Background
Extensible Modular Framework

Current Travel Patterns from Big Data

- Growth from Local / Regional Models
- Growth from Advanced Trip-Based Model
- Growth from Machine Learning Algorithms
- Growth from FHWA Freight & Long-Distance Models
- Growth from Activity-Based Model

Possible Future Phases

Phase 1
- Forecast A

Phase 2
- Forecast B
- Forecast C
- Forecast D
- Forecast E

etc.
Updating Base Year without Recalibration

Big Data: 2017 → 2020

Base ODs can be updated without recalibrating demand models

Growth from Local / Regional Models
Growth from Advanced Trip-Based Model
Growth from Machine Learning Algorithms
Growth from FHWA Freight & Long-Distance Models
Growth from Activity-Based Model

Forecast A
Forecast B
Forecast C
Forecast D
Forecast E

Phase 1
Phase 2
Possible Future Phases

etc.
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?

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

Big data from smartphone apps is the primary raw data source from which rMerge is derived.
Mobile Data Experience

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

2. **IDENTIFY TRIPS**
   - 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

3. **EXPAND TO REGION**
   - Remove bad devices & pings with noise reduction filters

4. **AGGREGATE**
   - 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

<table>
<thead>
<tr>
<th>LBS Data - October 2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sightings</td>
</tr>
<tr>
<td>Total Devices</td>
</tr>
<tr>
<td>Good Devices</td>
</tr>
<tr>
<td>Locations</td>
</tr>
<tr>
<td>Trips</td>
</tr>
</tbody>
</table>

- LBS data represents a sample of 8.3% of NC residents
Trips identified based on “stop” locations

1. PREPARE INPUT DATA

   a. CLUSTER PINGS
      - Remove pings with poor horizontal 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

2. IDENTIFY TRIPS

   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

3. EXPAND TO REGION

4. AGGREGATE
Expansion process matches regional counts and Census data

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**

- **Up to $25,000**
- **$25,000 to $49,999**
- **$50,000 to $74,999**
- **$75,000 to $100,000**
- **$100,000 or higher**

**AGE**

- **16-17 years**
- **18-24 years**
- **25-34 years**
- **35-44 years**
- **45-54 years**
- **55-64 years**
- **65-74 years**
- **75-84 years**
- **85 years or older**

[Bar charts showing income and age distributions for rMove/LBS and Census]
Data Verification: Duration vs. Smartphone Survey

The graph illustrates the frequency distribution of trip durations comparing LBS and rMove. The x-axis represents the trip duration in minutes, ranging from 5 to 65 minutes, while the y-axis shows the frequency percentage. The graph visually compares the two methods, with LBS represented in orange bars and rMove in gray bars. The data suggests a higher frequency of trips within the 5 to 20-minute range for both methods, with a noticeable decline in frequency as the duration increases.
Expansion process matches regional counts and Census data

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

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

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

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

**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

**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

**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

North Carolina Study Clusters
October, 2018
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: \( P = \frac{S_f}{S_b} B \)
2. Additive: \( P = B + (S_f - S_b) \)
3. Mixed / Average of above
# 8-Case Pivoting (Mixed)

<table>
<thead>
<tr>
<th>Case</th>
<th>Base (B)</th>
<th>Synthetic Base (S_b)</th>
<th>Synthetic Future (S_f)</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>&gt;0</td>
<td>Sf</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>&gt;0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4n</td>
<td>0</td>
<td>&gt;0</td>
<td>&gt;0 (&lt; X)</td>
<td>0</td>
</tr>
<tr>
<td>4e</td>
<td>0</td>
<td>&gt;0</td>
<td>&gt; 0 (&gt; X)</td>
<td>Sf - X</td>
</tr>
<tr>
<td>5</td>
<td>&gt;0</td>
<td>0</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>6</td>
<td>&gt;0</td>
<td>0</td>
<td>&gt;0</td>
<td>B + Sf</td>
</tr>
<tr>
<td>7</td>
<td>&gt;0</td>
<td>&gt;0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8n</td>
<td>&gt;0</td>
<td>&gt;0</td>
<td>&gt;0 (&lt;X)</td>
<td>B * (Sf/Sb)</td>
</tr>
</tbody>
</table>

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

1. Resident/Workplace Expansion
2. Single Factor Scaling
3. Iterative Screenline Fitting
4. Parametric Scaling
5. Constrained ODME
Big Data Expansion

- Resident/Workplace Expansion
- Single Factor Scaling
- Parametric Scaling
- Iterative Screenline Fitting
- Constrained ODME

Address penetration using Census data
Big Data Expansion

- Resident/Workplace Expansion
- Single Factor Scaling
- Parametric Scaling
- Iterative Screenline Fitting
- Constrained ODME

Match overall count volumes (reduce loading error)
Single Factor Scaling

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

<table>
<thead>
<tr>
<th>Statistic</th>
<th>All Vehicles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loading Error (%)</td>
<td>3.35</td>
</tr>
<tr>
<td>RMSE (%)</td>
<td>47.58</td>
</tr>
<tr>
<td>MAPE (%)</td>
<td>49.79</td>
</tr>
</tbody>
</table>
Big Data Expansion

- Resident/Workplace Expansion
  - Single Factor Scaling
  - Parametric Scaling
    - Iterative Screenline Fitting
    - Constrained ODME

Mitigate Trip Length Biases
Parametric Scaling

- Scaling by vehicle class
- Independent variable: Trip length

**Truck Scale by Distance**

Ratio of New to Old Trips
Big Data Expansion

Adjustments to Sample of Observed Counts
Iterative Screenline Fitting (ISF)

- 18 Screenlines
- 7 Cordons
- 32 Cutlines
- 5 Iterations
Iterative Screenline Fitting (ISF)

<table>
<thead>
<tr>
<th>Statistic</th>
<th>All Vehicles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loading Error (%)</td>
<td>0.23</td>
</tr>
<tr>
<td>RMSE (%)</td>
<td>41.66</td>
</tr>
<tr>
<td>MAPE (%)</td>
<td>44.62</td>
</tr>
</tbody>
</table>

Distribution of Expansion Weights

[Bar chart showing the distribution of expansion weights with expansion factors weighted by VMT.]

0% 10% 20% 30% 40% 50% 60% 70% 80% 90%
0 0.25 0.5 0.75 1 1.25 1.5 1.75 2 2.25 2.5 2.75 3 3.25 3.5 3.75 4

Expansion Factors (Weighted by VMT)
Big Data Expansion

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

- Resident/Workplace Expansion
- Single Factor Scaling
- Parametric Scaling
- Iterative Screenline Fitting
- Constrained ODME

Scaling Based on all Observed Counts
Constrained ODME

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Auto</th>
<th>Trucks</th>
<th>All Vehicles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loading Error (%)</td>
<td>-1.02</td>
<td>-9.85</td>
<td>-1.96</td>
</tr>
<tr>
<td>RMSE (%)</td>
<td>23.00</td>
<td>56.85</td>
<td>23.53</td>
</tr>
<tr>
<td>MAPE (%)</td>
<td>21.62</td>
<td>77.64</td>
<td>22.97</td>
</tr>
</tbody>
</table>

*After ODME there is a final single-factor scale to get error to about 0%*
Validation – Volume Group RMSE (%)

<table>
<thead>
<tr>
<th>AADT</th>
<th>Expansion</th>
<th>Guideline</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 5,000</td>
<td>57.53</td>
<td>-</td>
</tr>
<tr>
<td>5k-10k</td>
<td>33.18</td>
<td>45.0</td>
</tr>
<tr>
<td>10k - 20k</td>
<td>24.09</td>
<td>40.0</td>
</tr>
<tr>
<td>20k - 40k</td>
<td>16.48</td>
<td>35.0</td>
</tr>
<tr>
<td>&gt; 40k</td>
<td>8.87</td>
<td>30.0</td>
</tr>
<tr>
<td>Total</td>
<td>23.53</td>
<td>30-40</td>
</tr>
</tbody>
</table>
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