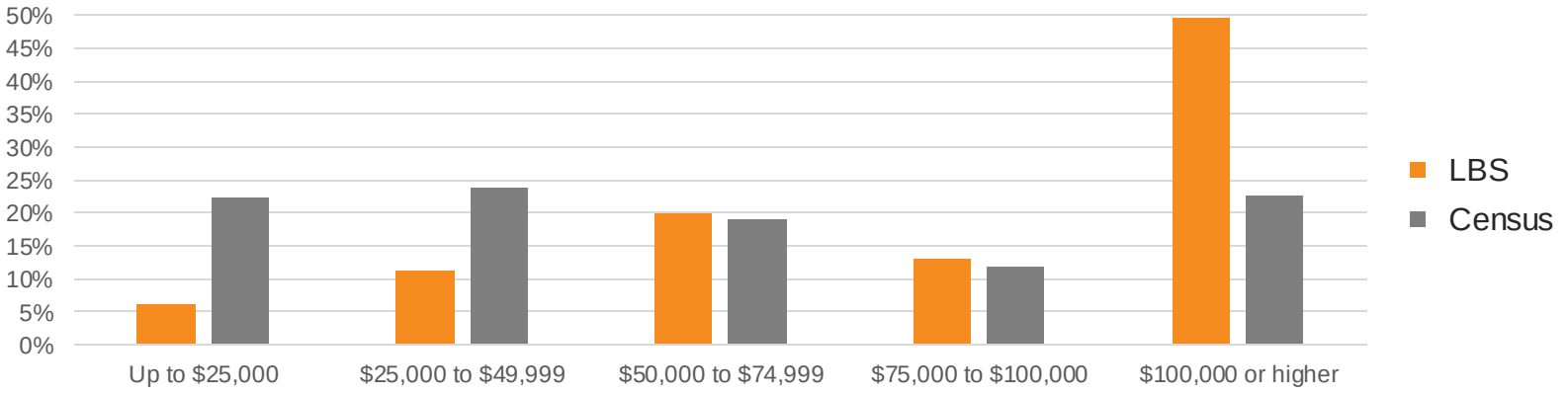
Geographic Coverage Gaps & Variations

SIGHTINGS AT GREAT SAND DUNES NATIONAL PARK IN JULY 2018

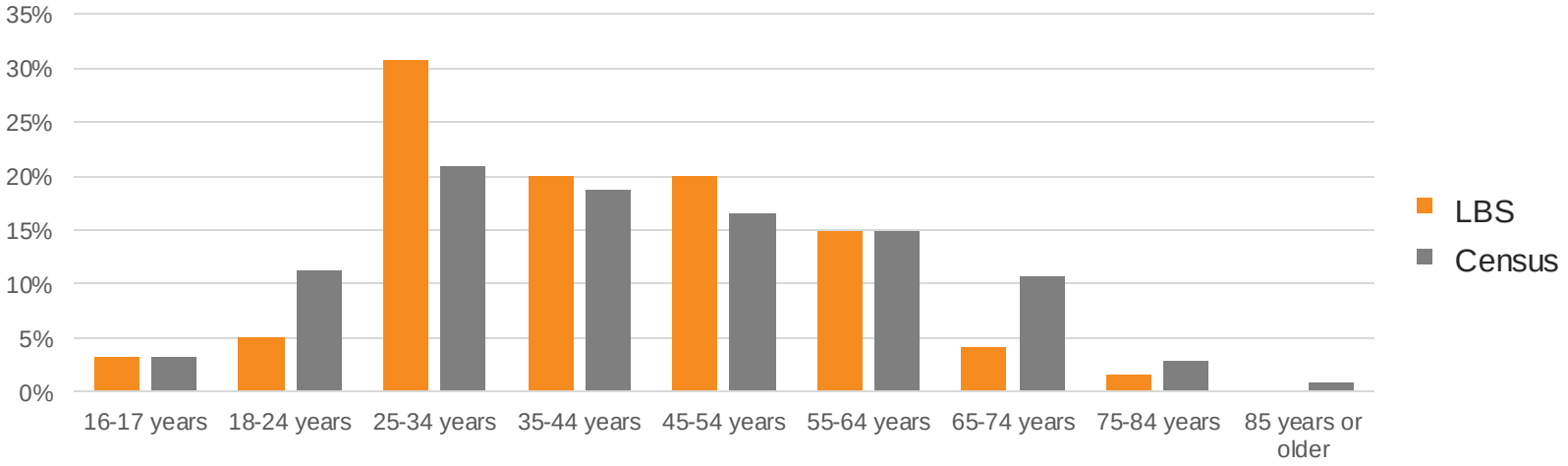


Data Verification: Demographics vs. Census

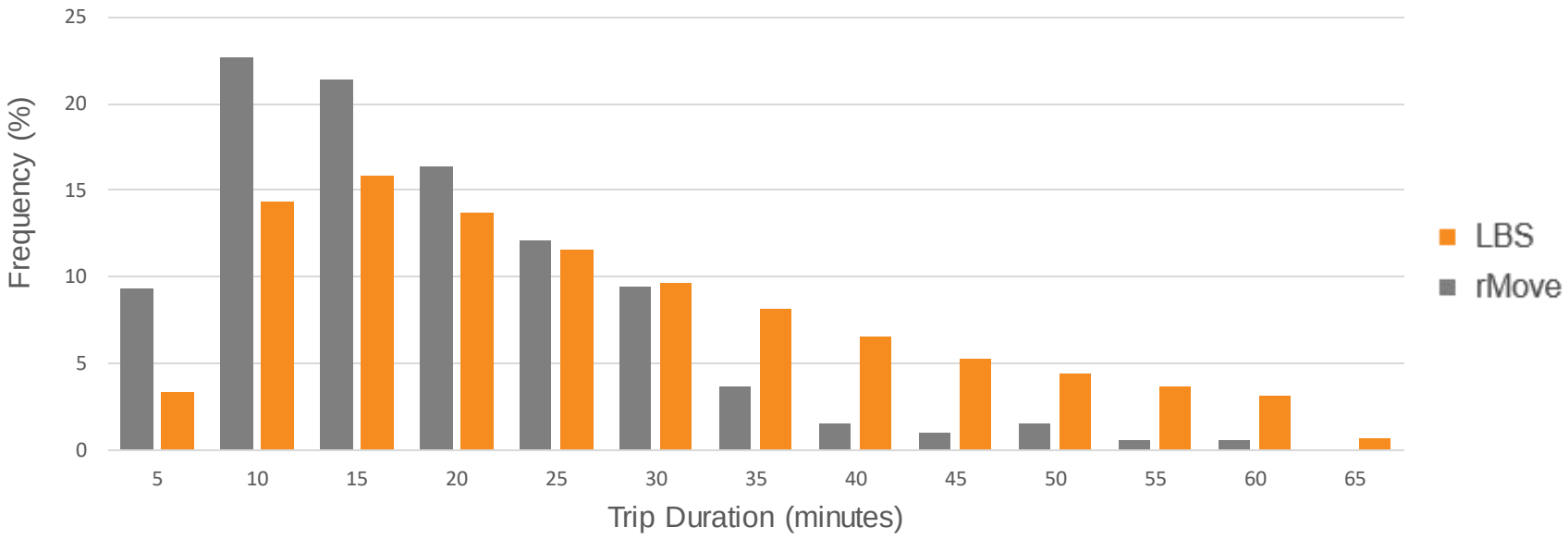
INCOME



AGE



Data Verification: Duration vs. Smartphone Survey



Expansion process matches regional counts and Census data

1 PREPARE INPUT DATA | 2 IDENTIFY TRIPS | 3 EXPAND TO REGION | 4 AGGREGATE TO O-D

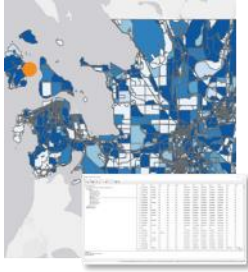
a RAKING TO CENSUS
- Rake number of residents and workers to Census estimates

b PARAMETRIC SCALING
- Create initial expansion factor using simple scaling to counts
- Apply expansion factor function (of trip/activity length)

c RAKING TO COUNTS
- Refine expansion factors with Iterative Screenline Fitting algorithm, a special form of raking or IPF

d LIMITED MATRIX ESTIMATION
- Apply Matrix Estimation (ODME) algorithm
  - Non-parametric expansion factors from comparison of loaded volumes from assignment to observed counts
  - **Minimum and maximum imposed on expansion factors**

Data aggregated to create OD tables



1 PREPARE INPUT DATA | 2 IDENTIFY TRIPS | 3 EXPAND TO REGION | 4 AGGREGATE TO O-D

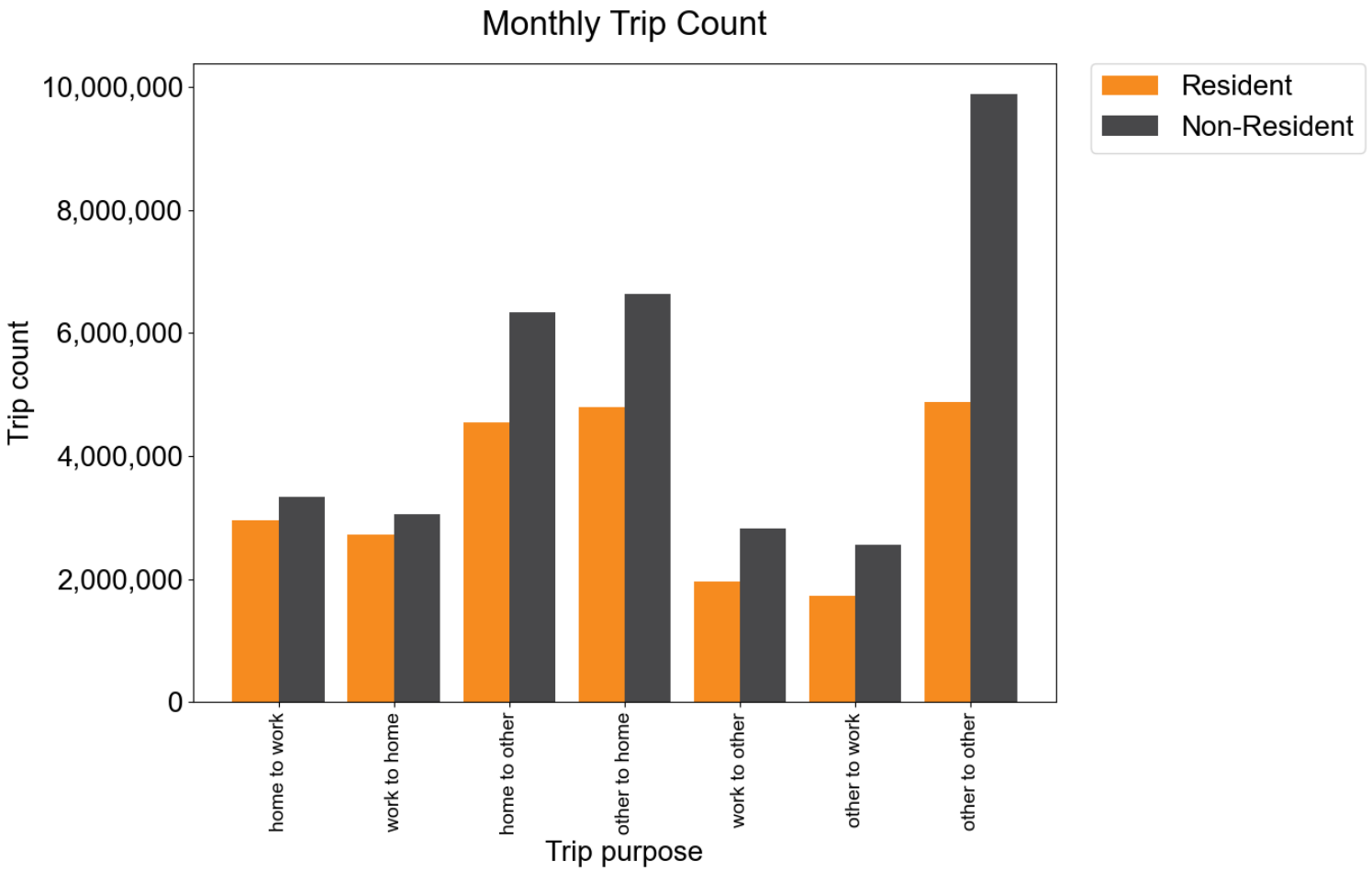
a **CLASSIFY TRIPS**

- Bin trips by resident and non-resident status
  - Calculated in trip-identification step from device "home" location
- Bin based on trip time period

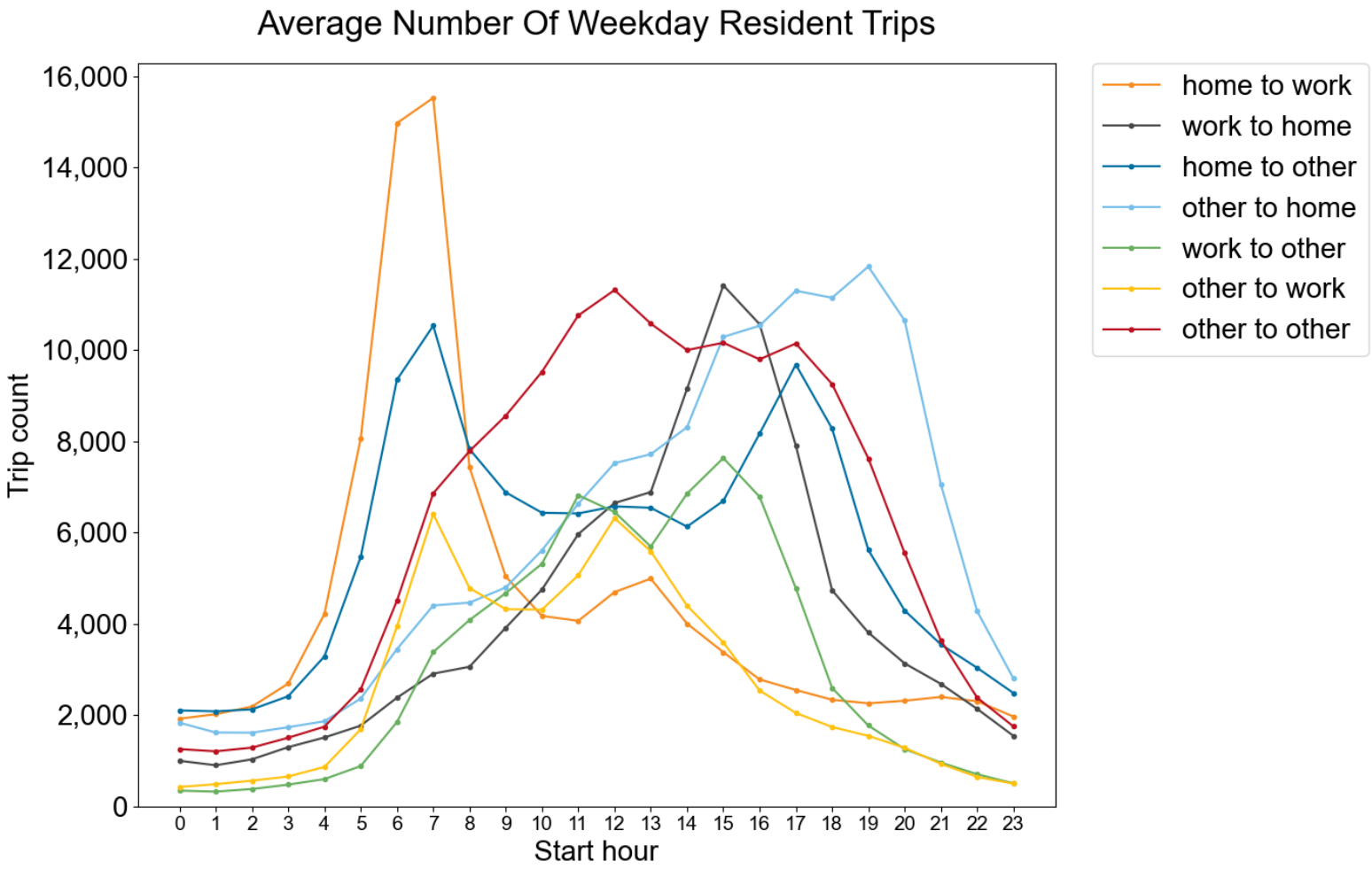
b **AGGREGATE TO MATRIX**

- Aggregate origins and destinations to model TAZ structure (or other designated geographies) to complete matrices

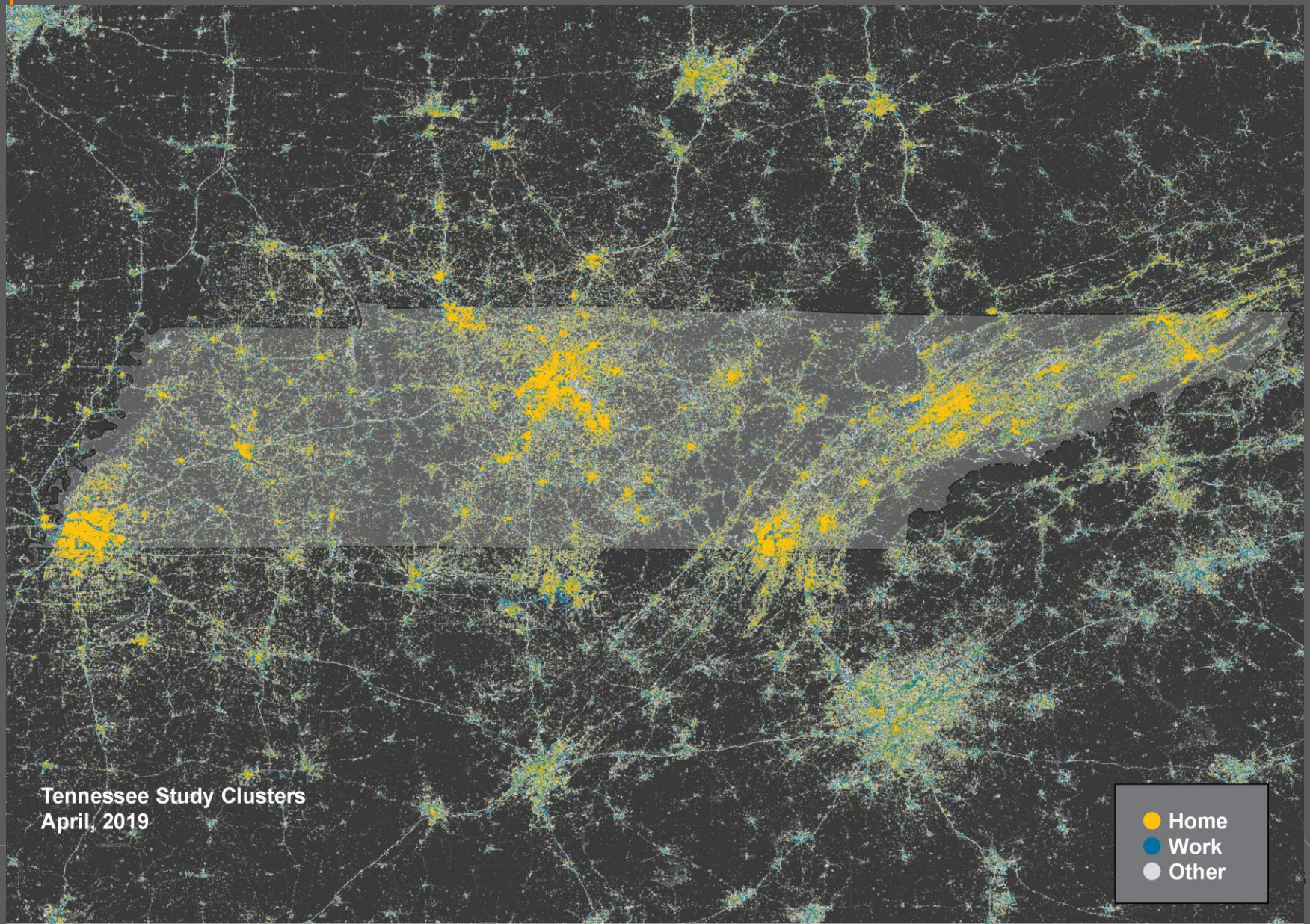
Trip Summary



Hourly Trip Distribution



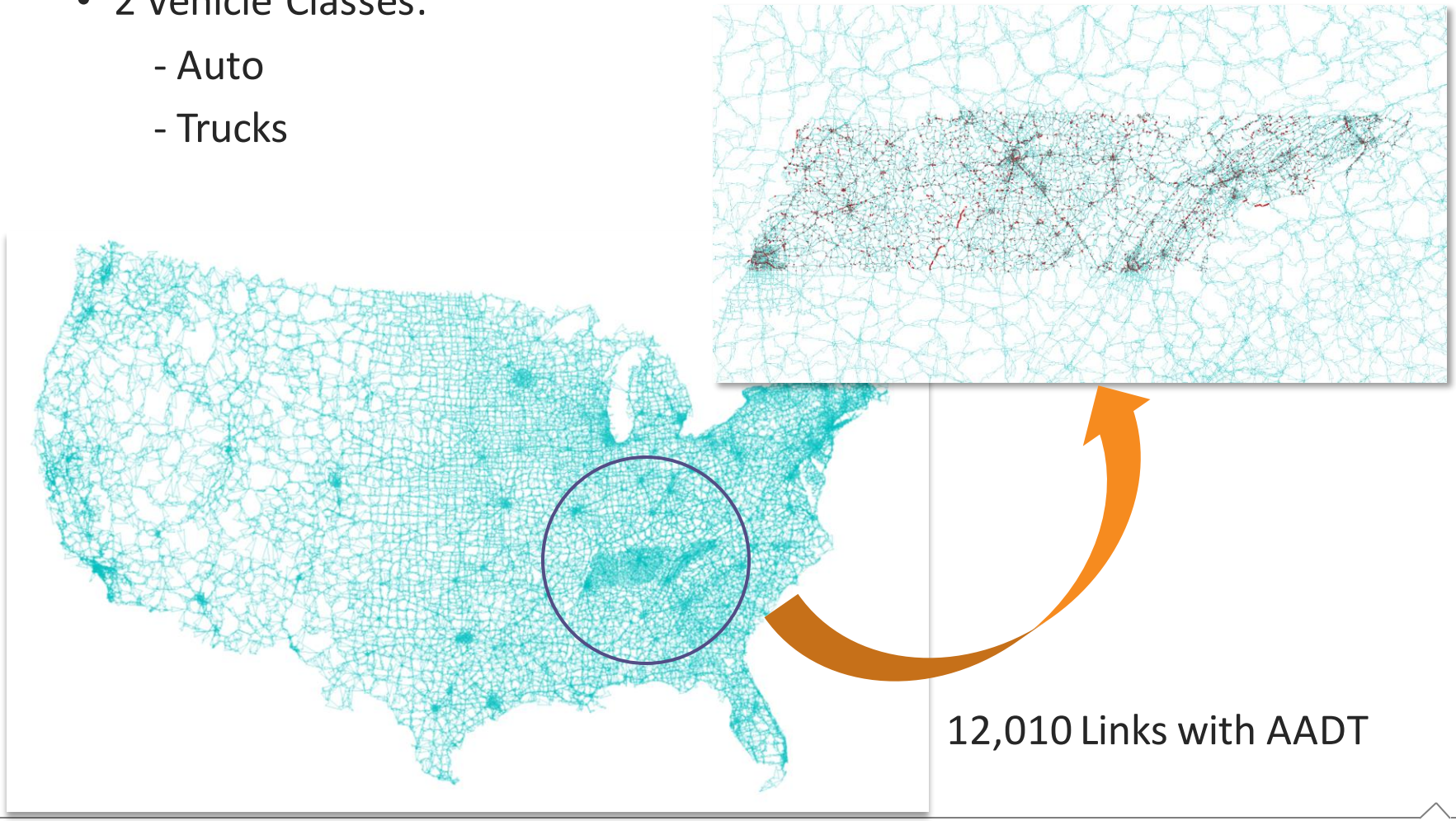
Device Observations in Tennessee



TN Big Data Application

Big Data Expansion

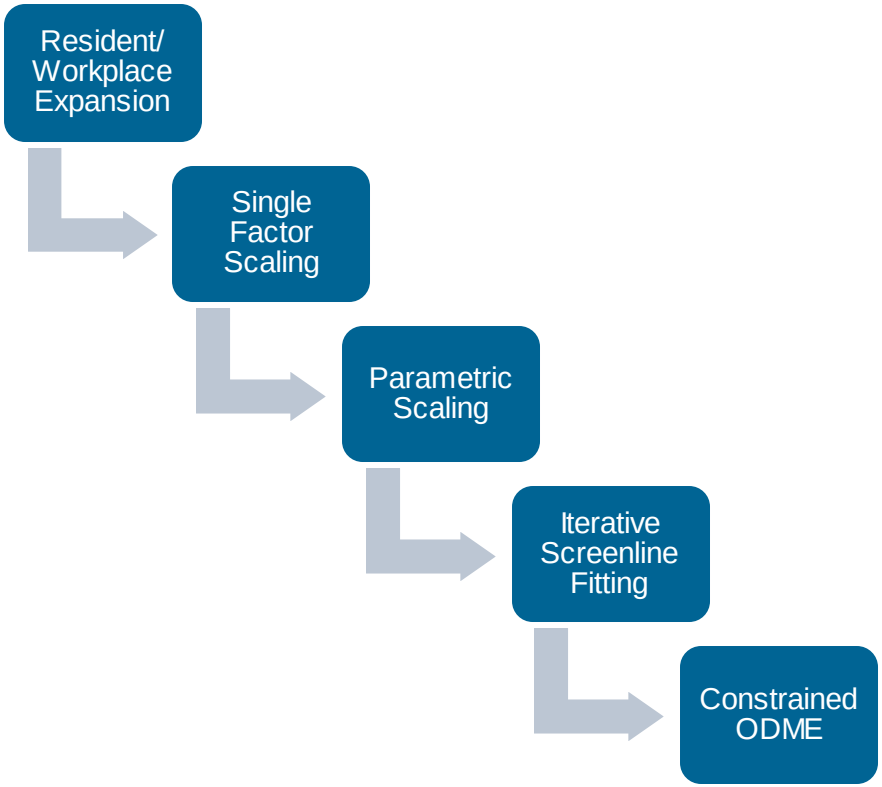
- Nationwide network with TN counts
- 2 Vehicle Classes:
  - Auto
  - Trucks



12,010 Links with AADT

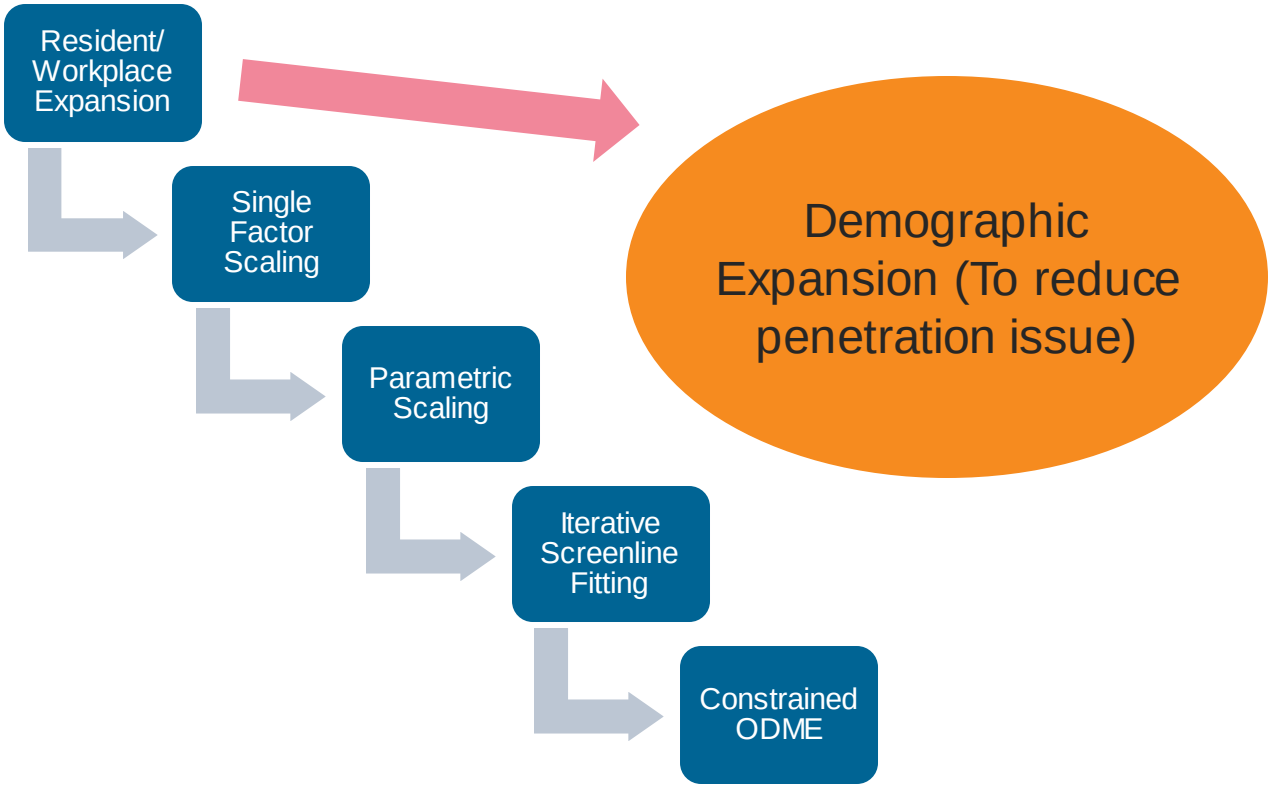
Big Data Expansion

A multistep process was used to develop the final expansion of the passive OD data



Big Data Expansion

A multistep process was used to develop the final expansion of the passive OD data



Big Data Expansion

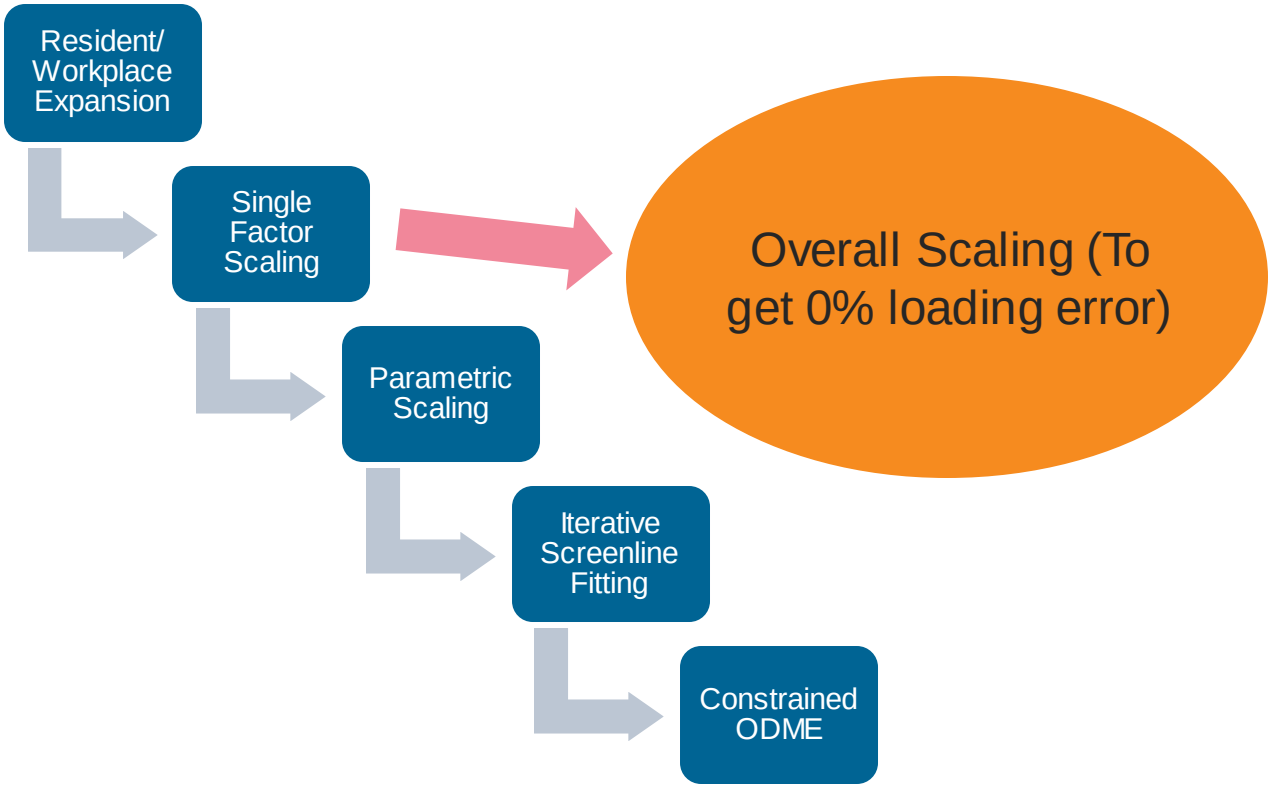
Resident/Workplace Expansion

Time Period	Trips
AM	1,961,744
MD	4,898,238
PM	2,429,418
NT	4,272,682
Total Daily	13,562,082

Statistic	All Vehicles
Total Vehicle Trips	13,562,082
Loading Error (%)	-2.7
RMSE (%)	63.3
MAPE (%)	83.5

Big Data Expansion

A multistep process was used to develop the final expansion of the passive OD data



Big Data Expansion

Single Factor Scaling

- Scaling by vehicle class
- Daily scaling factors
- Assignment by time period
- Iterative procedure
- Dampening factors after the 4th iteration

4 Iterations

Big Data Expansion

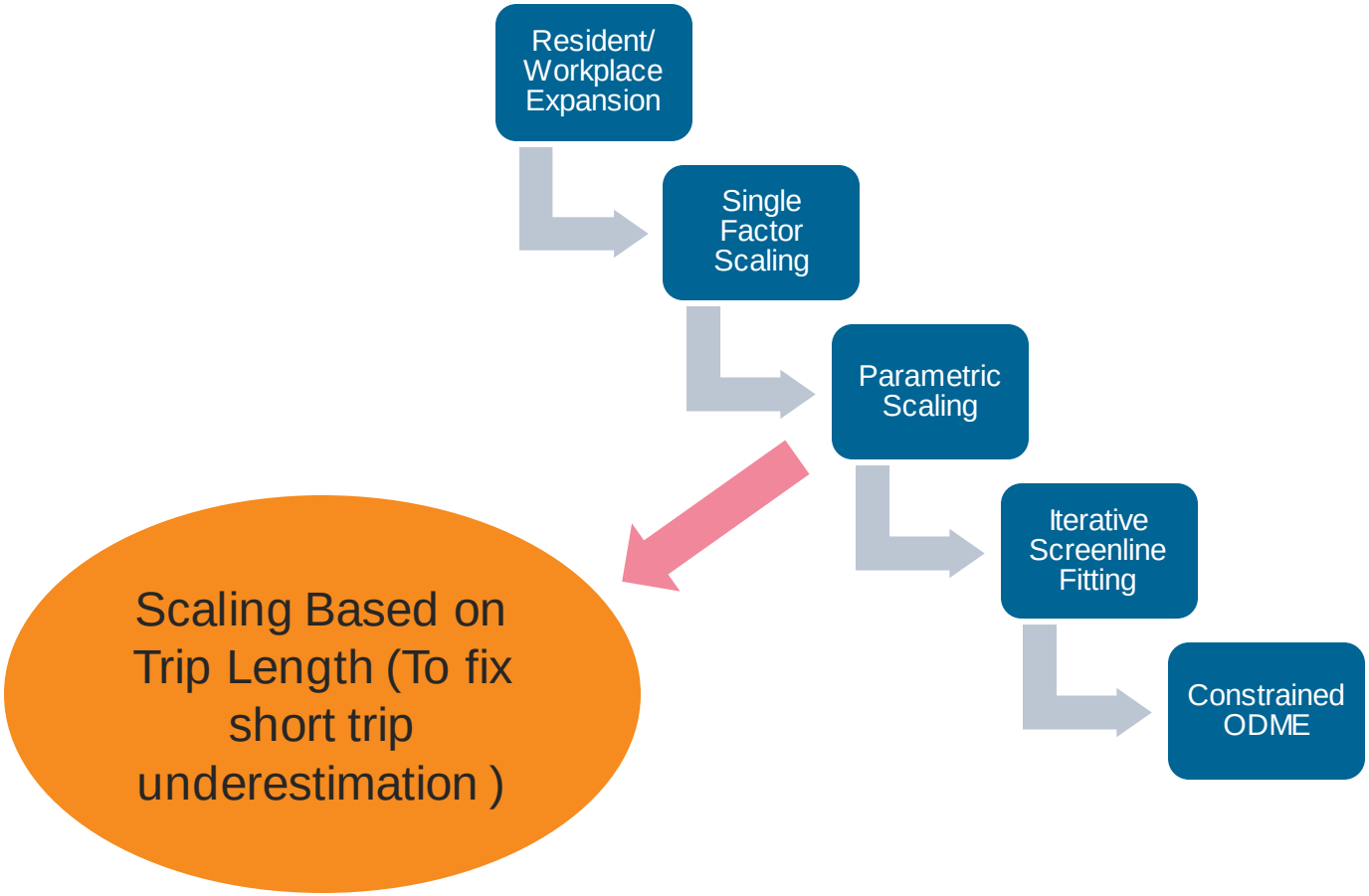
Single Factor Scaling

Time Period	Before	After	Change (%)
AM	1,961,744	2,307,444	17.6
MD	4,898,238	5,649,100	15.3
PM	2,429,418	3,007,323	23.8
NT	4,272,682	5,321,432	24.5
Total Daily	13,562,082	16,285,298	20.1

Statistic	Auto	Trucks	All Vehicles
Total Vehicle Trips	15,667,026	618,272	16,285,298
Loading Error (%)	0.4 (-34.2)	-3.7 (232.9)	1.7 (-2.7)
RMSE (%)	61.5 (83.1)	94.2 (383.5)	58.9 (63.3)
MAPE (%)	66.5 (59.4)	104.4 (583.8)	68.6 (83.5)

Big Data Expansion

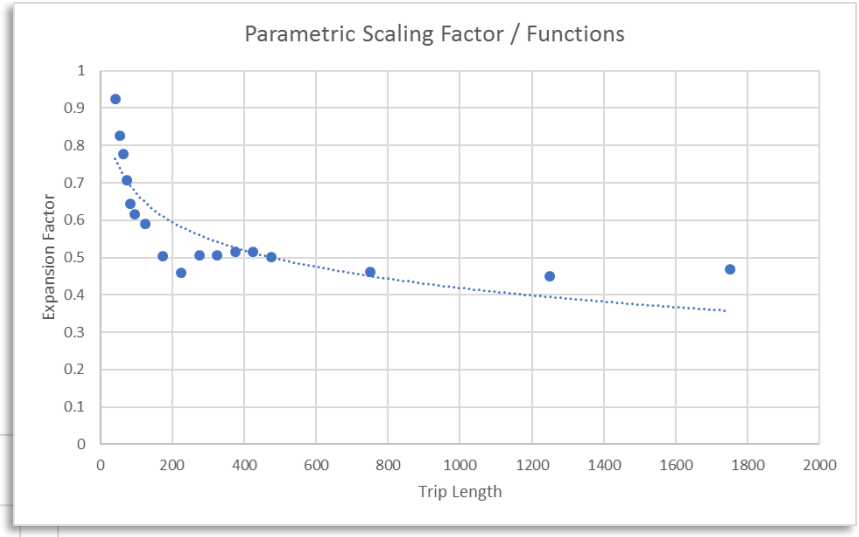
A multistep process was used to develop the final expansion of the passive OD data

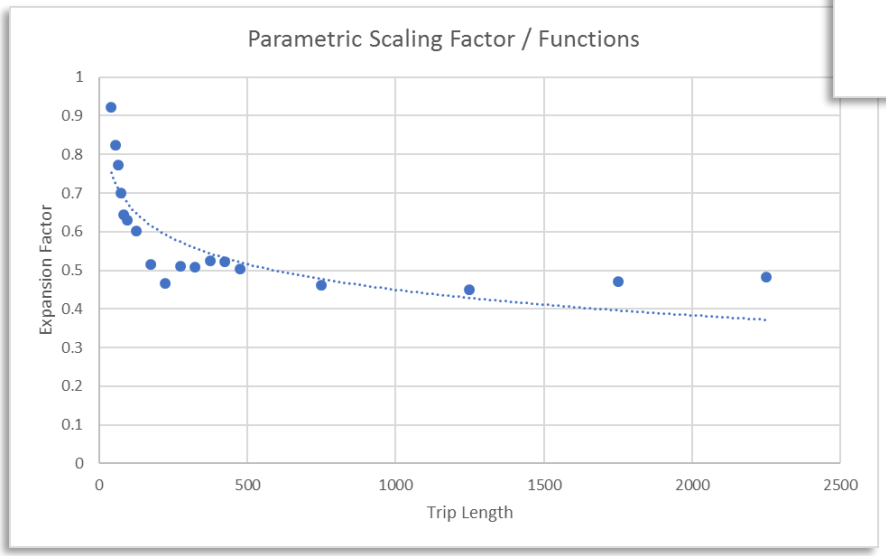


Big Data Expansion

Parametric Scaling

- Scaling by vehicle class
- Scaling factors by time of day
- Non-linear functions
- Independent variable: Trip length





Big Data Expansion

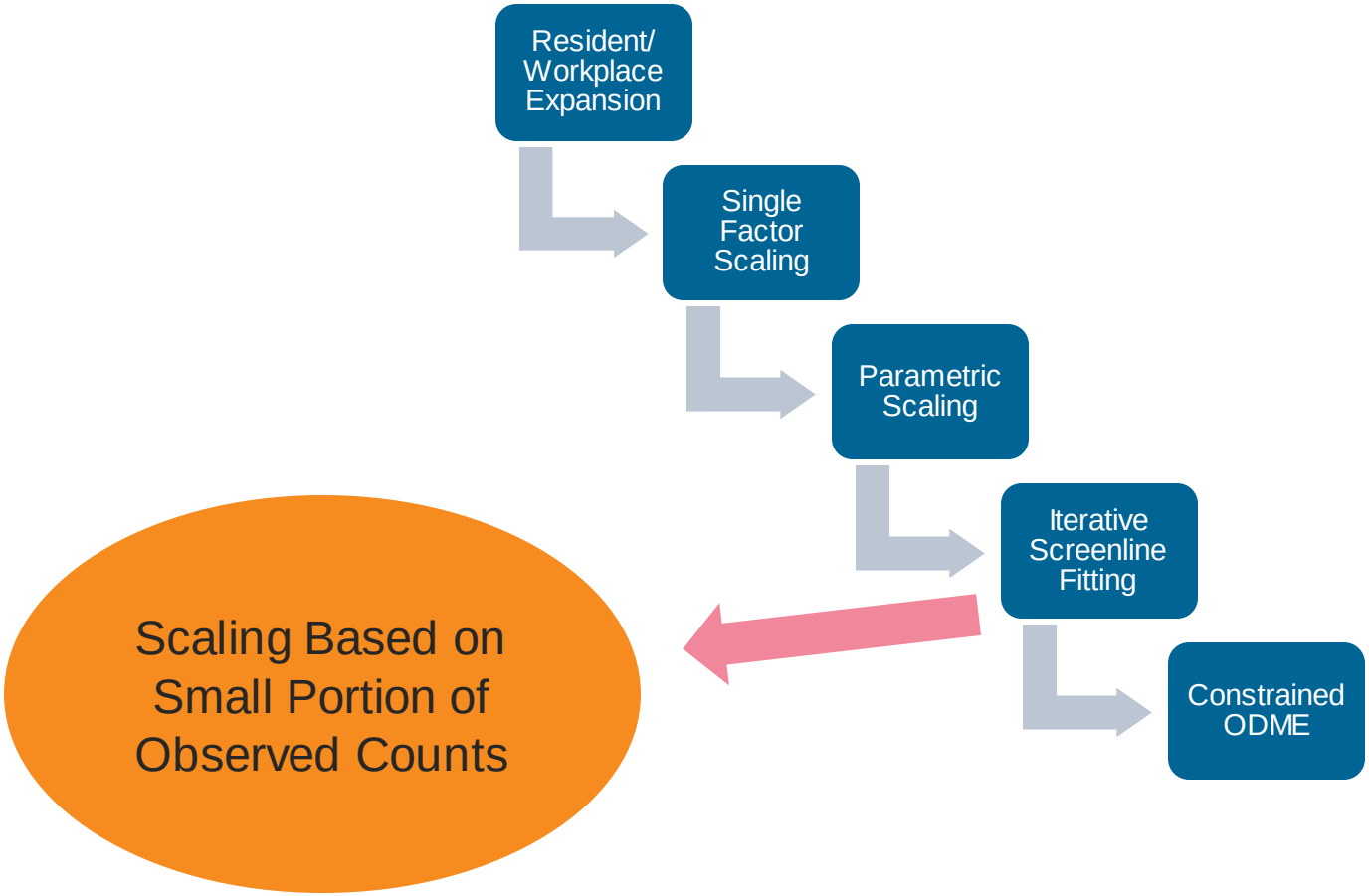
Parametric Scaling

Time Period	Before	After	Change (%)
AM	2,307,444	2,948,589	27.8
MD	5,649,100	7,246,366	28.3
PM	3,007,323	3,823,668	27.1
NT	5,321,432	6,500,158	22.2
Total Daily	16,285,298	20,518,781	26.0

Statistic	Auto	Trucks	All Vehicles
Total Vehicle Trips	19,449,080	1,069,702	20,518,781
Loading Error (%)	-2.19 (0.4)	0.1 (-3.7)	-0.1 (1.7)
RMSE (%)	53.7 (61.5)	93.6 (94.2)	51.3 (58.9)
MAPE (%)	72.5 (66.5)	93.7 (104.4)	73.4 (68.6)

Big Data Expansion

A multistep process was used to develop the final expansion of the passive OD data

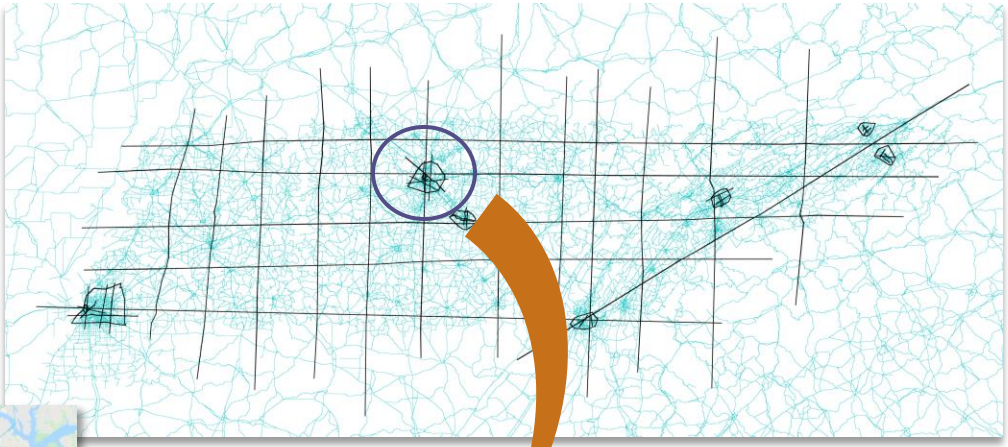


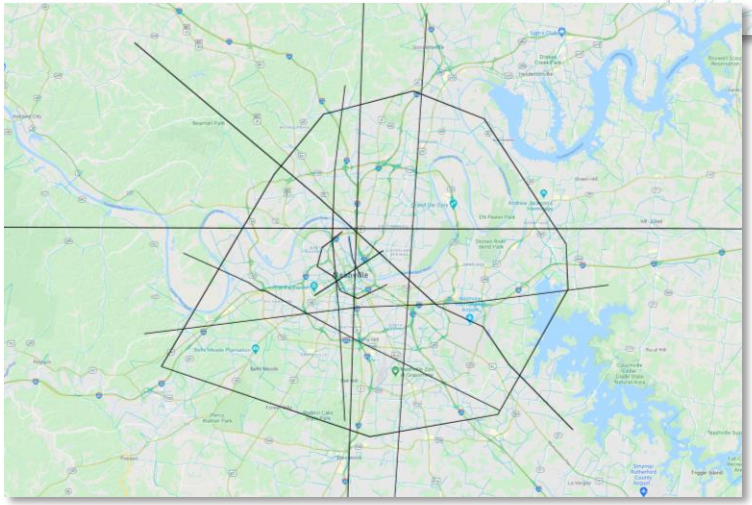
Big Data Expansion

Iterative Screenline Fitting (ISF)

- 18 Screenlines
- 7 Polygons
- 32 Cutlines

4 Iterations





Big Data Expansion

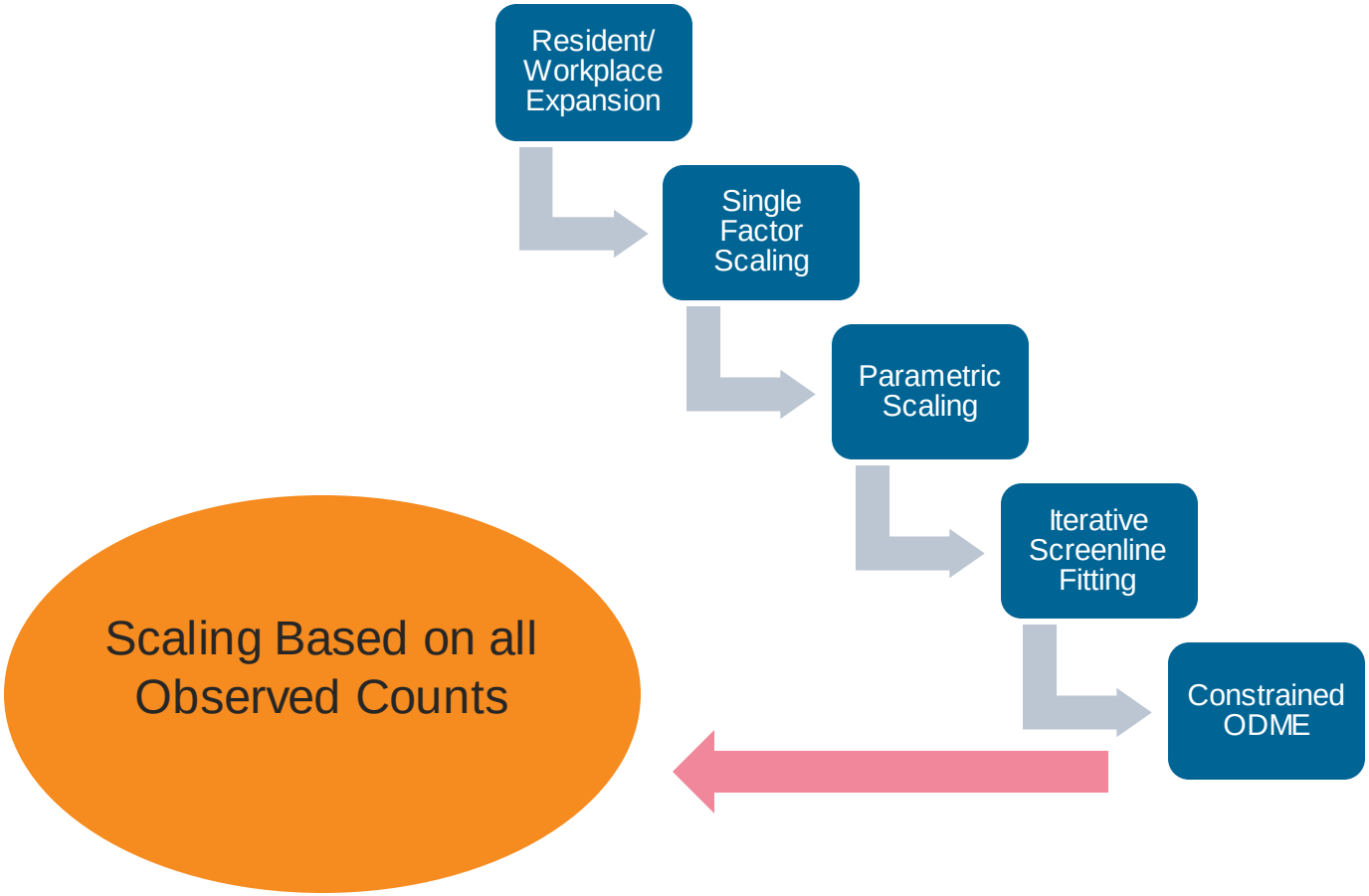
Iterative Scaling Fitting (ISF)

Time Period	Before	After	Change (%)
AM	2,948,589	2,947,800	0.0
MD	7,246,366	7,270,560	0.3
PM	3,823,668	3,805,977	-0.5
NT	6,500,158	6,466,803	-0.5
Total Daily	20,518,781	20,491,141	-0.1

Statistic	Auto	Trucks	All Vehicles
Total Vehicle Trips	19,260,543	1,230,597	20,491,141
Loading Error (%)	-1.1 (-2.19)	-4.5 (0.1)	0.5 (-0.1)
RMSE (%)	53.0 (53.7)	93.3 (93.6)	50.9 (51.3)
MAPE (%)	73.9 (72.5)	98.9 (93.7)	75.0 (73.4)

Big Data Expansion

A multistep process was used to develop the final expansion of the passive OD data



Big Data Expansion

Constrained ODME

- Scaling by vehicle class
- Scaling factors by time period
- Path building by time period
- Auto
  - Lower bound = 0.5
  - Upper bound = 2.5
- Trucks
  - Lower bound = 0.5
  - Upper bound = 3.5
  - Solid 0.1 trips as lower bound for any cell with very little number of trips
- 6 Iteration

Big Data Expansion

Constrained ODME

Time Period	Before	After	Change (%)
AM	2,947,800	3,040,394	3.1
MD	7,270,560	7,431,681	2.2
PM	3,805,977	3,914,725	2.9
NT	6,466,803	6,652,236	2.9
Total Daily	20,491,141	21,039,037	2.7

Statistic	Auto	Trucks	All Vehicles
Total Vehicle Trips	19,825,873	1,213,163	21,039,037
Loading Error (%)	1.0 (-1.1)	-4.7 (-4.5)	2.3 (0.5)
RMSE (%)	40.5 (53.0)	66.5 (93.3)	39.8 (50.9)
MAPE (%)	57.0 (73.9)	72.3 (98.9)	58.6 (75.0)

Data Expansion

Time Period	Single Factor Scaling	ODME	Change (%)
AM	2,307,444	3,040,394	31.8
MD	5,649,100	7,431,681	31.6
PM	3,007,323	3,914,725	30.2
NT	5,321,432	6,652,236	25.0
Total Daily	16,285,298	21,039,037	29.2

Statistic	Single Factor Scaling	ODME
Loading Error (%)	0.4	2.3
RMSE (%)	61.5	39.8
MAPE (%)	66.5	58.6

Validation Statistics

Facility Type	Loading Error (%)	RMSE (%)
Rural Interstates	1.67	13.11
Rural Principal Arterials	-0.69	28.99
Rural Minor Arterials	2.12	37.39
Rural Major Collectors	14.03	62.63
Rural Minor Collectors	38.89	127.5
Rural Local roads	3.38	86.49
Urban Interstates	2.58	14.65
Urban Other Freeways	1.79	21.30
Urban Principal Arterials	-7.93	33.15
Urban Minor Arterials	-0.23	43.23
Urban Collectors	16.35	88.71
Urban Local Roads	6.33	71.97

Validation Statistics

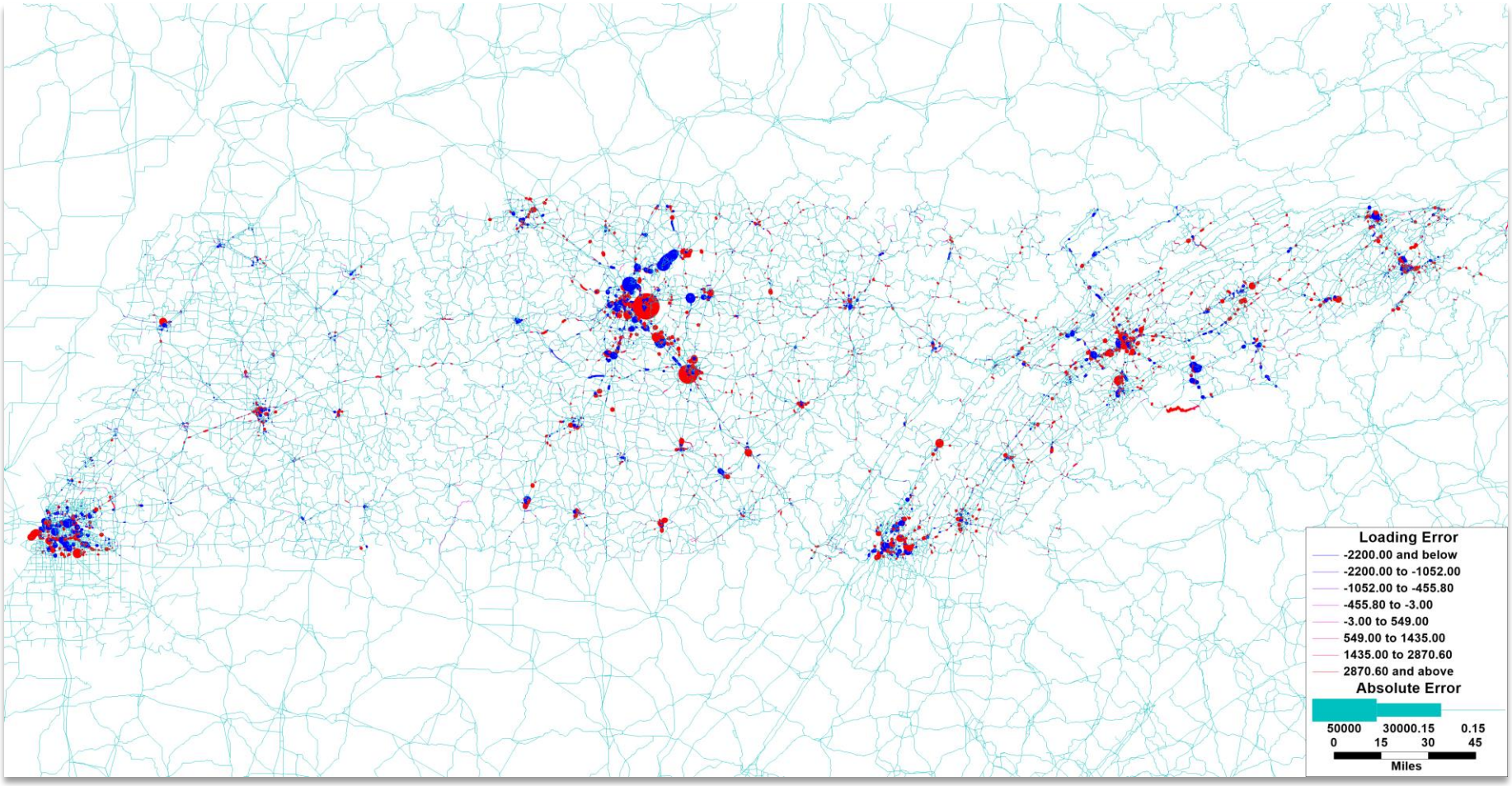
Facility Type	Loading Error (%)	RMSE (%)
Freeways	2.41	15.44
Arterials	-3.03	37.19
Collectors	20.10	94.46
Locals	5.80	74.66
All	2.30	39.79

Area Type	Loading Error (%)	RMSE (%)
Urban	1.04	36.35
Rural	7.09	43.06

Validation Statistics – RMSE (%)

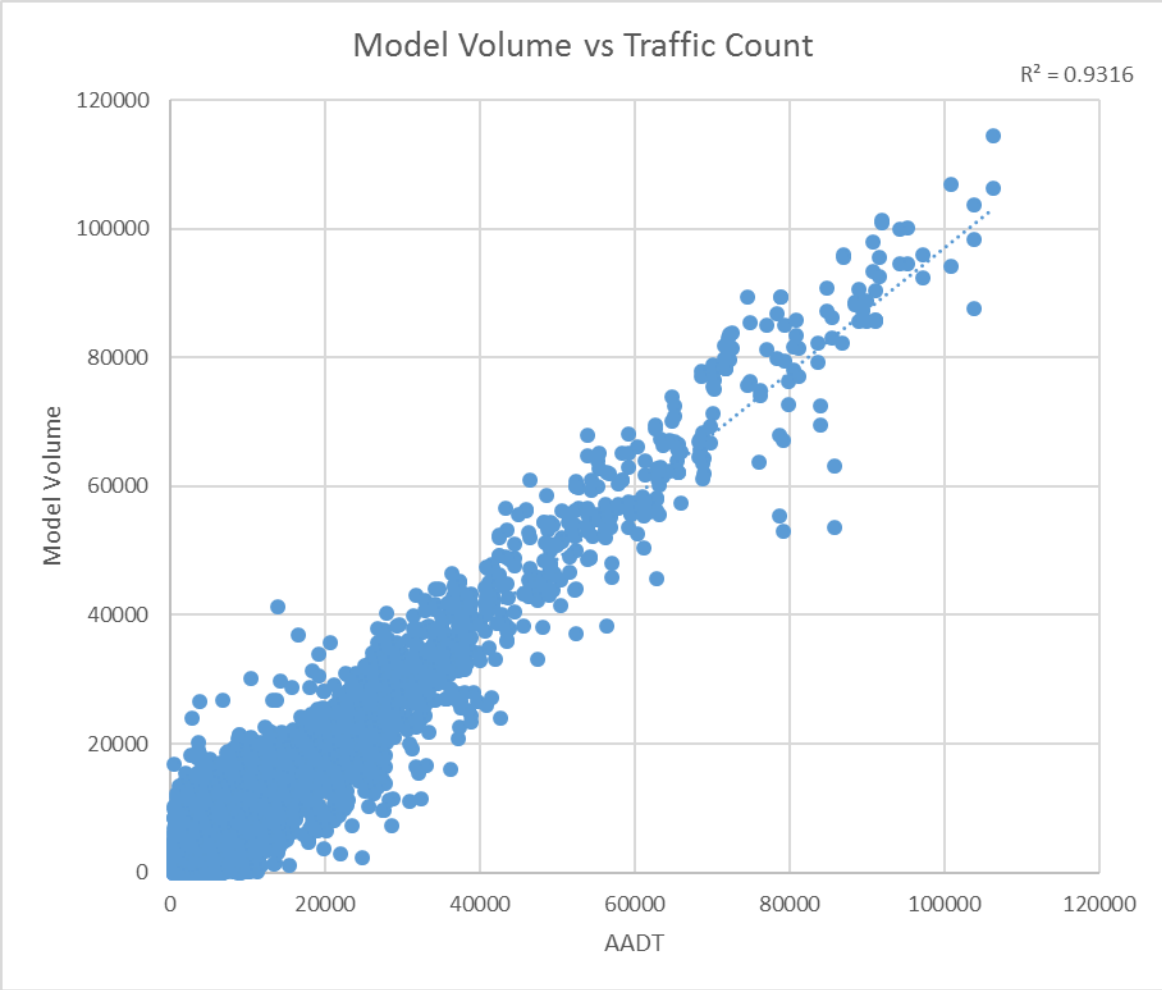
AADT	Expansion	Guideline
< 5,000	113.2	101.4
5k-10k	40.7	56.3
10k - 20k	26.2	51.4
20k - 30k	21.7	35.7
30k - 40k	18.7	32.0
40k - 50k	15.6	19.8
50k - 60k	10.5	20.5
> 60k	9.9	14.4
Total	**39.8**	**60.0**

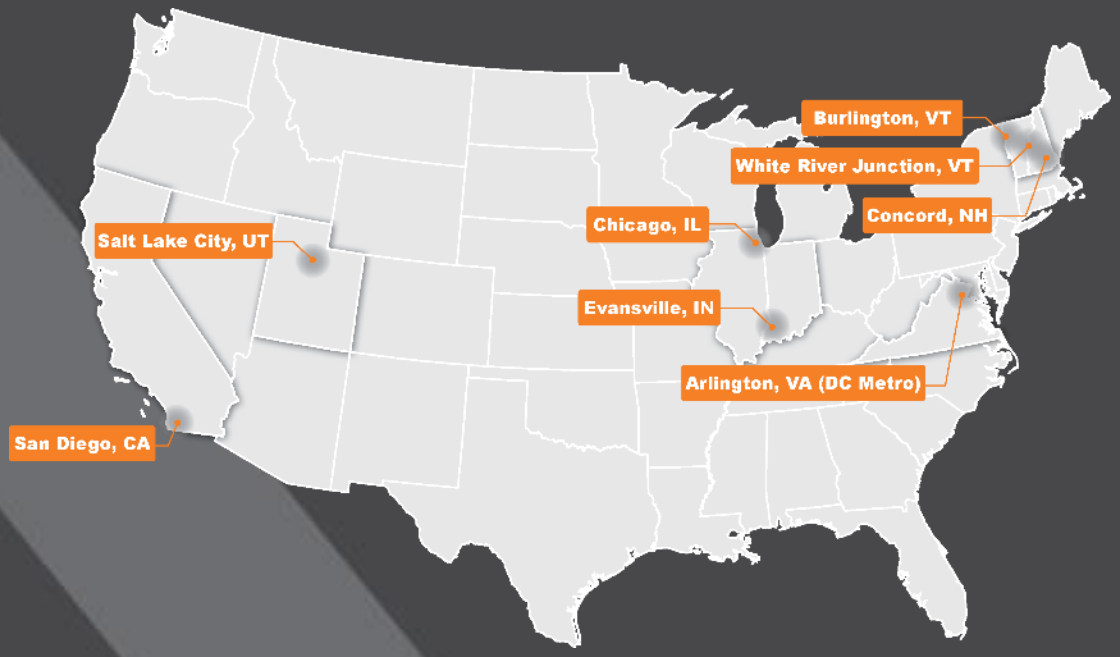
Model Loaded Network



Validation Statistics

Correlation = 0.93





Contact

www.rsginc.com

Steven Trevino
CONSULTANT

Steven.Trevino@rsginc.com
812.202.5760

Hadi Sadrsadat, PhD
CONSULTANT

Hadi.Sadrsadat@rsginc.com
240.283.0636