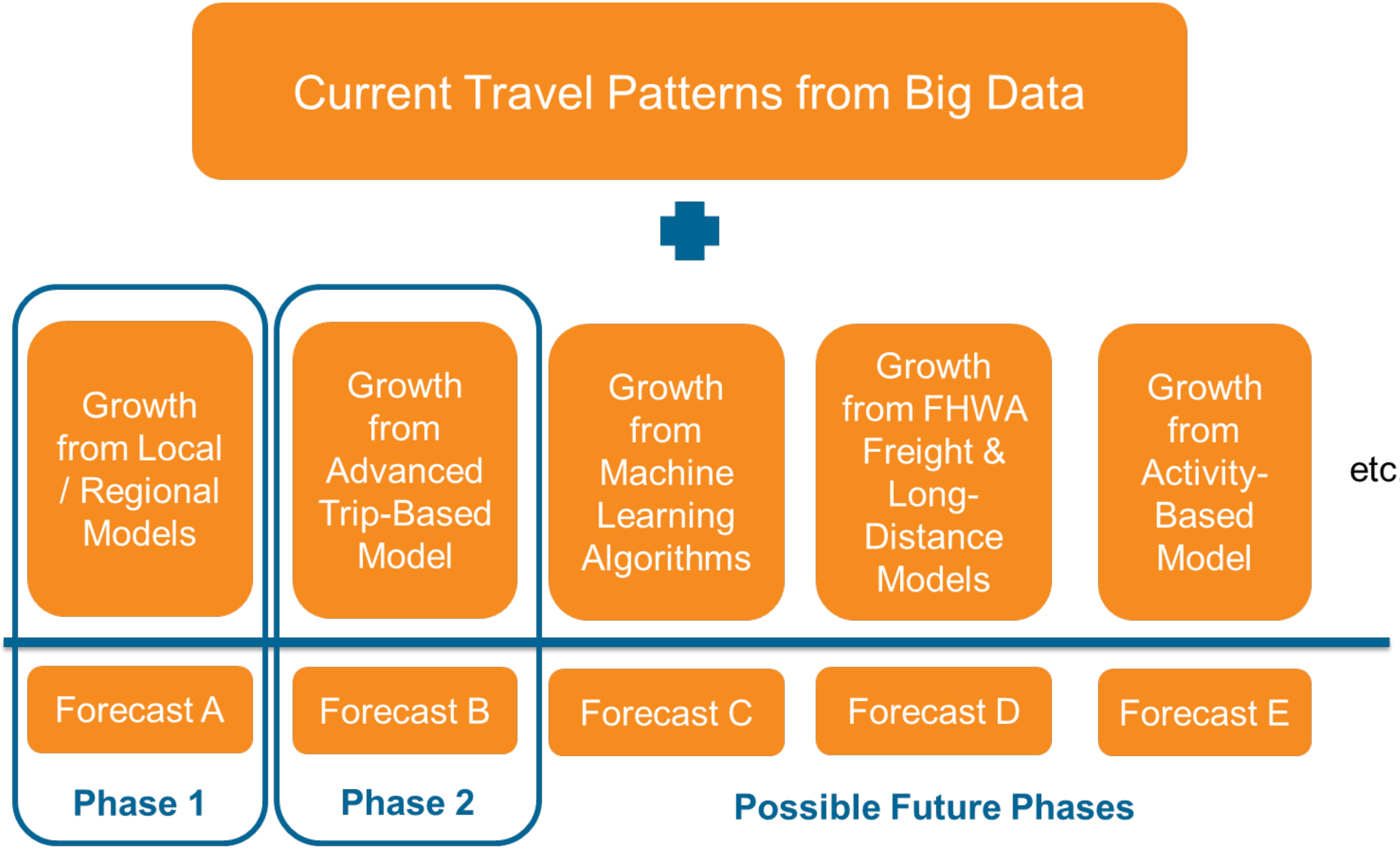
Big Data & Pivoting in the NCSTM

Steven Trevino & Stephen Tuttle

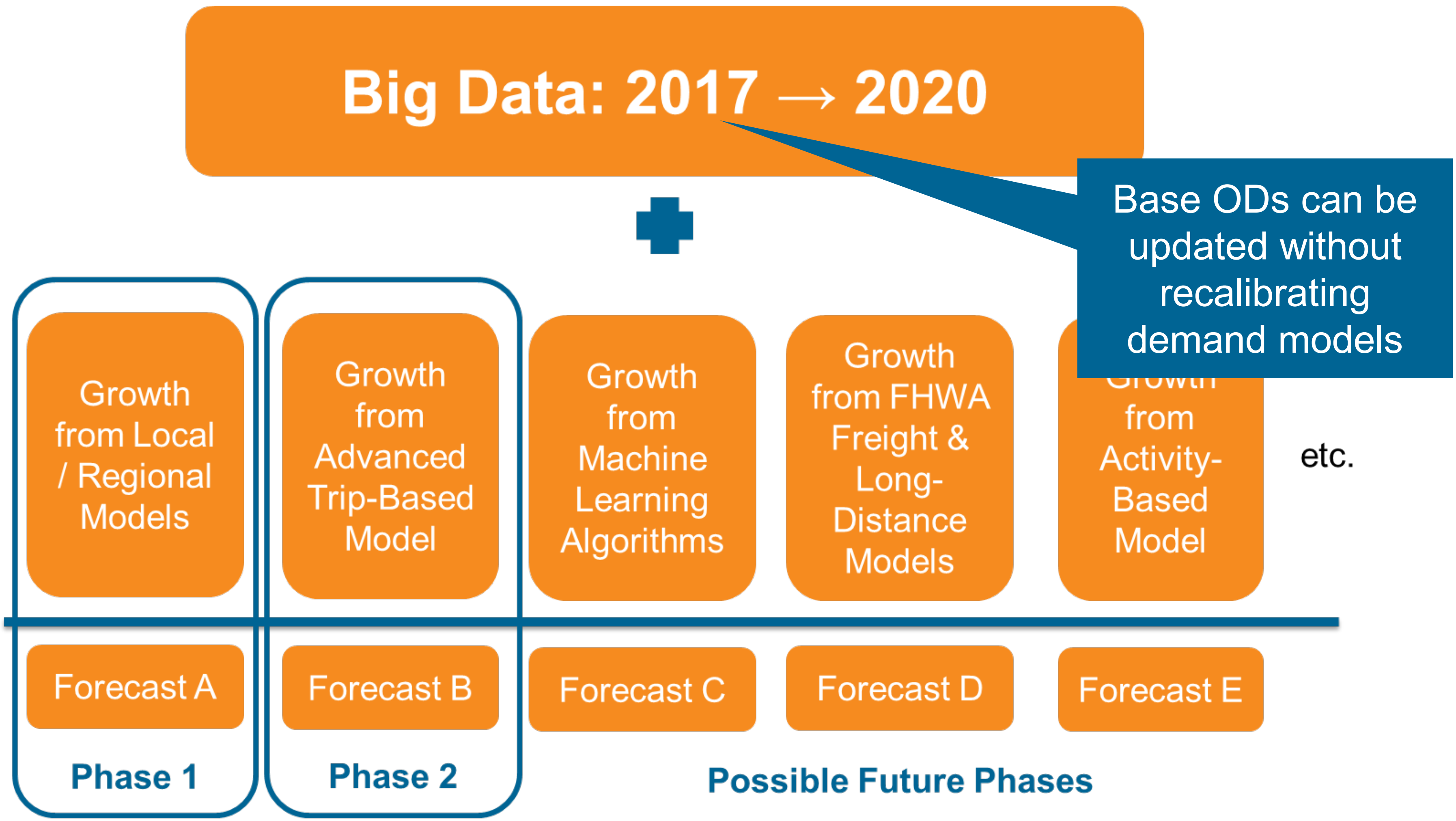
December 2, 2020

Background

Extensible Modular Framework



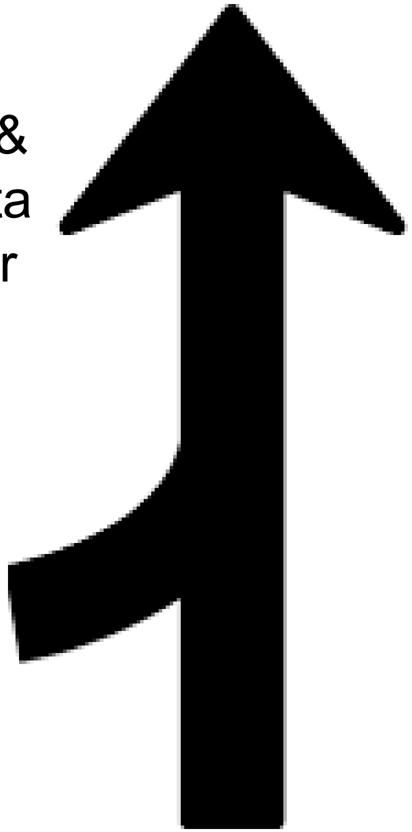
Updating Base Year without Recalibration



rMerge & LBS Data

RSG’s rMerge Platform

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



How is rMerge Applied?

RSG the science of insight

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

Census

Traffic Counts

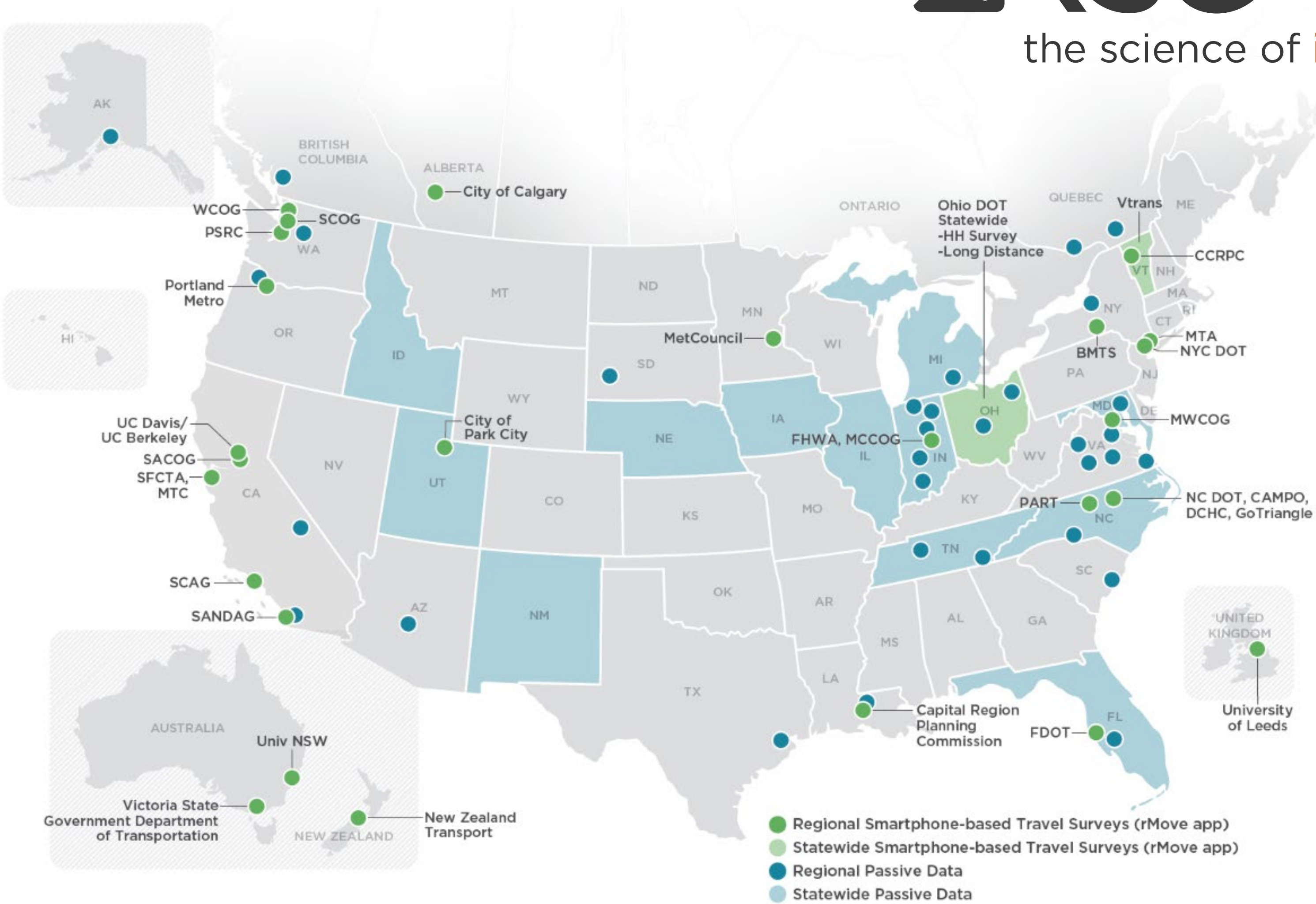
Travel Surveys

LBS data

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

Mobile Data Experience

RSG
the science of insight



Over 50 projects in over 25 states

How Big is this Big Data?

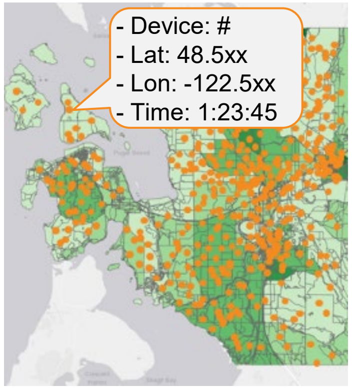
- 10-15% population on any given day (DAU)
- 50% of population over a month (MAU)
- ~ 3.8 million unique devices for NC during October 2018
- Larger sample than surveys or pure navigational GPS

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

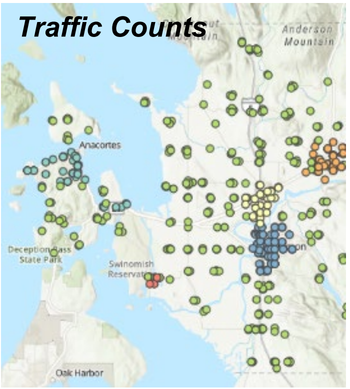


Billions of individual device location points from commercial LBS data are extracted, evaluated for basic metrics & cleaned*

2
IDENTIFY TRIPS

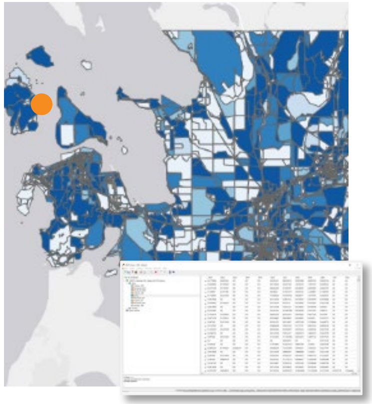
Points are clustered to identify stop locations, locations are classified (home, work, other) and linked to create trips

3
EXPAND TO REGION



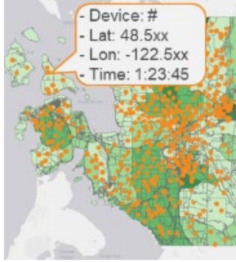
Trips are expanded to region based on Census and traffic count data, surveys and other sources to provide representative O-D flows

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

North Carolina LBS Data Summary

LBS Data - October 2018	
Sightings	1,268,125,349
Total Devices	3,873,300
Good Devices	1,290,589
Locations	9,676,084
Trips	32,432,463

– LBS data represents a sample of 8.3% of NC 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?

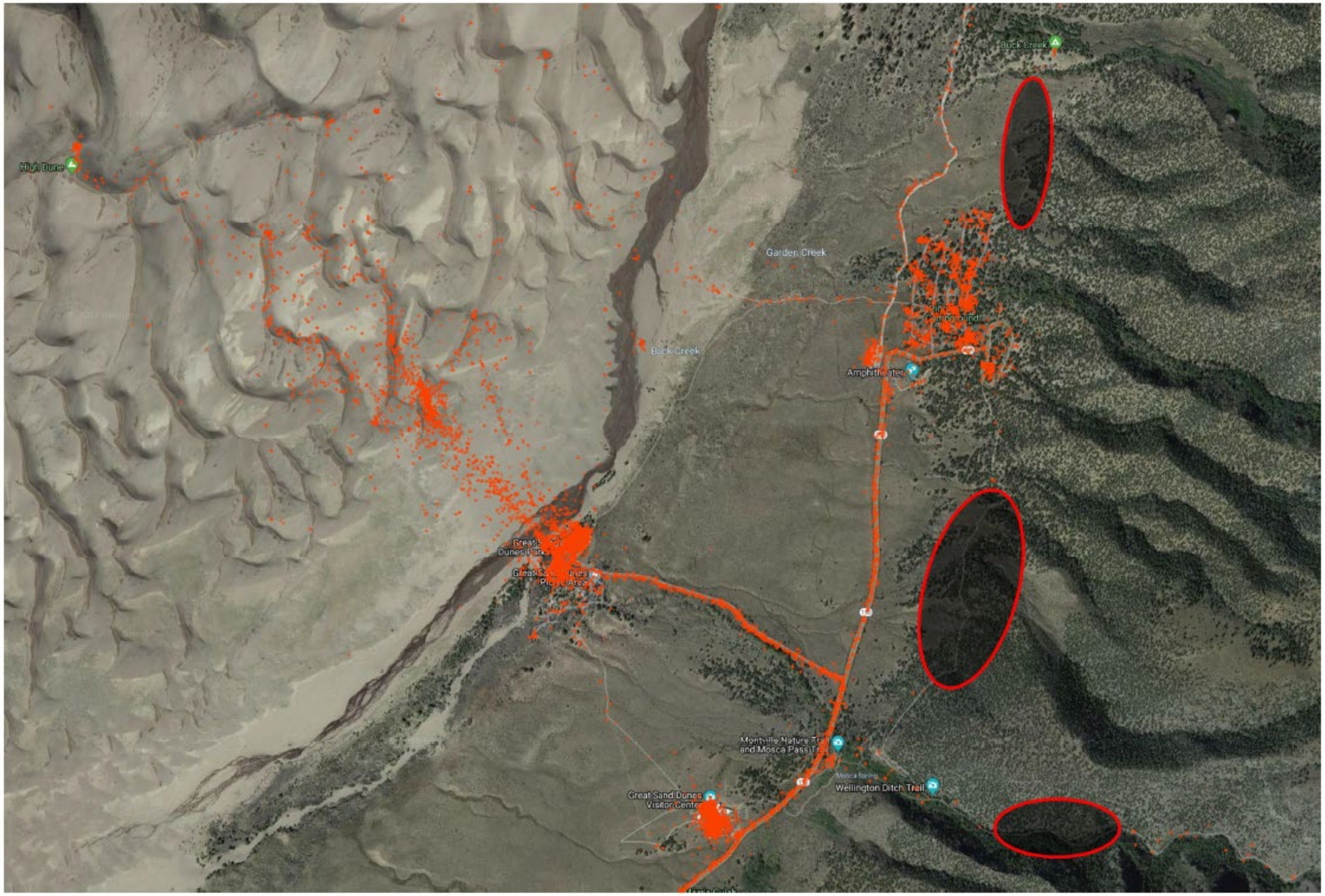
- Big data are large scale observations.
- But they are still only a sample of all travel.
- And they are NOT a random sample.
- Big data are known to have systematic **biases**.
- But if we can **measure** bias, we can **correct** for it.

What's Missing in Big Data?

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

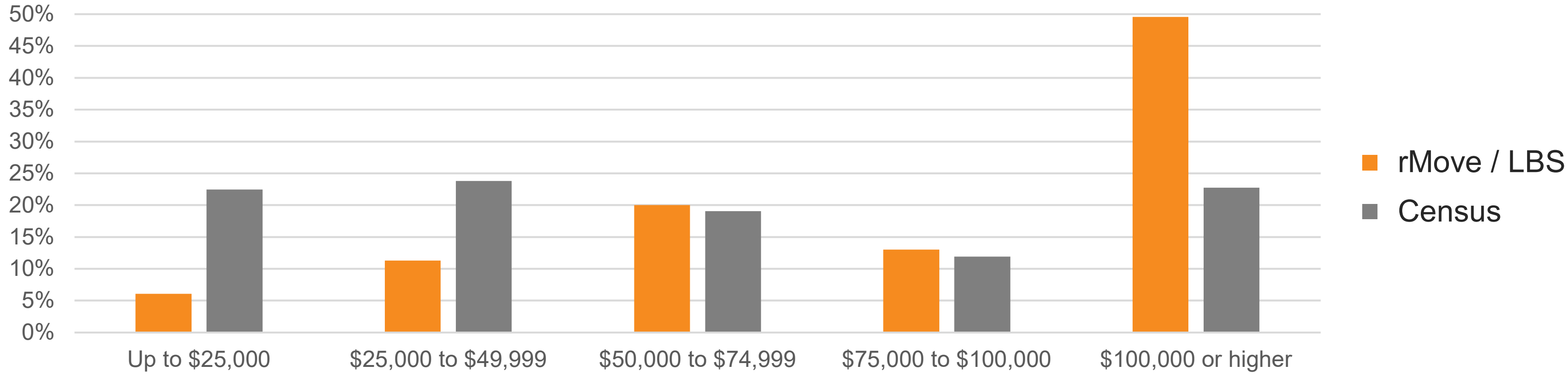
Geographic Coverage Gaps & Variations

SIGHTINGS AT GREAT SAND DUNES NATIONAL PARK IN JULY 2018

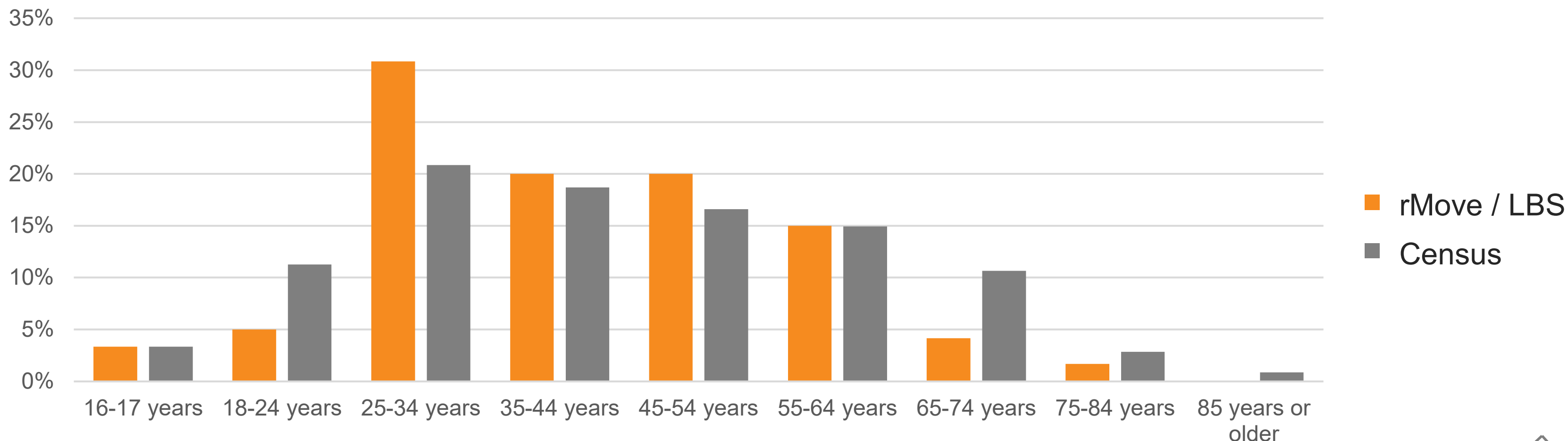


Data Verification: Demographics vs. Census

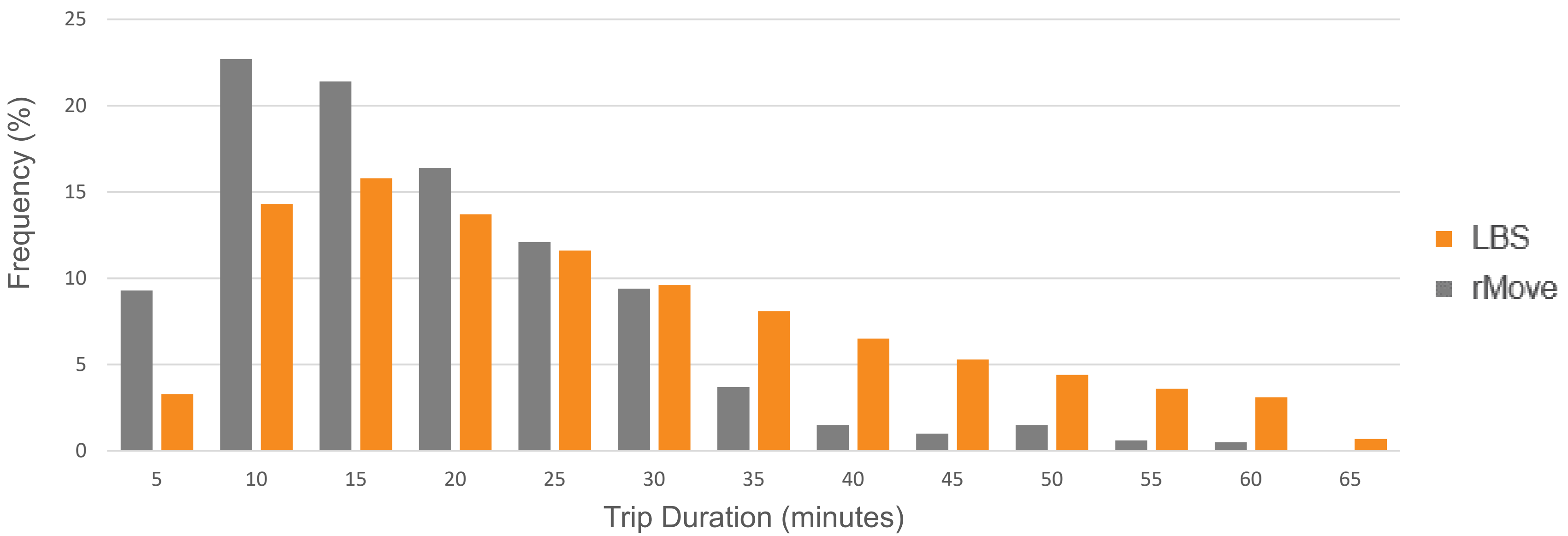
INCOME



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

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

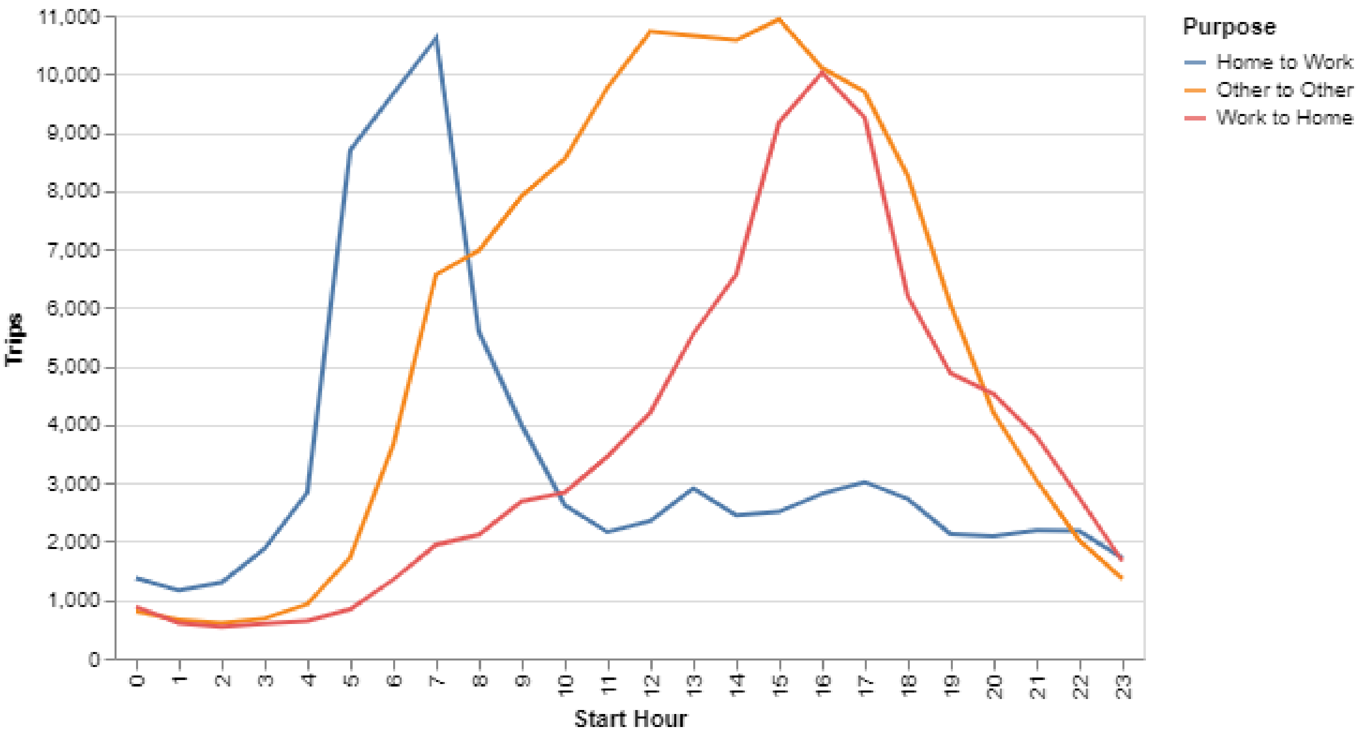
a — CLASSIFY TRIPS

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

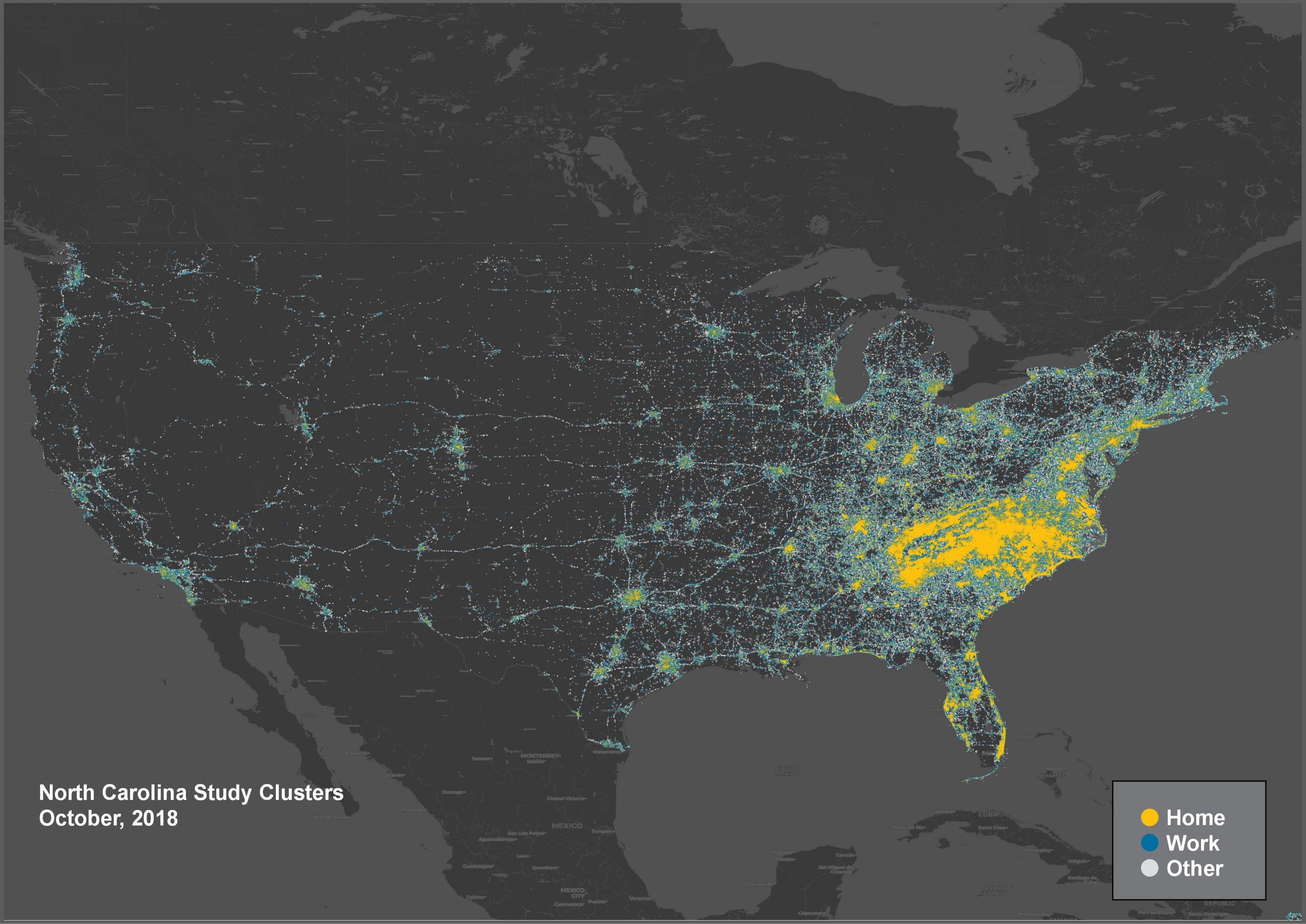
b — AGGREGATE TO MATRIX

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

Hourly Trip Distribution



Device Observations in NC



Big Data Pivoting

Why Pivot?

- Pivoting improves the accuracy of travel models by allowing the model to forecast changes from a known base

- Destination choice models still struggle to reproduce observed OD patterns

- Builds from recommended data-driven forecasting approaches (NCHRP 255 & 765)

- Pivoting requires accurate base year information

Types of Pivoting

- FHWA TMIP webinars by RAND Europe for Australia Forecasting (2015)
- *Pivoting in Travel Demand Models* (Daly, et al., 2012)

Goal: Combine model "synthetic" forecasts for base (S_b) and future (S_f) with base information (B) on flows

1. Multiplicative: $P = (S_f/S_b)\ B$
2. Additive: $P = B + (S_f - S_b)$
3. Mixed / Average of above

8-Case Pivoting (Mixed)

Case	Base (B)	Synthetic Base (S_b)	Synthetic Future (S_f)	Predicted
1	0	0	0	0
2	0	0	>0	Sf
3	0	>0	0	0
4n	0	>0	>0 (< X)	0
4e	0	>0	> 0 (> X)	Sf - X
5	>0	0	0	B
6	>0	0	>0	B + Sf
7	>0	>0	0	0
8n	>0	>0	>0 (<X)	B * (Sf/Sb)

Base matrix (B) : data derived base year OD demand
Synthetic Base (S_b) : base year demand model output
Synthetic Future (S_f) : future year demand model output
Switching Point (X) : parameter used to identify high growth

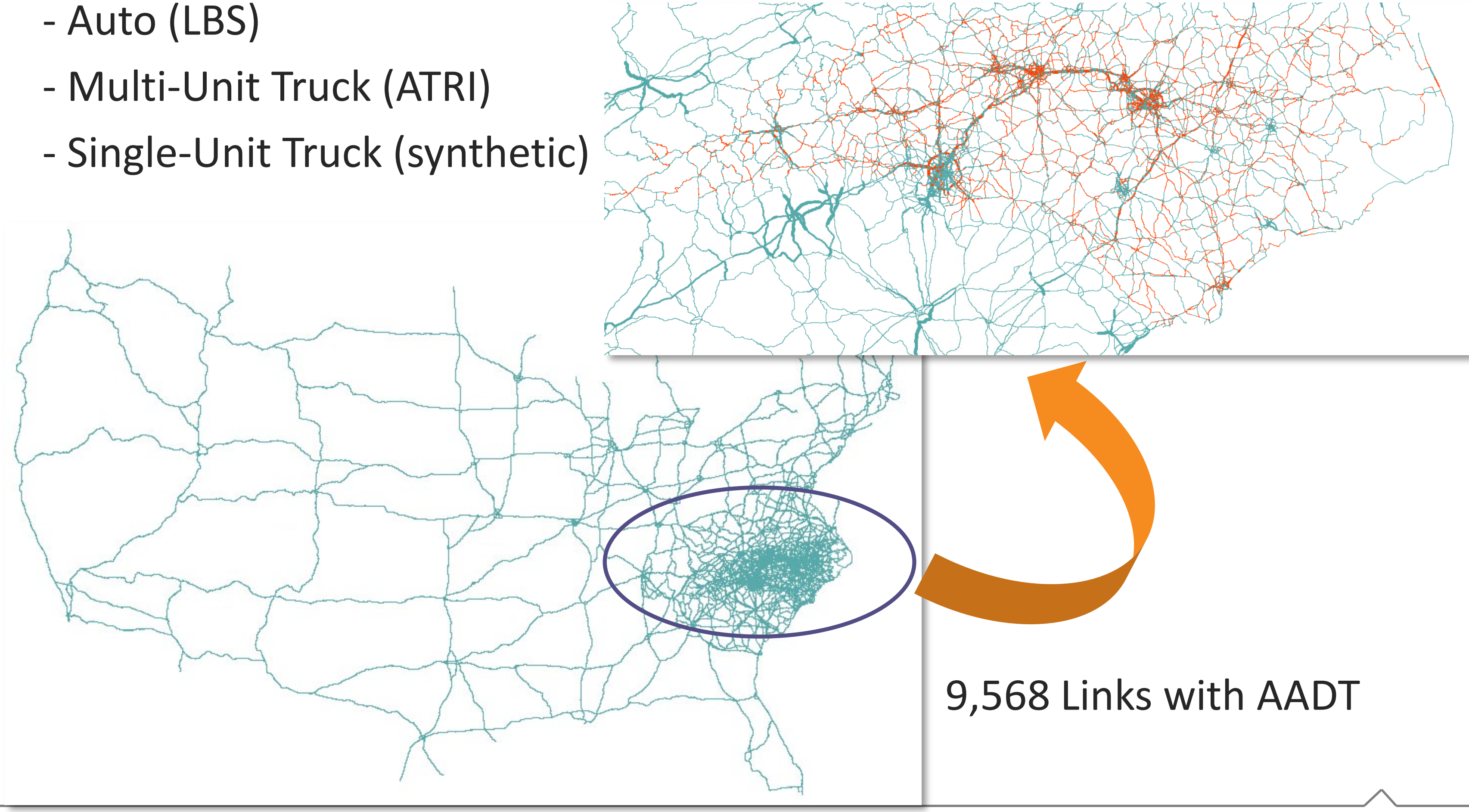
Pivoting Pitfalls

- Applying multiplicative factors can be challenging
  - Defining & calibrating switching point (X)
  - Base year model & passive data alignment
  - Base year errors can be amplified by future SE
- Model growth is often interpolated and applied uniformly

NCSTM Big Data Application

Big Data Expansion

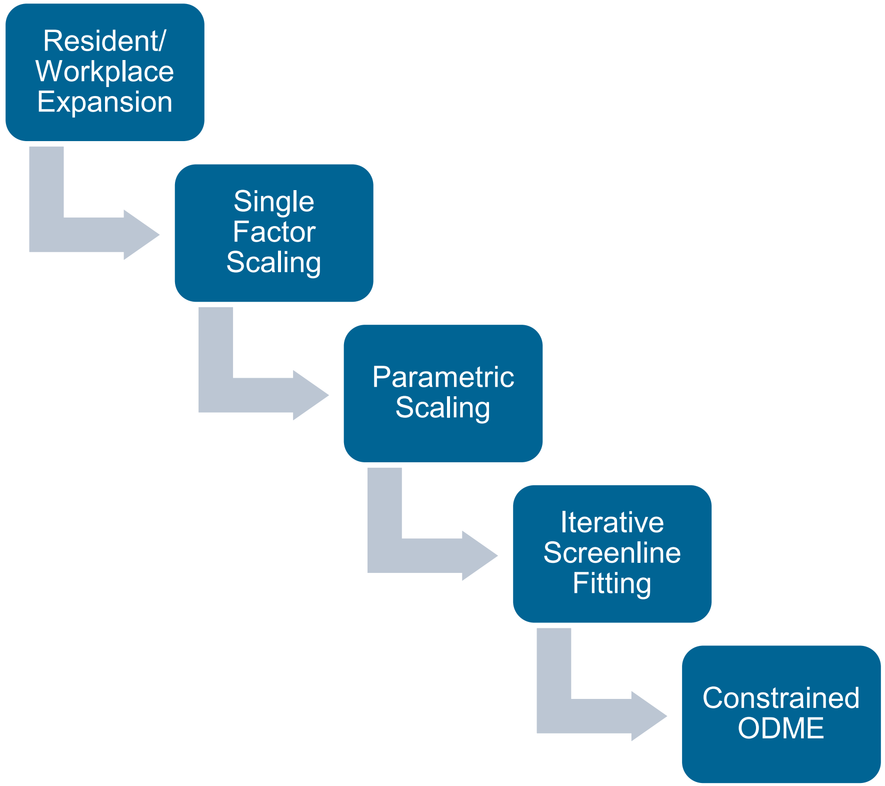
- Nationwide network with NC counts
- 3 Vehicle Classes:
  - Auto (LBS)
  - Multi-Unit Truck (ATRI)
  - Single-Unit Truck (synthetic)



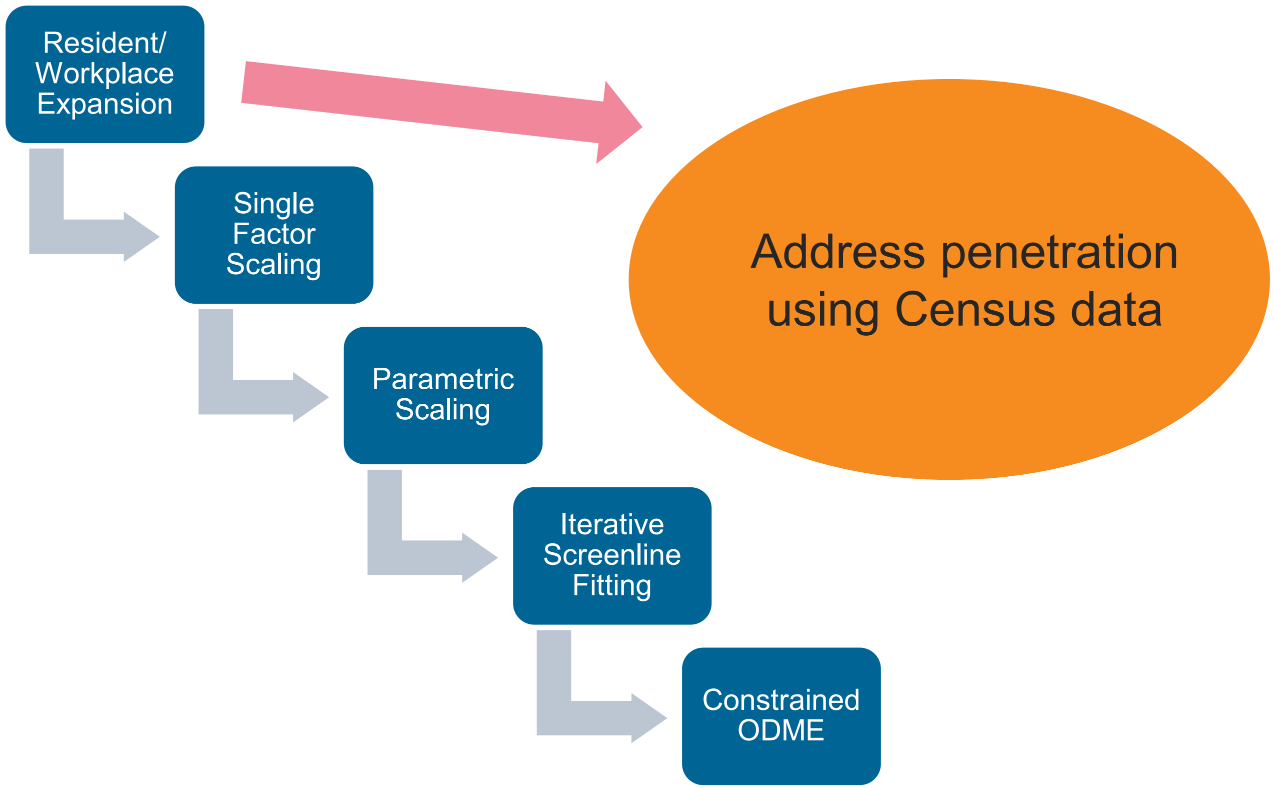
9,568 Links with AADT

Big Data Expansion

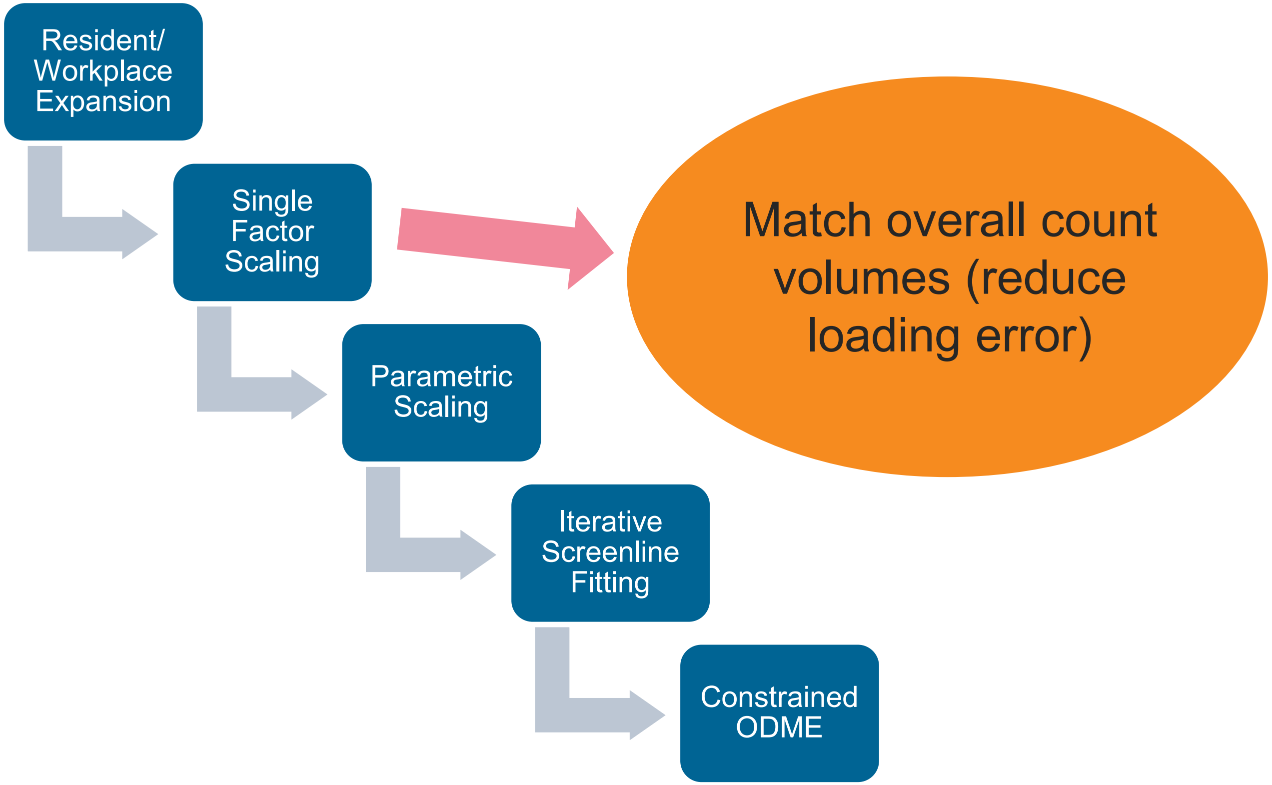
A multistep process to perform expansion of the passive LBS data



Big Data Expansion



Big Data Expansion

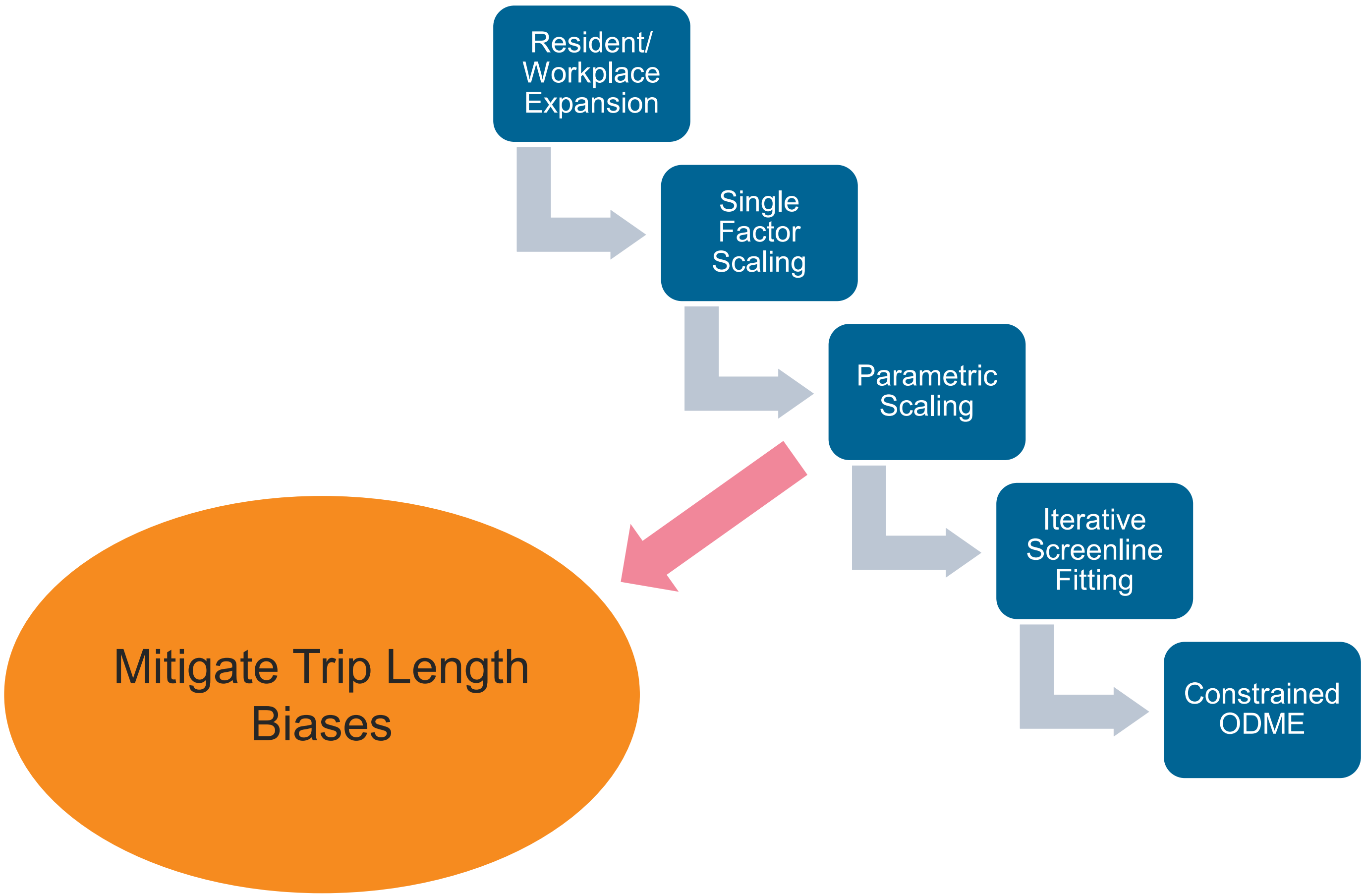


Single Factor Scaling

- Scaling by vehicle class
- Daily scaling factors
- Assignment by time period
- Iterative procedure

Statistic	All Vehicles
Loading Error (%)	3.35
RMSE (%)	47.58
MAPE (%)	49.79

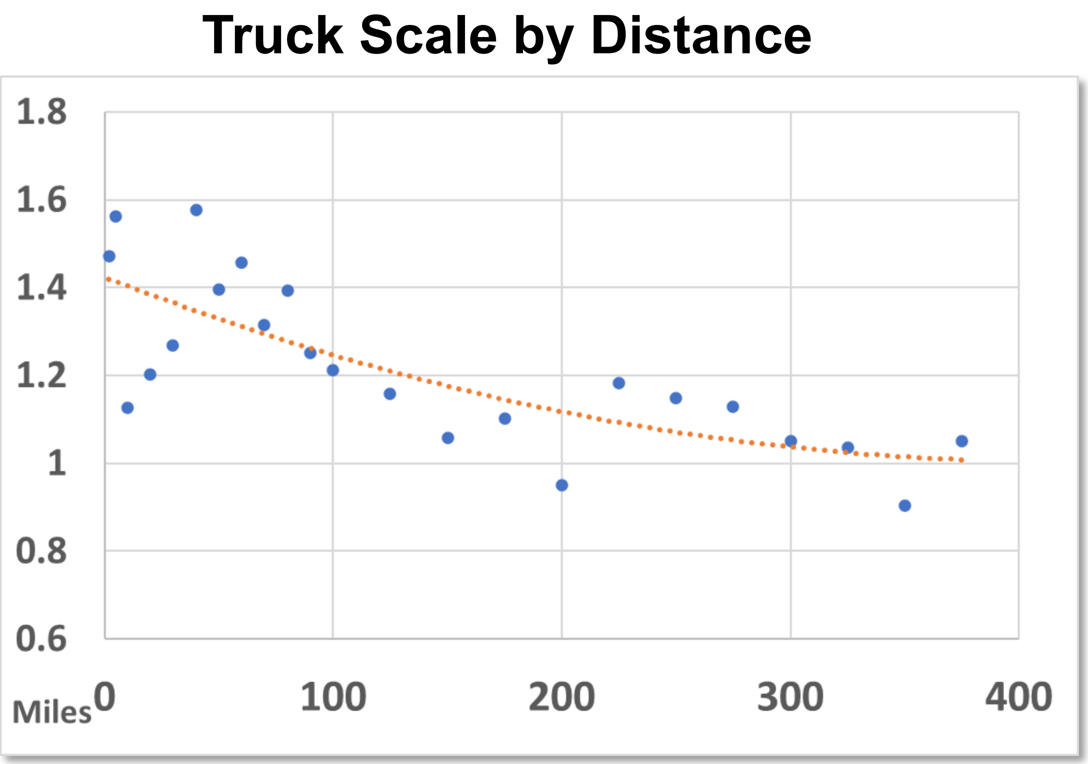
Big Data Expansion



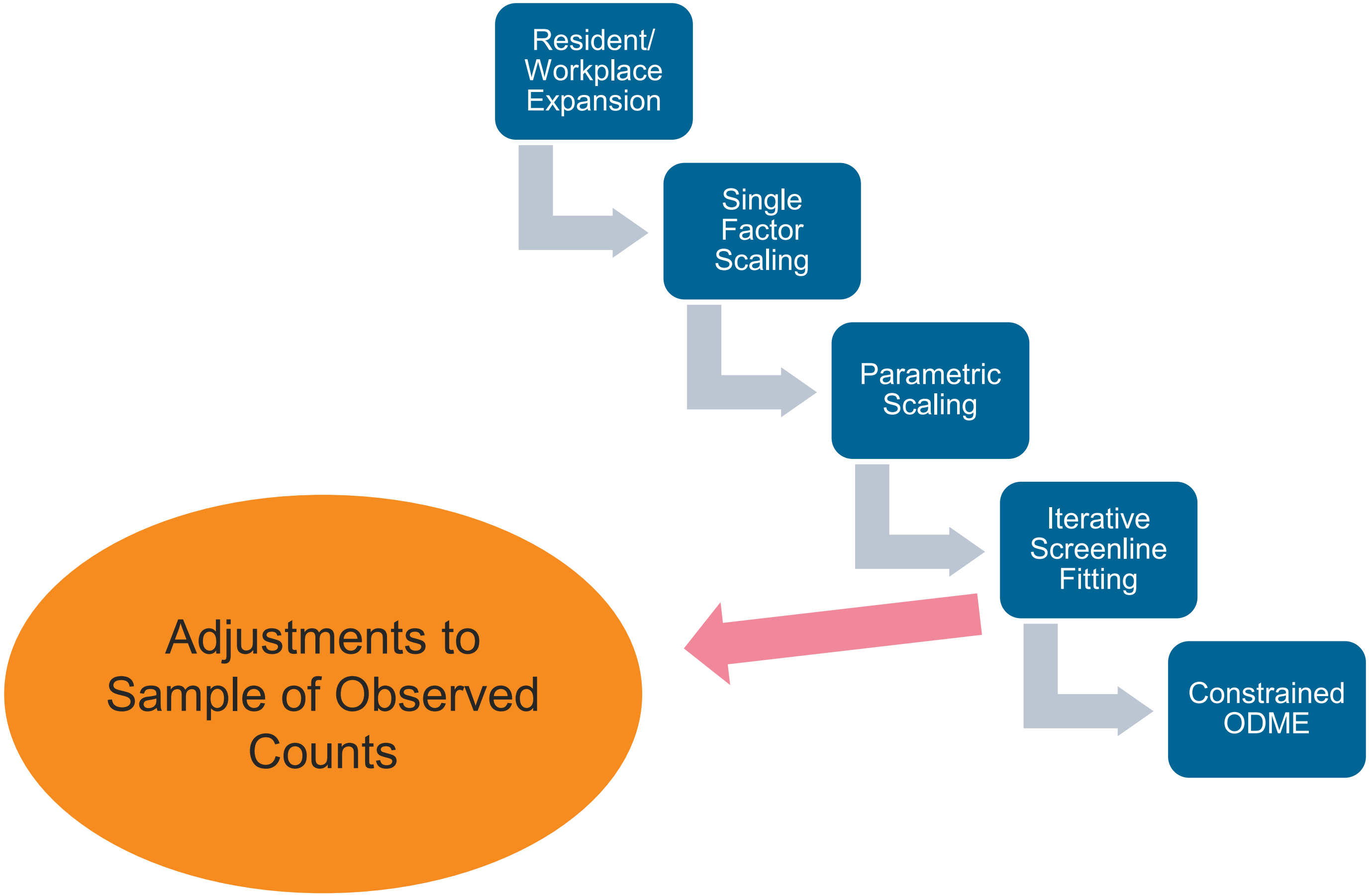
Parametric Scaling

- Scaling by vehicle class
- Independent variable: Trip length

Ratio of
New to Old Trips



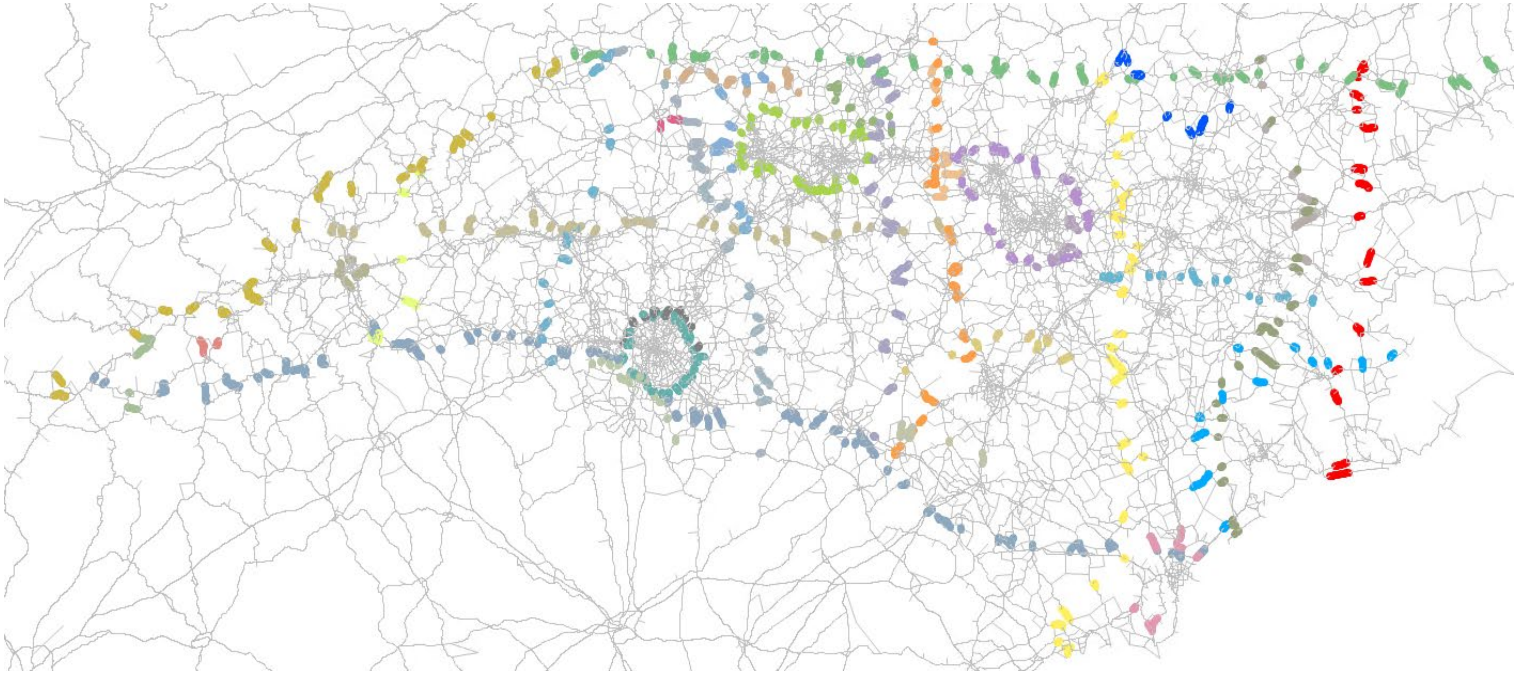
Big Data Expansion



Iterative Screenline Fitting (ISF)

Iterative Screenline Fitting (ISF)

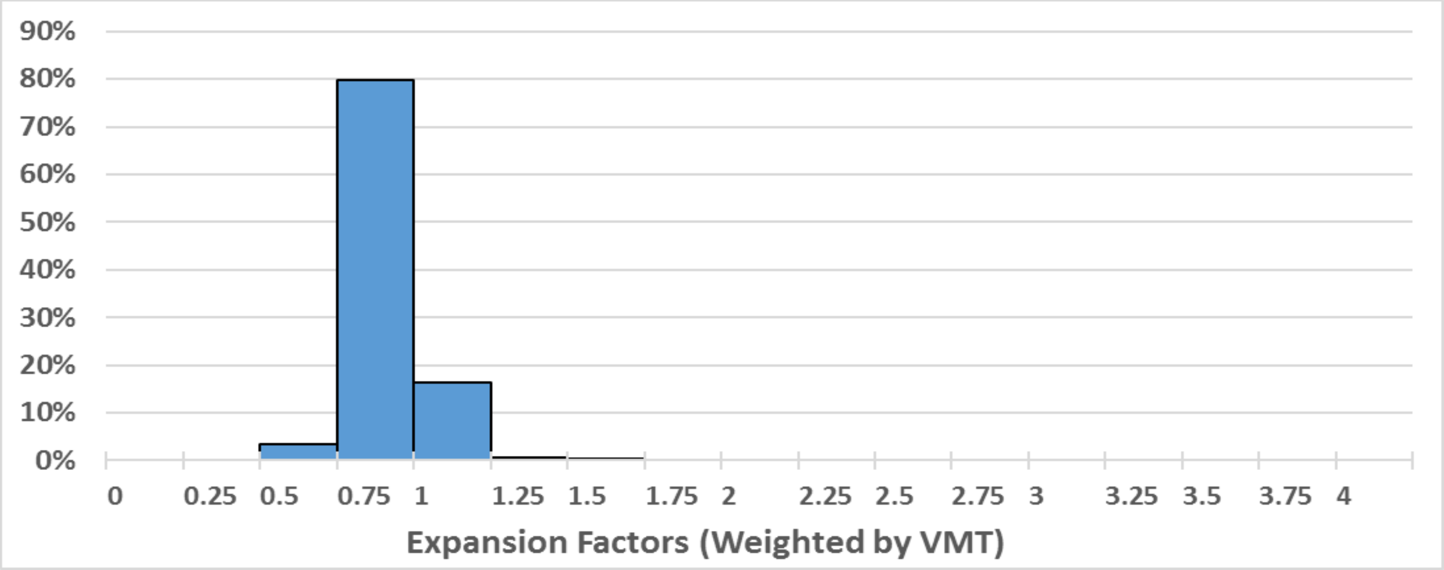
- 18 Screenlines
- 7 Cordons
- 32 Cutlines
- 5 Iterations



Iterative Screenline Fitting (ISF)

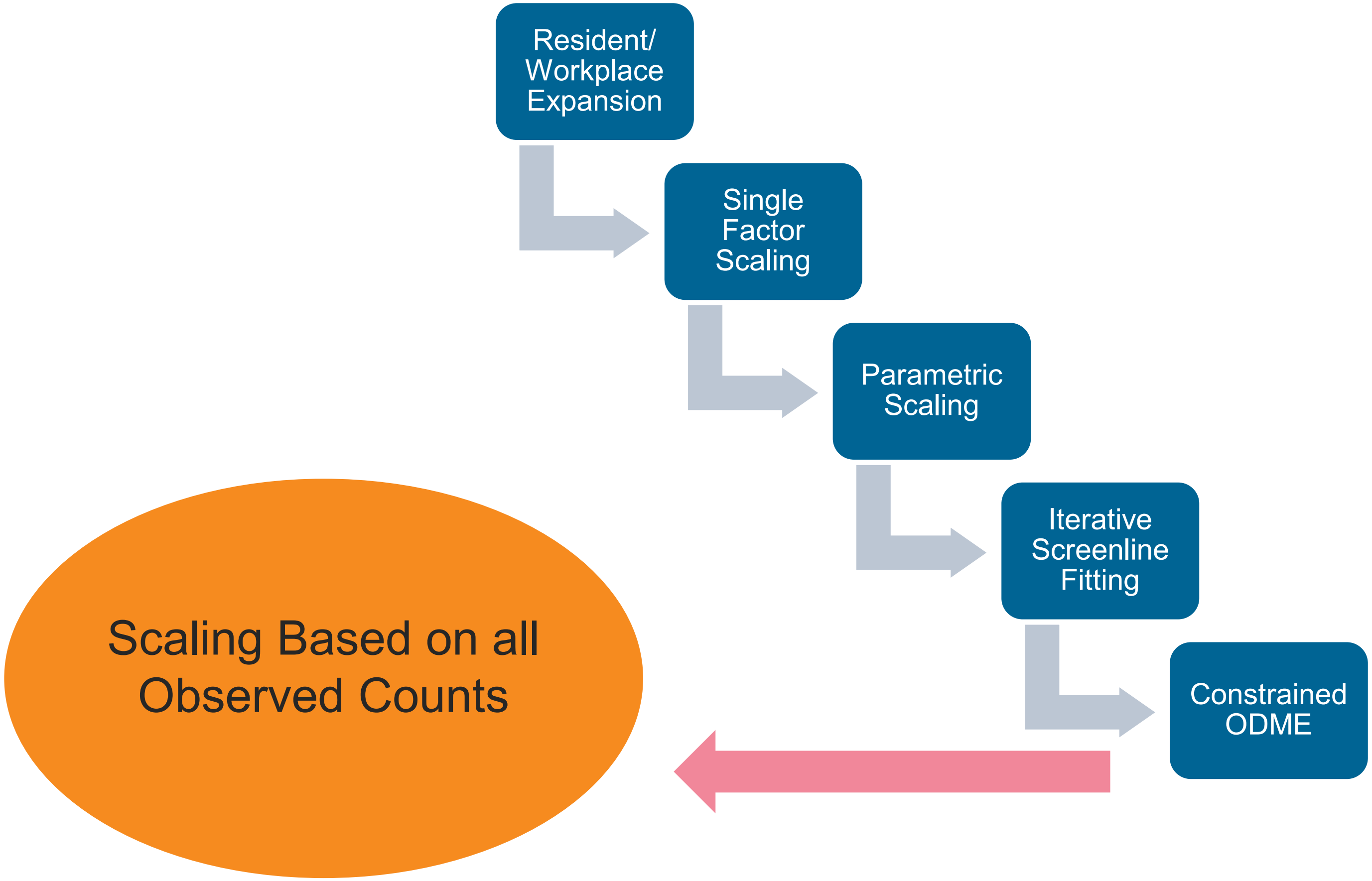
Statistic	All Vehicles
Loading Error (%)	0.23
RMSE (%)	41.66
MAPE (%)	44.62

Distribution of Expansion Weights



Big Data Expansion

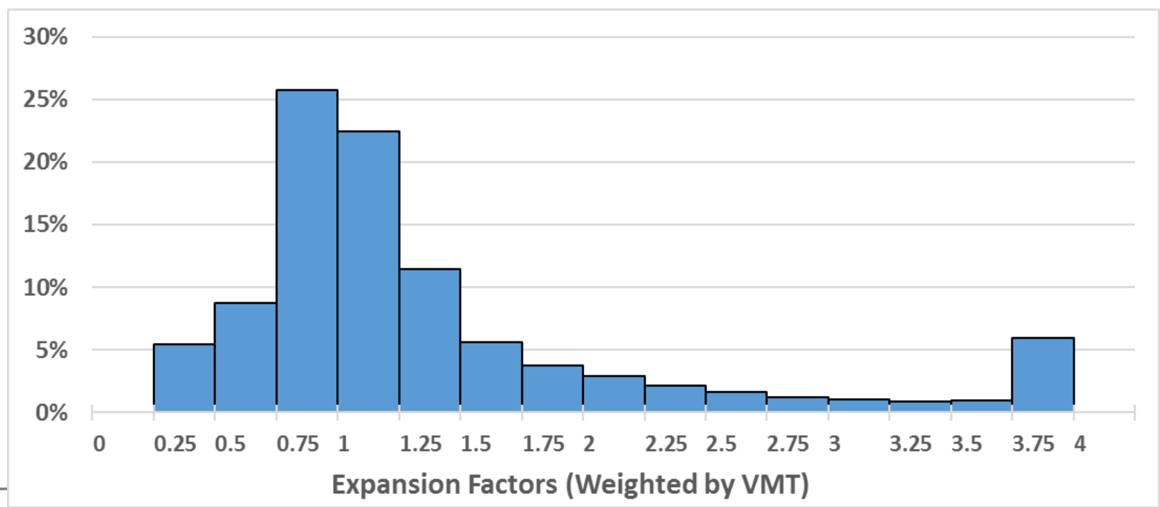
A multistep process was used to develop the final expansion of the passive OD data



Constrained ODME

Statistic	Auto	Trucks	All Vehicles
Loading Error (%)	-1.02	-9.85	-1.96
RMSE (%)	23.00	56.85	23.53
MAPE (%)	21.62	77.64	22.97

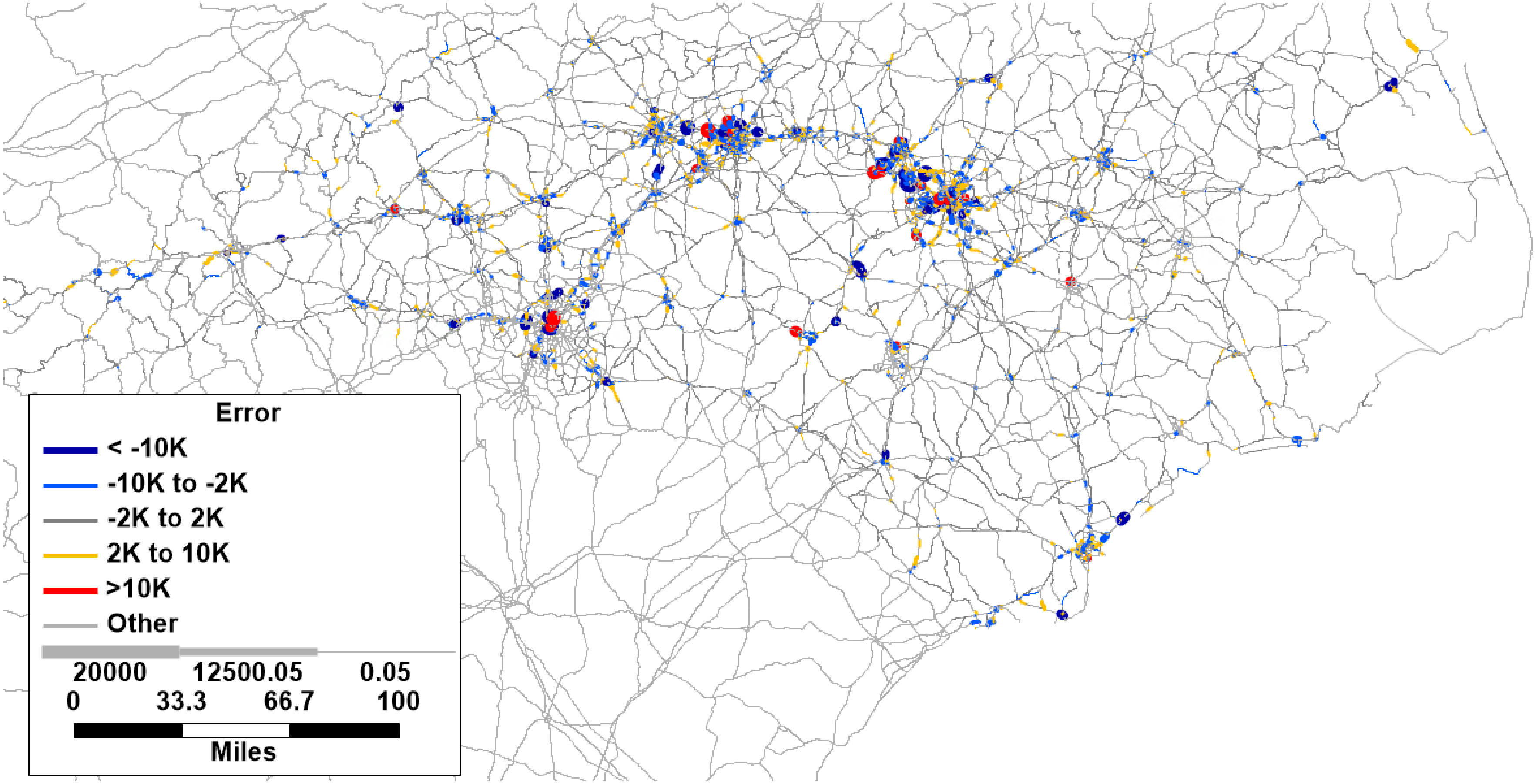
**After ODME there is a final single-factor scale to get error to about 0%*



Validation – Volume Group RMSE (%)

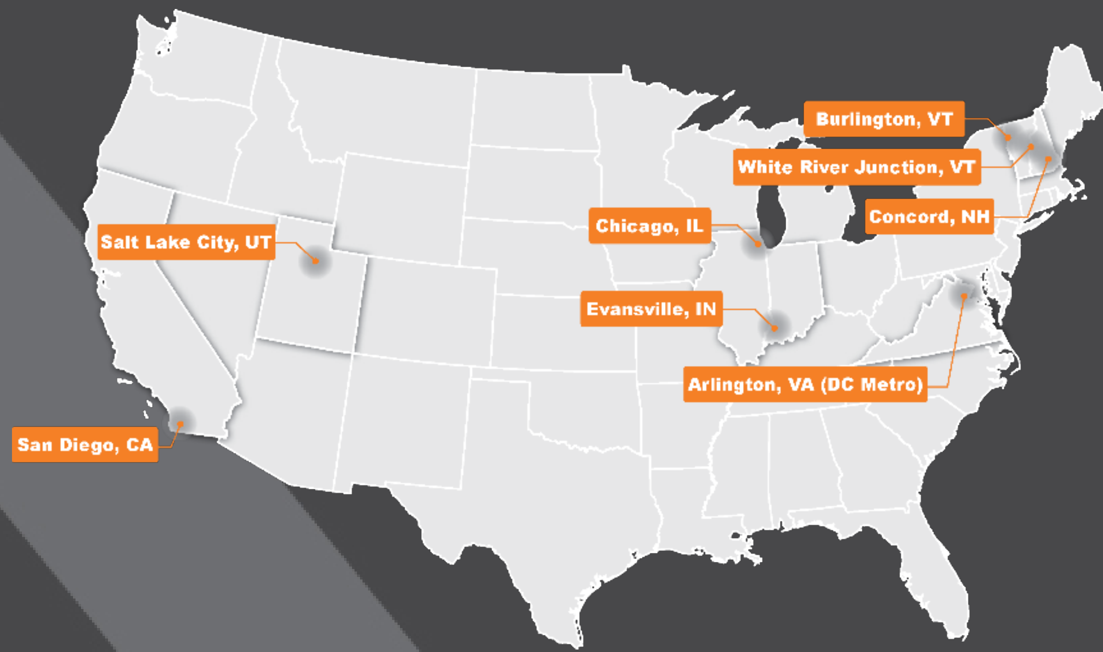
AADT	Expansion	Guideline
< 5,000	57.53	-
5k-10k	33.18	45.0
10k - 20k	24.09	40.0
20k - 40k	16.48	35.0
> 40k	8.87	30.0
Total	23.53	30-40

Model Loaded Network



Next Steps / COVID

- Passive Data can help measure & monitor changes due to the pandemic
- Metrics
- New Trip Matrices (efficient model update)
- Observe before/after changes in:
  - Quantity of trip Productions & Attractions by purpose (HBW, HBO, NHB, Long, Short, visitor, etc.)
  - Percent Stay-at-home
  - E-commerce/deliveries
  - Trip Distances
  - Spatial Distribution of Trips
  - Time of Day Distribution



Contact

www.rsginc.com

Steven Trevino
CONSULTANT
Steven.Trevino@rsginc.com

Stephen Tuttle
SENIOR CONSULTANT
Stephen.Tuttle@rsginc.com