TMIP

Better Methods. Better Outcomes.

How-To: Develop Big Data-Driven Demand for | Traffic Forecasting

Date:

February 21, 2018

Vince Bernardin, Ph.D.

Nagendra Dhakar, Ph.D.

U.S. Department of Transportation

Federal Highway Administration

About TMIP

- The Travel Model Improvement Program (TMIP) is a program within the FHWA Office of Planning, Environment and Realty (HEP).
- TMIP has conducted research, provided technical assistance, and delivered training to local, regional, and state transportation planning professionals since 1994.
- Today, TMIP continues its mission of improving analysis practices to ensure that transportation professionals are well equipped to inform and support strategic transportation decisions.

Disclaimer

The views expressed in this document do not represent the opinions of FHWA and do not constitute an endorsement, recommendation or specification by FHWA. The document is based solely on the research conducted by RSG.

Acknowledgement

This volume is a collaboration between transportation professionals at FHWA, FTA, the Tennessee Department of Transportation, and RSG.

Introduction

Data Driven Traffic Forecasting

- Not a new idea, data driven methods in NCHRP 255 and NCHRP 765
  - Pivoting off of traffic counts
  - Using traffic counts to improve OD matrices

Data Driven Traffic Forecasting

- Not a new idea, data driven methods in NCHRP 255 and NCHRP 765
  - Pivoting off of traffic counts
  - Using traffic counts to improve OD matrices

- So, what has changed?

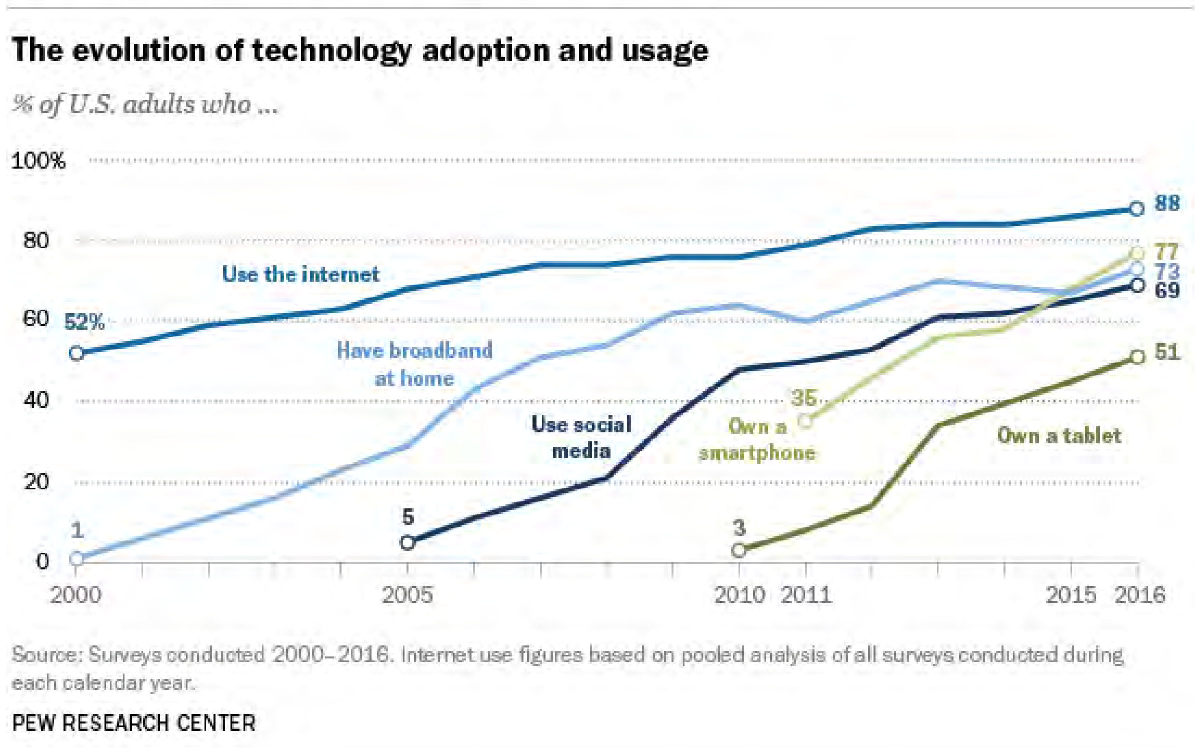
TMIP

Smartphone Ownership

- Pew Research Center Data
  - 92% of those 18-29, 88% of those 30-49 have smartphones

Percent of US Adults (18+) who own a smartphone

Nov 2016	77%
June 2015	68%
Jan 2014	58%
May 2013	56%
Feb 2012	46%
May 2011	35%



TMIP

Data Driven Traffic Forecasting

- Not a new idea, data driven methods in NCHRP 255 and NCHRP 765
  - Pivoting off of traffic counts
  - Using traffic counts to improve OD matrices

- So, what has changed?
  - Mobile devices now passively provide an entirely new kind of big data for forecasting

Focus on Demand (OD Flows)

- OD Flows, where people are going (to & from) – modeled with either gravity or destination choice models – is the largest source of error in travel forecasting models (by far). [Zhao & Kockelman, 2002]

- Huge solution space – at least 500k up to > 10 Million

- Limited explanatory variables for modeling

Why Getting OD Flows Right Matters

- We have to know where people are going to and from in order to know:
  - If they would pay a toll
  - If they might change modes (and walk, ride transit…)
  - If deadheading restrictions on automated vehicles would be effective

Data Fusion

- Traffic counts are still necessary for expanding passive OD data and ensuring its representativeness

- Traffic Counts < Passive ODs < Passive ODs + Traffic Counts

Overview of the How-To

- Traffic Counts
- Passive OD Data
- Combining Counts & OD Data
- Data Driven Traffic Forecasting & Modeling

- TDOT Statewide Model Proof of Concept

Traffic Counts

The Need for Data Validation

- Traffic counts provide information on the total traffic on a road
- Errors in count data
  - Sample error
  - Counter devices/technology
  - Data processing
- Need to validate count data before usage
- Checks for consistency of the traffic count data

Count Consistency Checking Tool

- Three types of checks
  - Logical consistency with other roadway attributes
  - Internal temporal consistency **(not in the tool)**
  - Internal spatial consistency
- Automated methods to identify inconsistencies in a highway network count database
- Tennessee Department of Transportation (TDOT) travel demand model
- TransCAD and GISDK scripting

Logical Consistency with Other Roadway Attributes

$$\{Threshold\ low\} * capacity <= count <= \{Threshold\ high\} * capacity$$

Output

Msg	Label
0	No count
1	Count is reasonable
2	Count is low
3	Count is high
4	Count/capacity is not available
5	Count is on unexpected direction

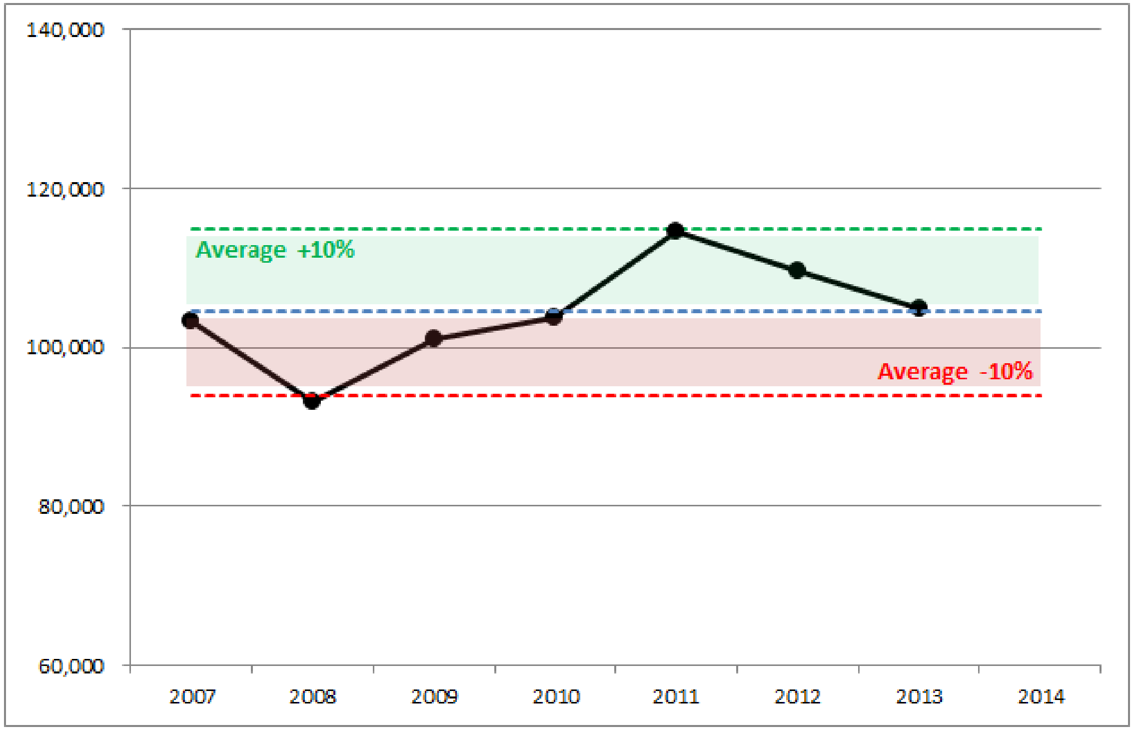
Internal Temporal Consistency

- First, throw out any bad years
  - For each station calculate front weighted mean
    - 2012 = 5   2011 = 4   2010 = 3   2009 = 2   2008 = 1
- Compare each year's count with the weighted mean for possible removal
  - Volume < 1,000 - acceptable error = +/- 200%
  - Volume < 2,500 - acceptable error = +/- 100%
  - Volume < 5,000 - acceptable error = +/- 50%
  - Volume < 10,000 - acceptable error = +/- 25%
  - Volume < 25,000 - acceptable error = +/- 20%
  - Volume < 50,000 - acceptable error = +/- 15%
  - Volume > 50,000 - acceptable error = +/- 10%
- Second, throw out bad / erratic stations
  - Coefficient of Variation (CV) was calculated once all the outlier AADTs for each station and each year had been removed.
  - For stations with only 2012 data, Coefficient of Variation (CV) was calculated by adding the year 2013 data.
  - Stations were dropped if CV was > 15% and if standard deviation was > 100.

TDOT Internal Temporal Consistency Results

- A total of 213 stations were removed (out of 12,297) due to this process as either being outliers or otherwise suspicious data

STATION_ID = 37000317; I-40 in Davidson County

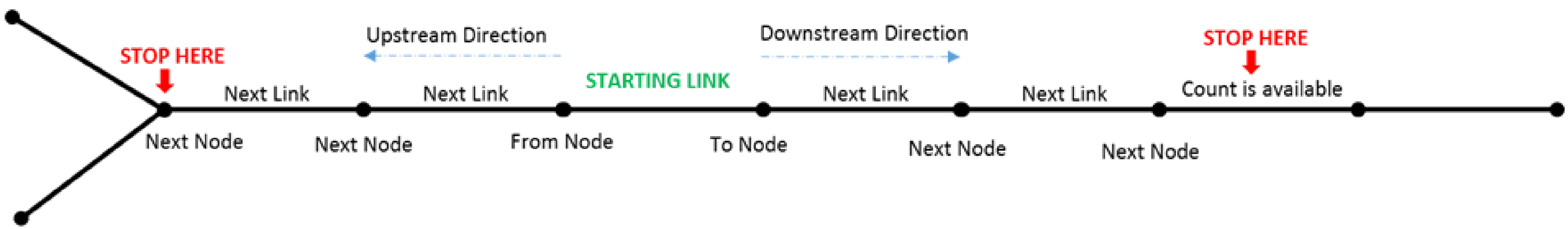


Internal Spatial Consistency

- Count propagation
- Conservation of flow based checks
    - Intersection-level check
    - Intersection approach-level check #1
    - Intersection approach-level check #2
    - Intersection turning movement check
- Guidance for intersections with missing data

Count Propagation

- Assign (propagate) counts from coded links to other links of the roadway segment



Output

Msg	Label
0	No count
1	Existing coded count
2	Propagated count
3	Conflicting counts

Count Propagation - Coverage Statistics

1. Network-level
   - Count type
   - Link type
2. After count propagation
   - Functional class
   - Link type
3. Interchanges
   - Interchange type
   - Count availability
4. ODME Summary

NETWORK SUMMARY

Existing Counts

LinkType	NumLinks	(%)	LaneMiles	(%)	Avg_AADT_AB	Avg_AADT_BA	Avg_Cap_AB	Avg_Cap_BA
With Counts	012,830	10.90%	02,658.36	08.07%	03,152.74	04,588.72	24,500.50	20,610.48
Without Counts	104,891	89.10%	30,289.84	91.93%	00,000.00	00,952.00	25,041.21	20,508.15
TOTAL	117,721	100.00%	32,948.20	100.00%				

Propagated Counts

LinkType	NumLinks	(%)	LaneMiles	(%)	Avg_AADT_AB	Avg_AADT_BA	Avg_Cap_AB	Avg_Cap_BA
With Counts	077,097	65.49%	25,113.96	76.22%	02,815.59	04,147.70	23,306.73	20,002.47
Without Counts	040,624	34.51%	07,834.24	23.78%	00,000.00	00,000.00	28,162.17	21,500.15
TOTAL	117,721	100.00%	32,948.20	100.00%				

By Functional Class

FUNCCLASS	NumLinks	(%)	LaneMiles	(%)	Avg_AADT_AB	Avg_AADT_BA	Avg_Cap_AB	Avg_Cap_BA
01	002,362	78.65%	01,252.34	90.60%	18,465.63	16,829.20	053,466.23	044,889.23
02	005,154	67.42%	01,823.20	69.49%	04,106.39	04,241.36	037,350.84	026,009.62
06	007,113	72.64%	02,543.03	76.60%	02,354.94	02,381.50	018,351.37	016,824.81
07	010,736	80.84%	04,369.09	84.33%	00,985.19	00,988.26	015,703.39	015,536.34
08	020,213	87.89%	09,396.85	88.59%	00,417.47	00,418.38	015,634.53	015,617.88
09	000,289	74.10%	00,134.29	78.57%	00,635.34	00,635.34	015,685.61	015,791.05
10	000,011	00.64%	00,001.94	00.76%	01,703.14	01,992.33	011,216.31	006,479.87
11	002,525	54.49%	00,577.92	68.12%	41,089.37	38,714.37	069,510.96	057,408.74
12	000,929	50.49%	00,190.02	61.39%	16,330.54	19,865.86	056,375.33	041,913.59
14	007,731	50.87%	01,262.00	56.60%	09,575.67	09,558.75	034,988.62	024,171.94
16	010,154	55.42%	01,684.07	61.26%	04,429.45	04,531.83	020,727.35	017,934.51
17	009,390	68.93%	01,783.77	74.15%	01,698.71	01,705.06	015,155.38	014,446.36
19	000,456	74.15%	00,093.58	79.99%	01,479.78	01,479.99	014,912.52	014,607.55
20	000,018	00.41%	00,001.47	00.27%	10,991.22	05,252.50	012,128.39	002,400.65
91	000,016	27.59%	00,000.39	33.44%	04,087.58	02,748.88	006,594.76	001,637.11
TOTAL	077,097	65.49%	25,113.96	76.22%				

INTERCHANGE SUMMARY

Type	CountsOnAll	(%)	MissingOneApproach	(%)	MissingMore	(%)	TOTAL
3-way	01,125	07.26%	04,377	28.23%	10,003	64.51%	15,505
4-way	00,179	03.74%	00,580	12.11%	04,030	84.15%	04,789
5-way	00,000	00.00%	00,000	00.00%	00,026	100.00%	00,026

ODME SUMMARY

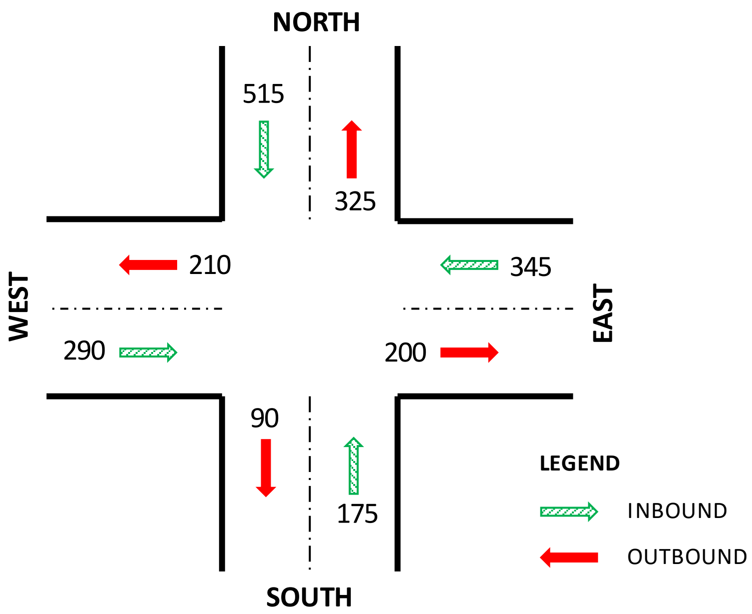
	Avg. Number of Shortest Paths
With Counts	055,055
Without Counts	052,007

NOTE: these summaries are from TN Statewide Model Network

Conservation of Flow Checks

- Intersection-level check

"Total flow entering an intersection is equal to total flow exiting the intersection."



Total inbound flow = 1,325
(515+345+175+290)

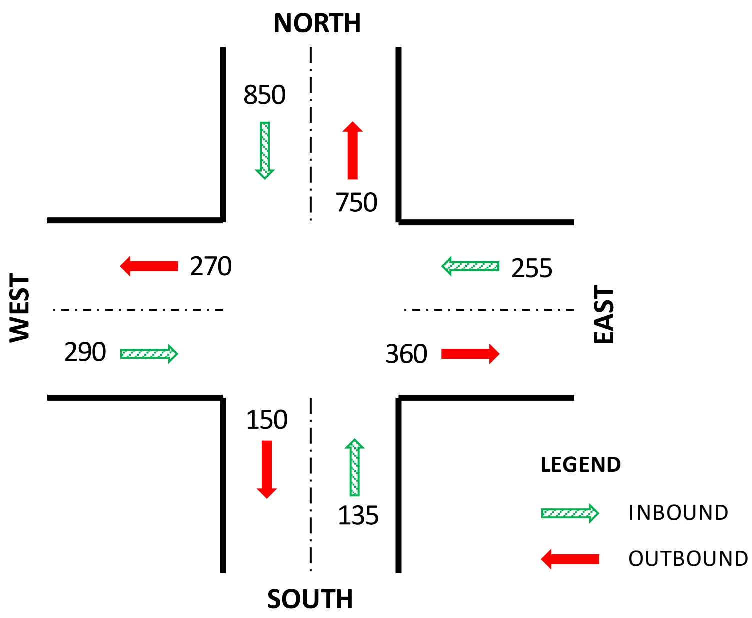
Total outbound flow = 825
(325+200+90+210)

Message code =1 ("Total flow entering the junction is not equal to the total flow exiting the junction")

Conservation of Flow Checks

- Intersection approach-level check #1

"Inbound AADT from a leg is less than the summation of outbound AADTs from other legs of that intersection."



Total inbound flow (1530) = Total outbound flow (1530)

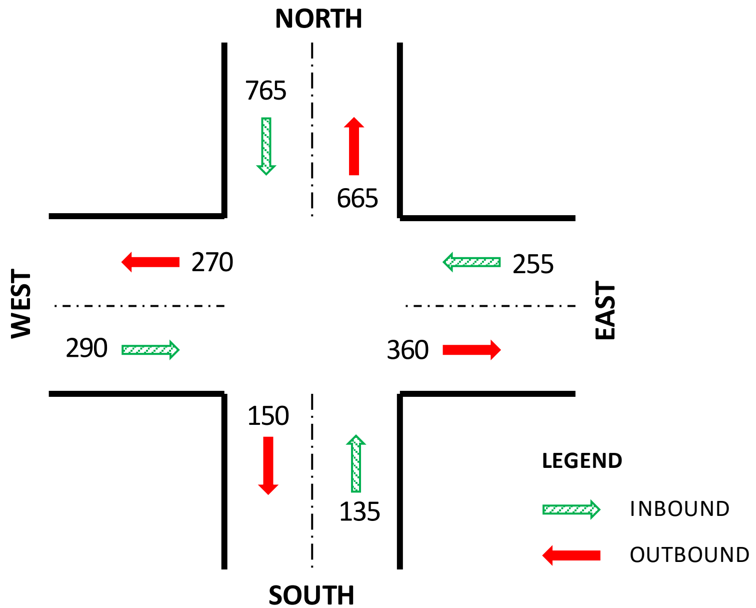
North leg inbound flow (850) is > Sum of outbound flows (780) from East, South, and West legs

Message code =2 ("Inbound flow is not less than the sum of outbound flows from other legs")

Conservation of Flow Checks

- Intersection approach-level check #2

"The ratio of inbound AADT from a particular leg and the summation of outbound AADTs from other legs is significantly less than one."



Total inbound flow (1445) = Total outbound flow (1445)

Inbound flows from any leg is less than the sum of outbound flows from other legs

(Inbound flow from North leg) / (sum of outbound flows from the other legs = 0.981 > 0.9 (threshold)

Message code =3 ("Ratio of inbound flows and sum of outbound flows from other legs is too high")

Conservation of Flow Checks

Output

Msg	Label
0	Passed intersection checks
1	Total flow entering the junction is not equal to the total flow exiting the junction
2	Inbound flow is not less than the sum of outbound flows from other legs
3	Ratio of inbound flows and sum of outbound flows from other legs is too high

Conservation of Flow Checks

- Intersection turning movement check

$$Gap = \sum_{j=1}^{N}\left[1-\left(OUT\_AADT_j / \sum_{i=1}^{N} Turn\ Movement_i\right)\right]$$

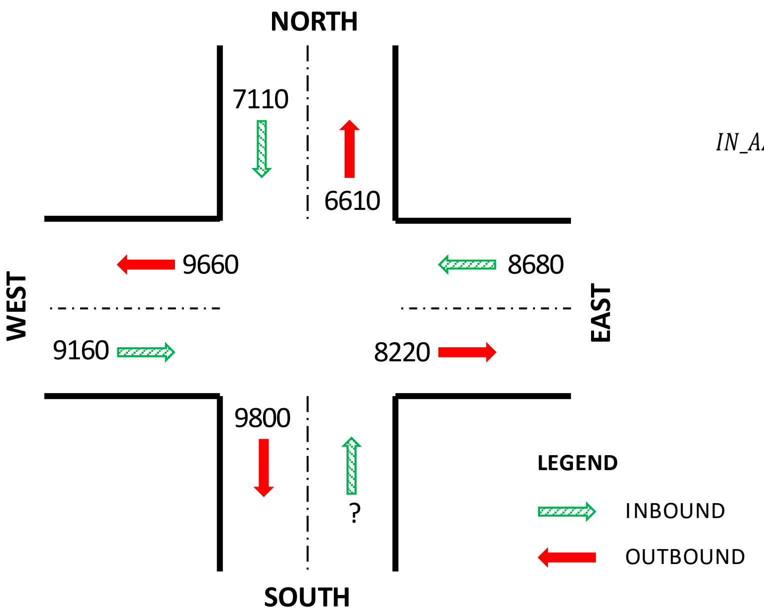
Output

Msg	Label
0	Completed
1	Flows from one or more legs are too high to calculate turning movements
2	Turning movements cannot be calculated; please check input flows

Intersections with Missing Data

- A missing inbound count

$$IN\_AADT_i = \sum_{j=1}^{N} OUT\_AADT_j - \sum_{i \neq j, j=1}^{N} IN\_AADT_j$$



$$IN\_AADT_S = (OUT\_AADT_N + OUT\_AADT_E + OUT\_AADT_W + OUT\_AADT_S) - (IN\_AADT_N + IN\_AADT_E + IN\_AADT_W)$$

$$IN\_AADT_S = (6{,}610 + 8{,}220 + 9{,}660 + 9{,}800) - (7{,}110 + 8{,}680 + 9{,}160)$$

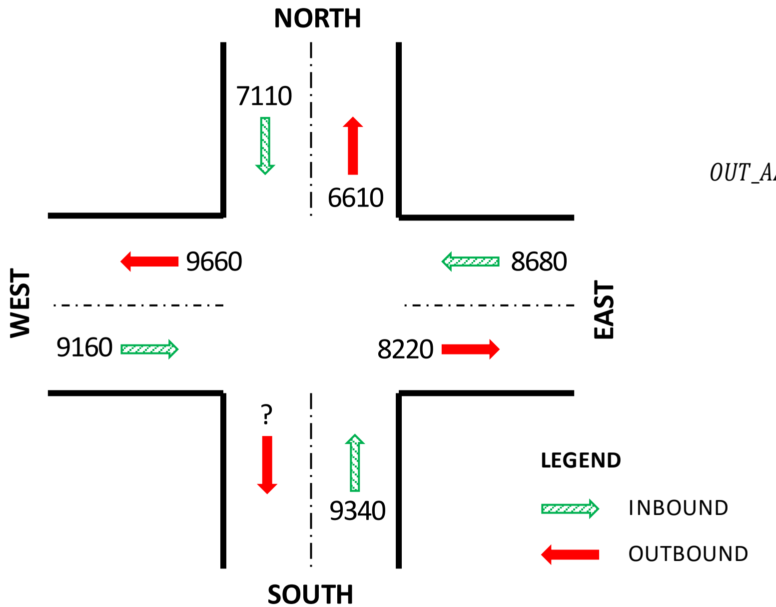
$$IN\_AADT_S = 9{,}340$$

If negative value, message code =2 ("Cannot calculate a missing inbound count—total outbound is less than total inbound")

Intersections with Missing Data

- A missing outbound count

$$OUT\_AADT_i = \sum_{j=1}^{N} IN\_AADT_j - \sum_{i \neq j, j=1}^{N} OUT\_AADT_j$$



$$OUT\_AADT_S = (IN\_AADT_N + IN\_AADT_E + IN\_AADT_W + IN\_AADT_S) - (OUT\_AADT_N + OUT\_AADT_E + OUT\_AADT_W)$$

$$OUT\_AADT_S = (7{,}110 + 8{,}680 + 9{,}160 + 9{,}340) - (6{,}610 + 8{,}220 + 9{,}660)$$

$$OUT\_AADT_S = 9{,}800$$

If negative value, message code =4 ("Cannot calculate a missing outbound count—total inbound is less than total outbound")

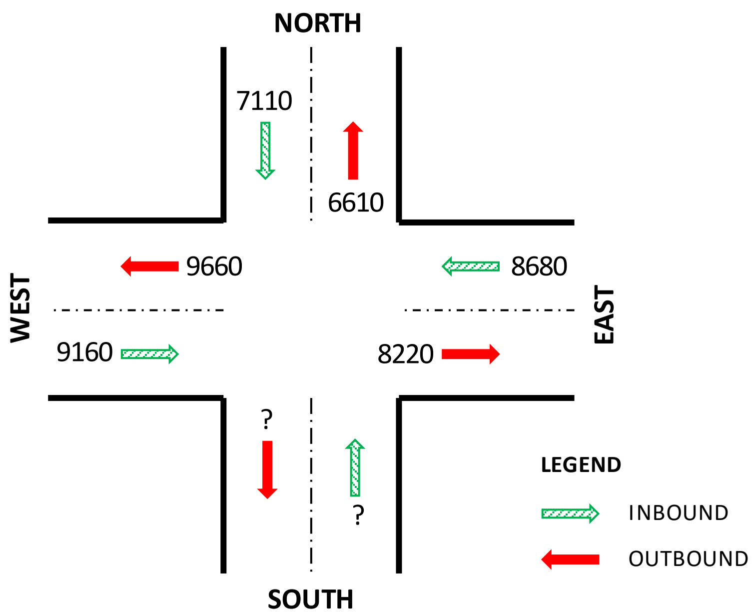
Intersections with Missing Data

- Missing one approach counts

$$n\times\Big(\sum_{j=1}^{N} OUT\_AADT_j\Big) \le IN\_AADT_i \le m\times\left(\sum_{j=1}^{N} OUT\_AADT_j\right), where\ 0<n<m<1\ and\ j\#i$$

and

$$n\times\Big(\sum_{j=1}^{N} IN\_AADT_j\Big) \le OUT\_AADT_i \le m\times\Big(\sum_{j=1}^{N} IN\_AADT_j\Big), where\ 0<n<m<1\ and\ j\#i$$



For n=0.1 and m=0.9

$$2{,}449 \le IN\_AADT_S \le 22{,}041 \quad and \quad 2{,}495 \le OUT_{AADT_S} \le 22{,}455$$

For n=0.2 and m=0.4

$$4{,}898 \le IN\_AADT_S \le 9{,}796 \quad and \quad 4{,}990 \le OUT_{AADT_S} \le 9{,}980$$

Message code =5 ("Calculated missing approach counts (both inbound and outbound")

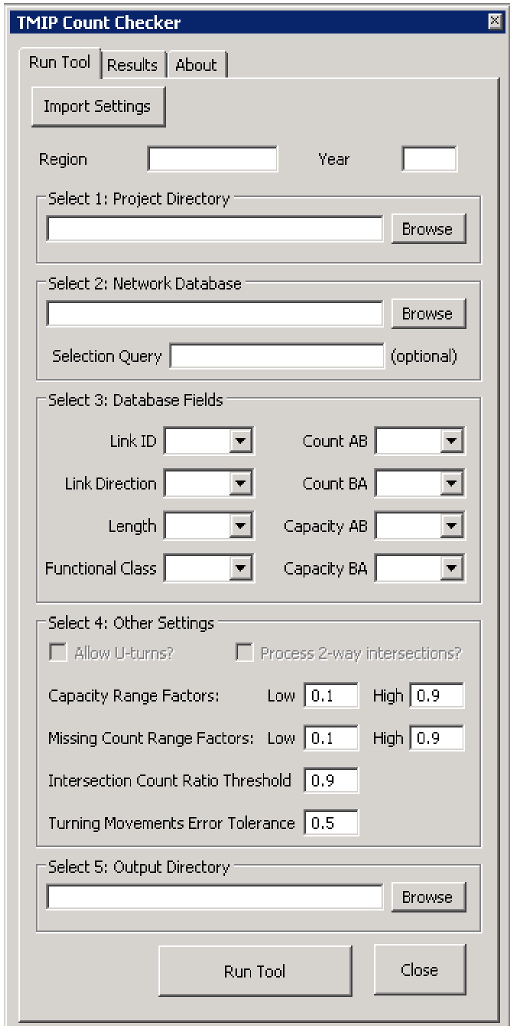
Intersections with Missing Data

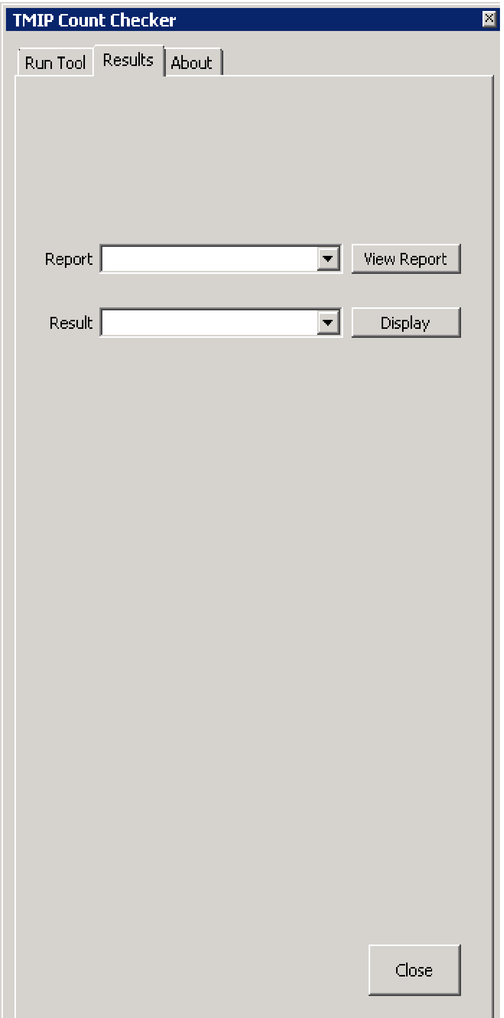
- Output (missing one approach counts)

Msg	Label
1	Calculated a missing inbound count
2	Cannot calculate a missing inbound count: total outbound is less than total inbound
3	Calculated a missing outbound count
4	Cannot calculate a missing outbound count: total inbound is lower than total outbound
5	Calculated missing approach counts (both inbound and outbound)
6	Cannot calculate: inbound/outbound flow is not available

- Output (missing more than one approach counts)
  - Link ids

The Tool


TMIP Count Checker
Run Tool
Results
About
Import Settings
Region
Year
Select 1: Project Directory
Browse
Select 2: Network Database
Browse
Selection Query
(optional)
Select 3: Database Fields
Link ID
Count AB
Link Direction
Count BA
Length
Capacity AB
Functional Class
Capacity BA
Select 4: Other Settings
Allow U-turns?
Process 2-way intersections?
Capacity Range Factors:
Low 0.1
High 0.9
Missing Count Range Factors:
Low 0.1
High 0.9
Intersection Count Ratio Threshold 0.9
Turning Movements Error Tolerance 0.5
Select 5: Output Directory
Browse
Run Tool
Close


TMIP Count Checker
Run Tool
Results
About
Report
View Report
Result
Display
Close

Passive Origin-Destination Data

Types of Passive OD Data

- Four Types of Passive OD Data
  - Cellular Tower Signaling
  - LBS (Location Based Services / App Data)
  - GPS (Global Positioning Systems
  - Bluetooth
- Each type of data has advantages and disadvantages
  - The best dataset can depend on the application
  - Key considerations (including those presented here) vary both across regions and over time

Types of Passive OD Data

Description	Cell-Tower Signaling	LBS	GPS	Bluetooth
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; multihour periods within the day	Average weekday or average weekend or individual day of week; multihour 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; sometimes disaggregate traces also available but with restricted use	Aggregate trip ODs; sometimes disaggregate traces also available but with restricted use	Disaggregate trip ODs
OD Travel Time Data (Including Reliability)	Not possible	Not commercially available	Available with varying degrees of processing effort depending on provider	Generally produced as part of the processing of trips

Precision and Penetration

Precision and 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	1 month to multiple years depending on provider and pricing	1 month to multiple years depending on provider and pricing	Typically <1 month
Coverage Issues	Poor coverage in some (mostly rural) areas	--	--	Coverage limited—requires mounting detector devices

Representativeness

Representativeness and Expansion	Cell-Tower Signaling	LBS	GPS	Bluetooth
Trip-Length / Duration Bias	Confirmed	Confirmed	Confirmed	Not suspected
Demographic Bias	Present but mild and easily corrected	Moderate Age and Income Biases	Severe Income Bias and Some Age Bias; difficult to correct	Not well understood, believed to be moderate but difficult to correct
Included/Default Expansion	Residence market share-based; generally requires further correction	None/single count-based factor, generally requires further correction	None/single count-based factor, generally requires further correction	Typically expanded to counts

Segmentation and Applications

Segmentation and 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	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	Depending on provider may be possible	Depending on provider may be possible	Possible as a postprocess
Residency Information	Premium options for regional residents vs. nonresidents or home block groups	Premium options for regional residents vs. nonresidents	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 available	From some providers Heavy and medium trucks, private vehicles	Generally not possible

Resource Requirements

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

Considerations on Types of Big OD Data

- Different types of data are different
  - Important to know what can and cannot be done with each type
  - Do you want / need direct or indirect corridor level info?
  - Are long-distance or visitor trips important? Traveler demographics? Mode?
- Sample penetration and/or sample penetration x time period is how much information you are getting
  - This is what you're paying for
  - Precision limits what you can do with it
  - Sample penetration may vary by region & over time

Considerations on Types of Big OD Data

- Looking forward, think about datasets that are most likely to support data future and serve as a baseline for future retrospectives
- Representativeness is the big 'gotcha'
  that you will have to fix
- Realize you have to budget for data expansion
- Consider full cost, including processing, not just data
- Buy what you need, not more, not less

TDOT Passive OD Data Sets

TDOT Datasets & Processing

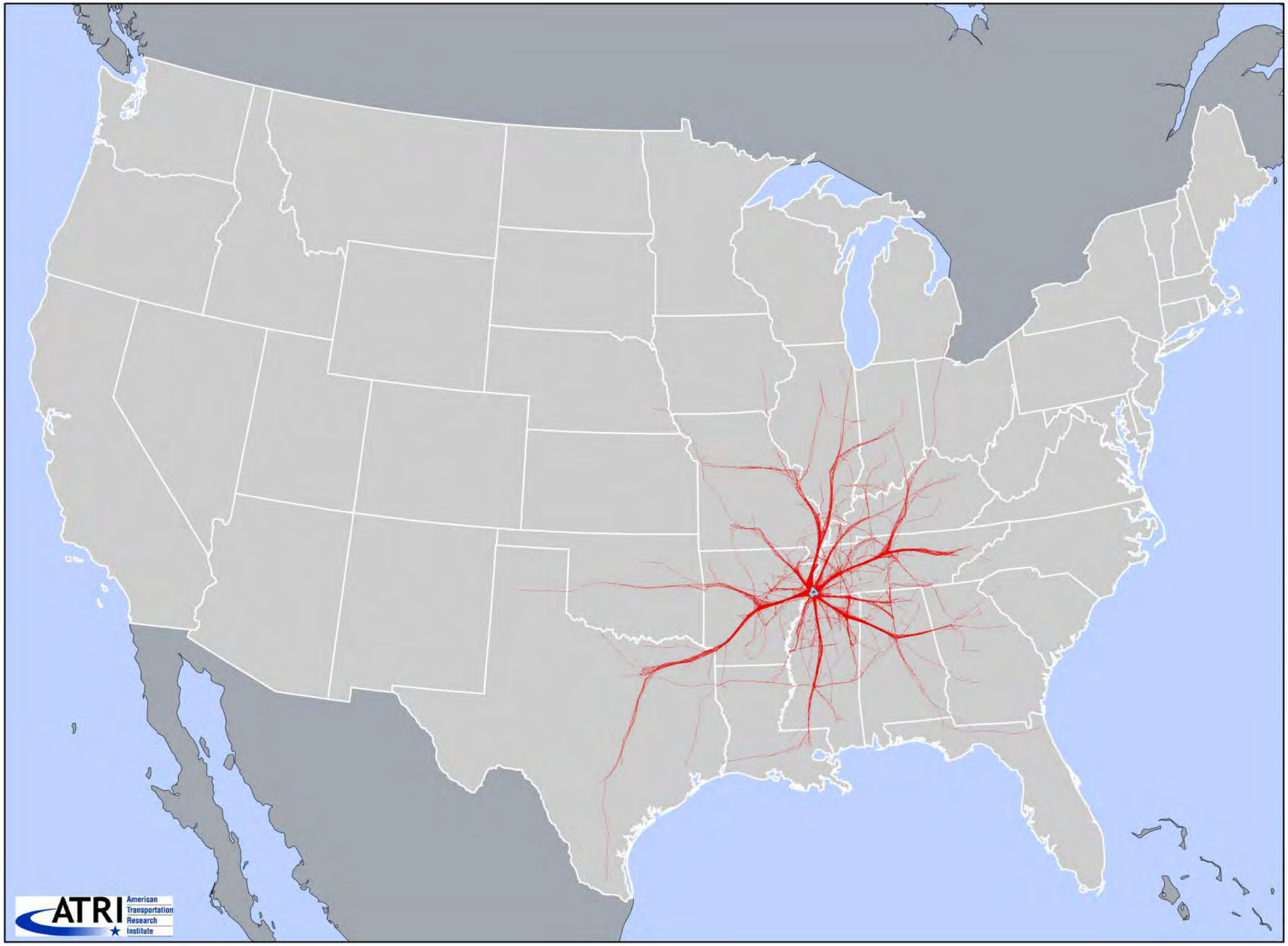
- Total OD matrix from AirSage with demographic expansion
- Truck GPS trace data from ATRI processed to ODs
- Removal of trucks from total ODs to estimate passenger ODs
- Expansion of both truck and passenger ODs to correct for trip duration bias

TDOT ATRI Dataset

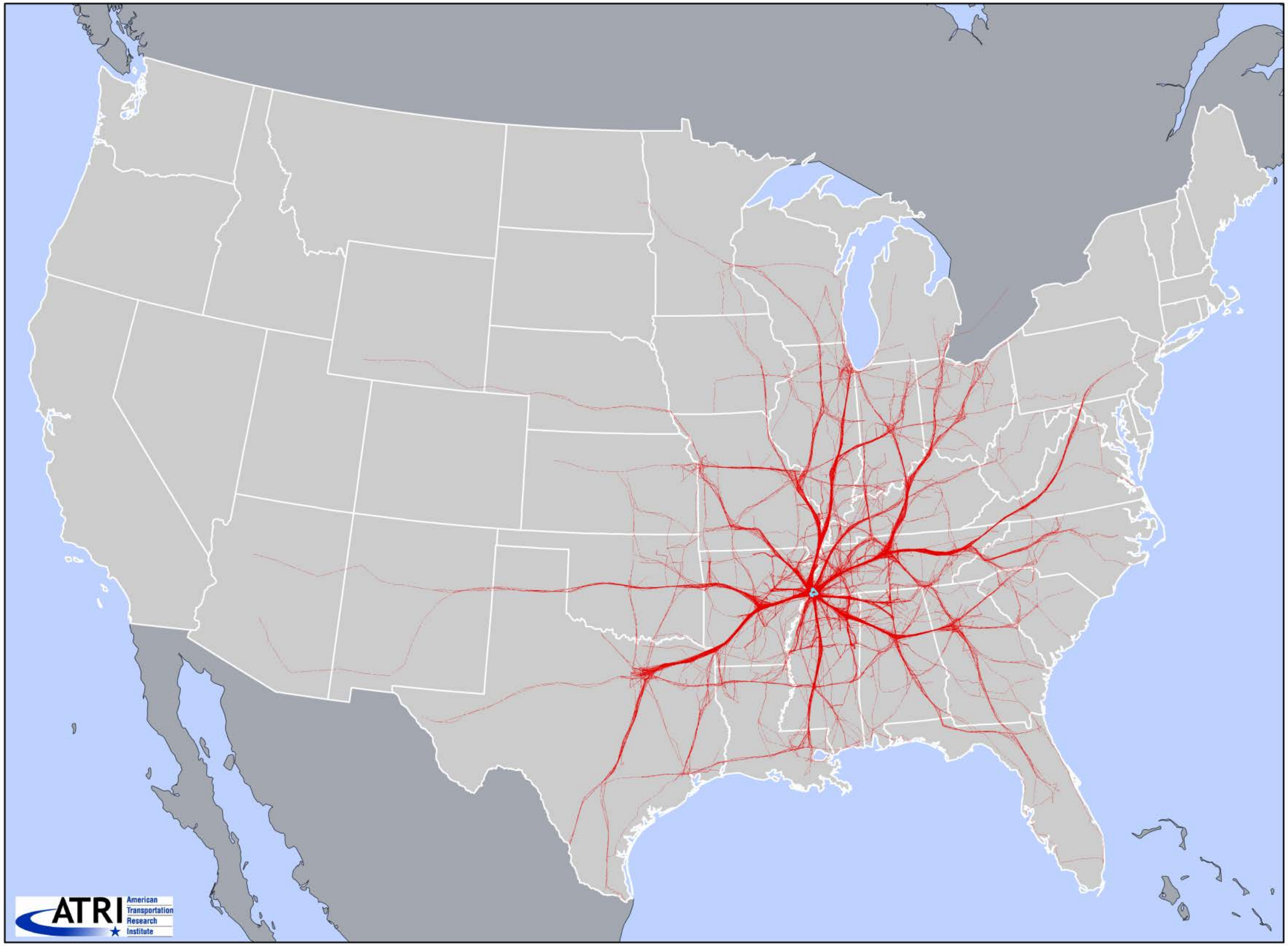
- Four 2-week samples over 2013 quarters
- 235,000 unique trucks
- 138 million records processed to 5.8 million trips
- 84,147,030 truck VMT within TN
- **Sample rate of 10.7% of multi-unit trucks**

Memphis - 1,000 Truck Sample

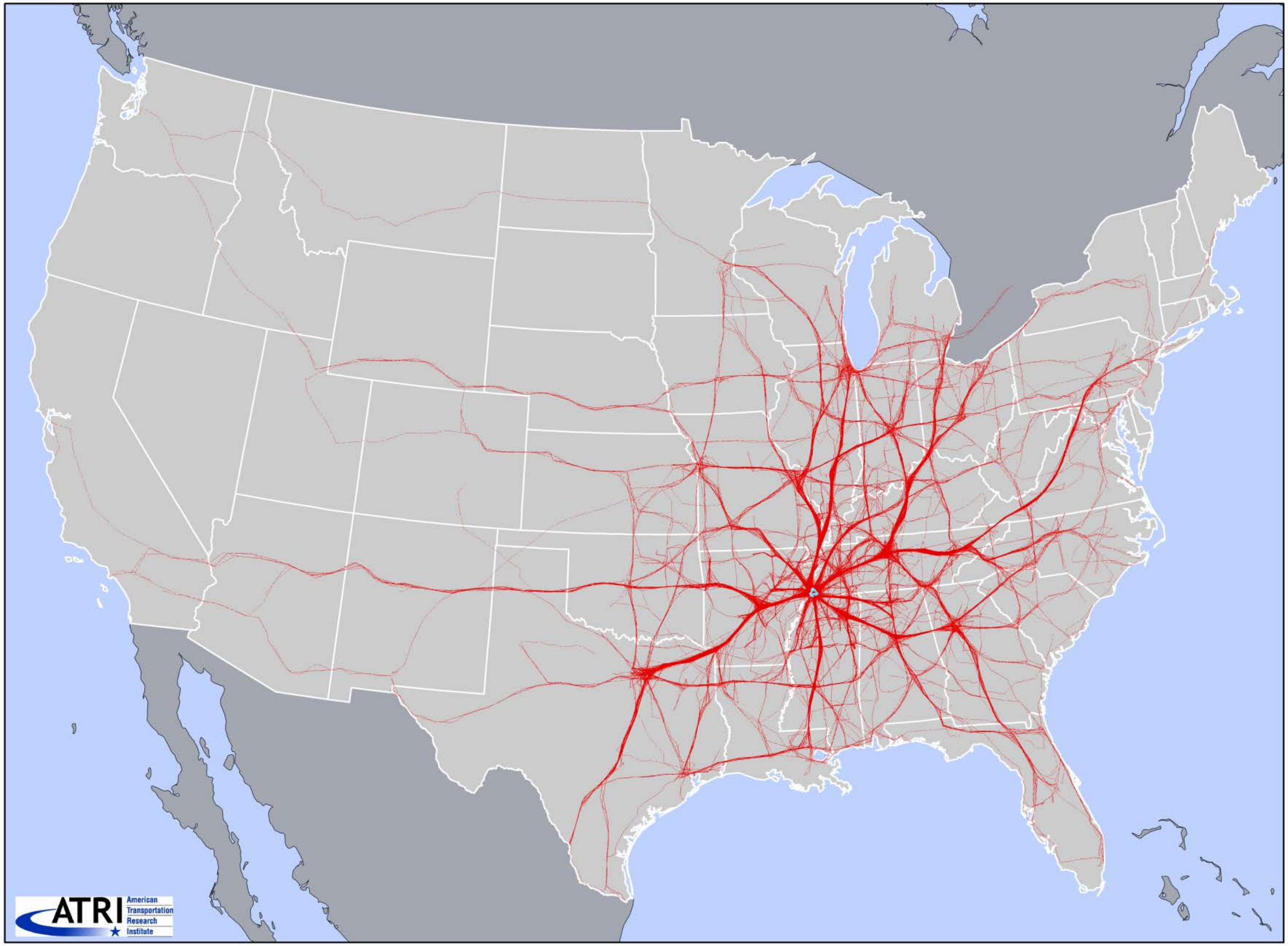
Same 1,000 Trucks After 24 Hours



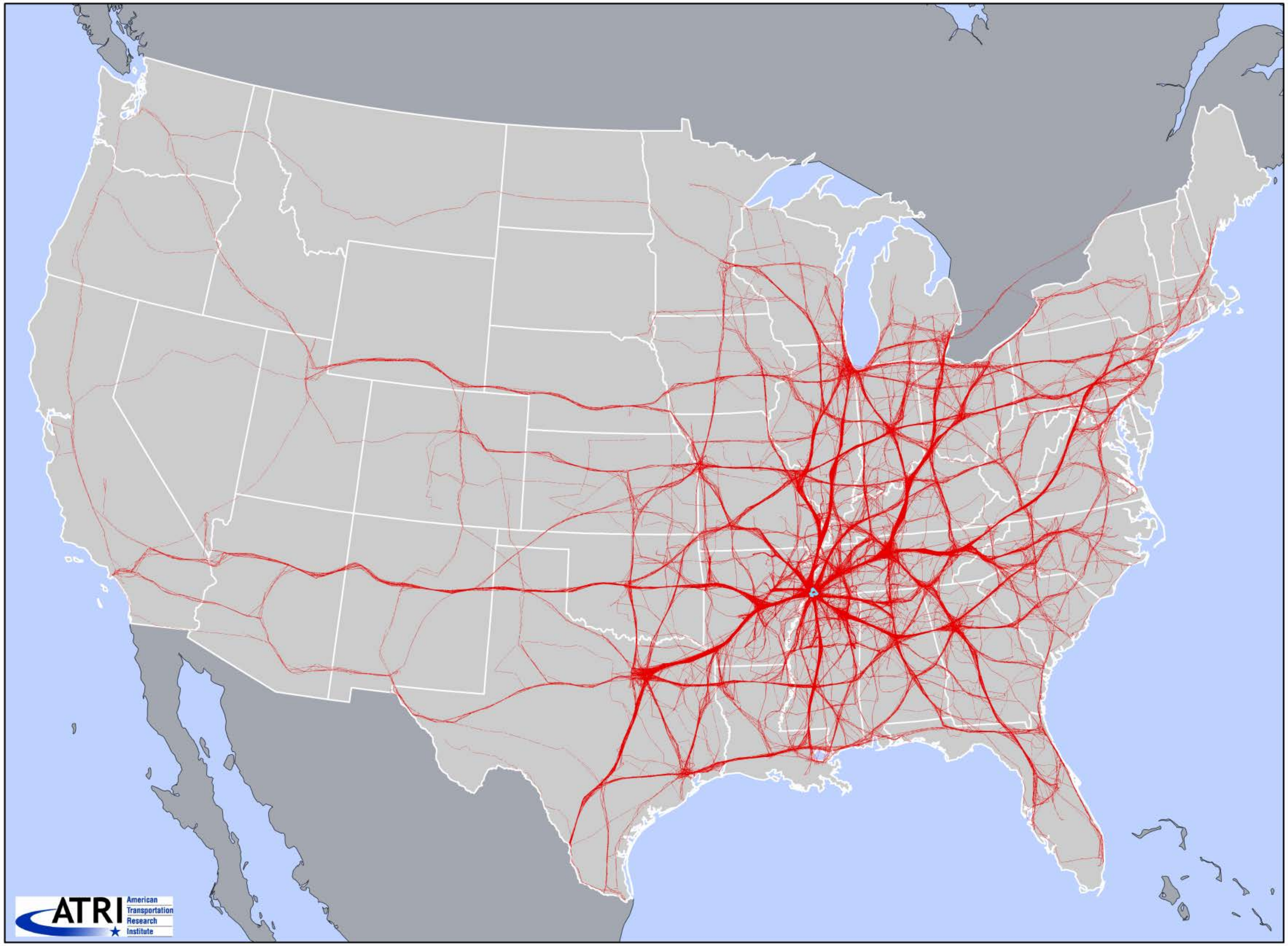
Same 1,000 Trucks After 48 Hours



Same 1,000 Trucks After 72 Hours

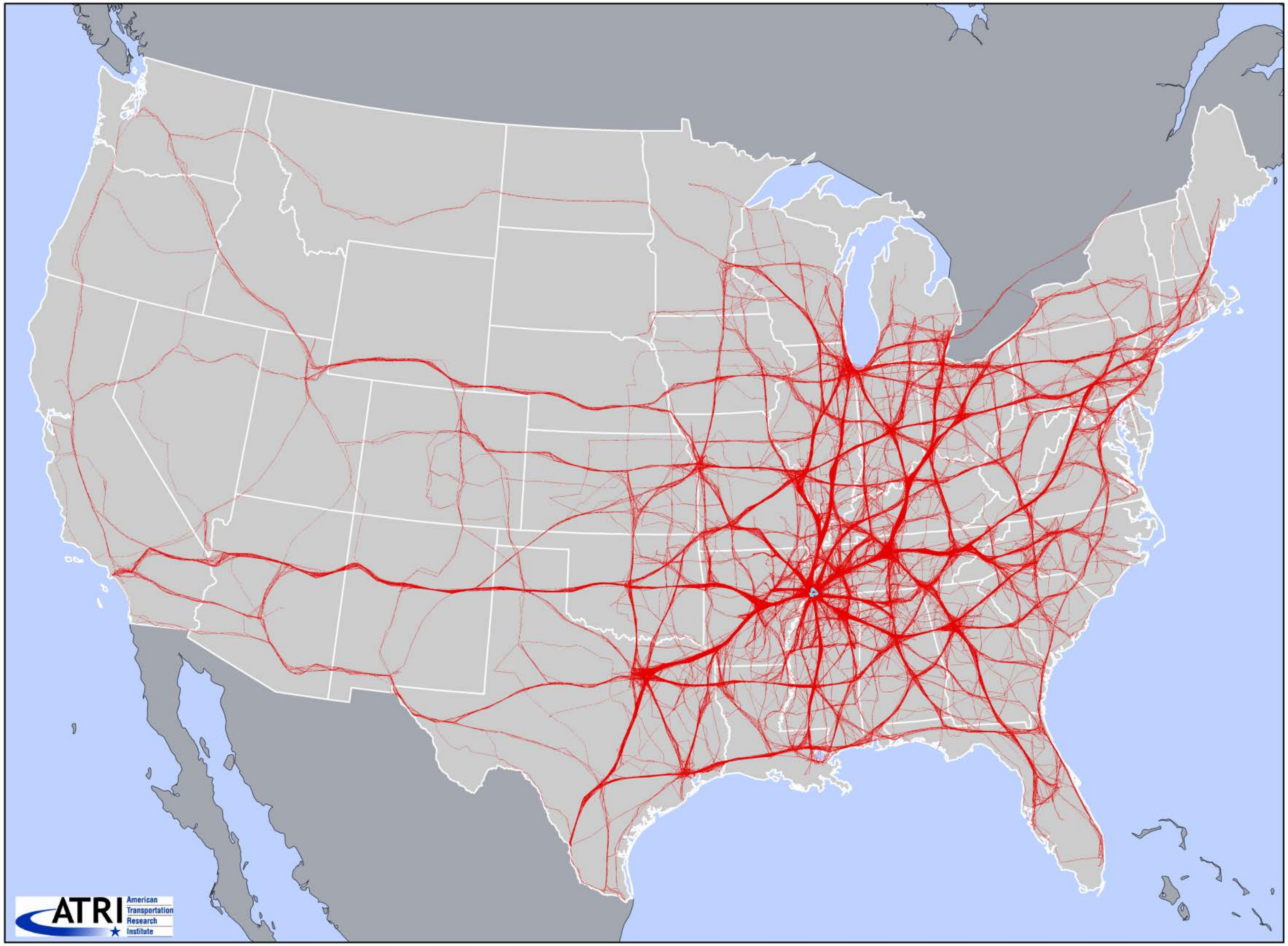


Same 1,000 Trucks After 5 Days



MIP

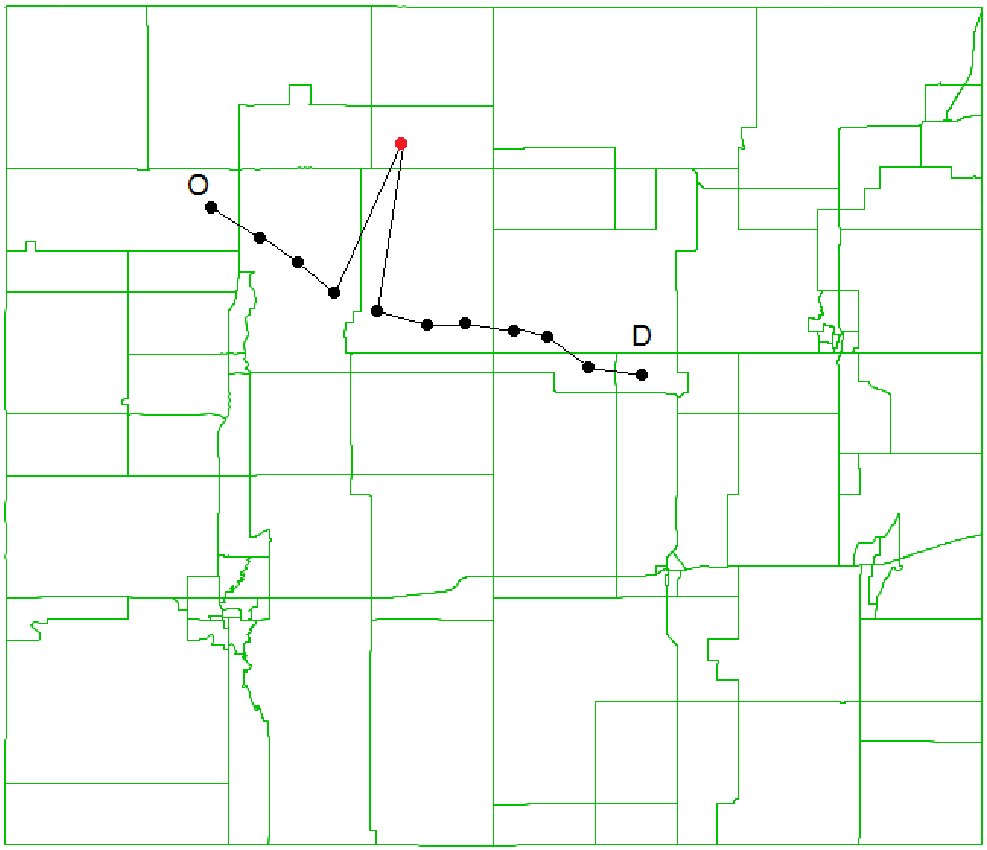
Same 1,000 Trucks After 7 Days



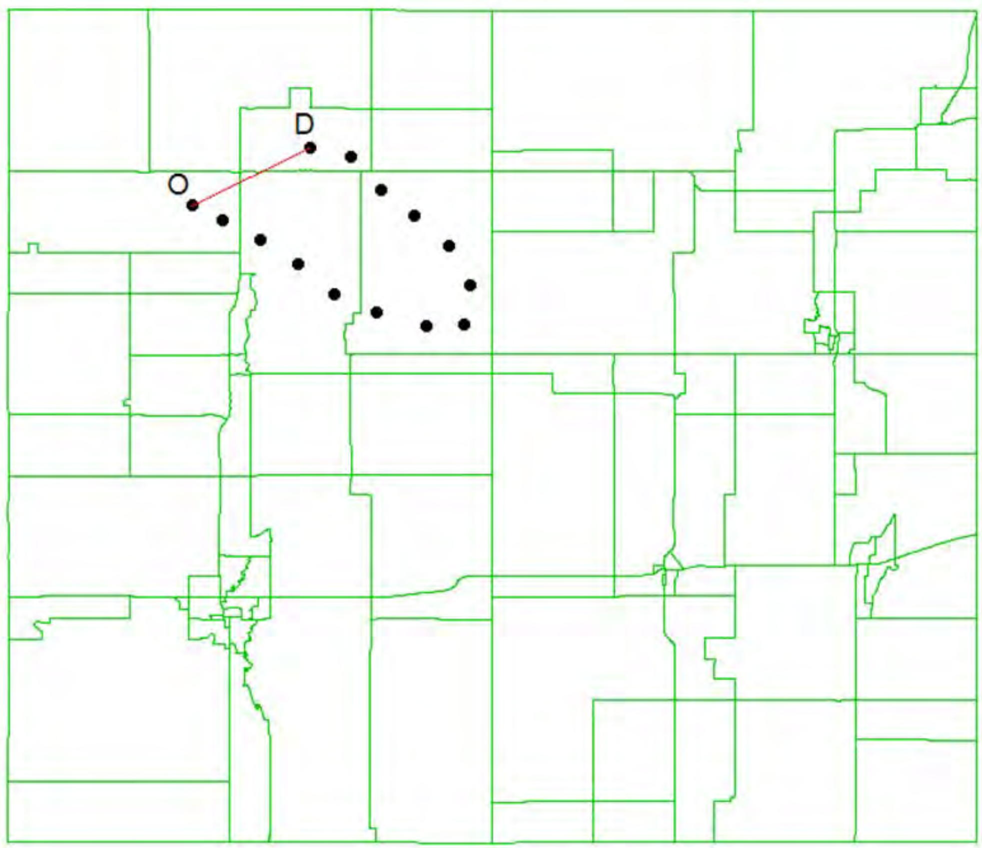
Data Cleaning

- Data Filtering:
  - GPS jumps – urban canyons, mountains, spatial joins, etc.
  - Study period edges – trips in progress
  - Duration & OD mismatch – missed stops, GPS jumps
- Applied conservative filtering methods

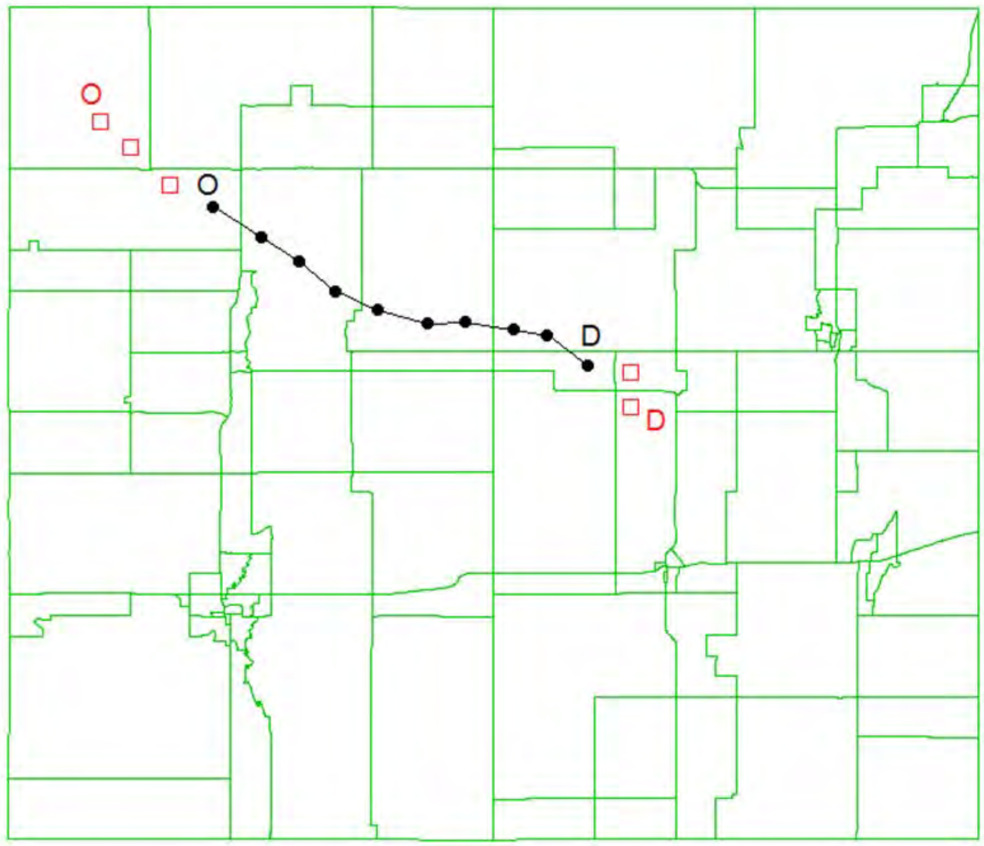
GPS Blips



Circuity



Truncation



Processing Data to Identify Stops

Processing Data to Identify Stops

- Classify trace data records into moving and stopped
- Aggregate moving records into trip records

from TAZ	to TAZ	distance	time	elapsed time	speed	status1	status2
10	101032	66.0	57.7	57.7	68.6	moving	moving
101032	101033	16.3	14.3	72.0	68.6	moving	moving
101033	**101015**	26.8	27.9	99.9	57.5	moving	moving
101015	101015	0.0	5.0	5.0	0.0	stopped	stopped
101015	101015	0.2	2.7	7.7	5.2	stopped	stopped
101015	101015	0.3	9.8	17.5	2.0	stopped	stopped
101015	101015	0.1	0.3	0.3	28.2	moving	*stopped?*
101015	2035	37.1	60.0	60.3	37.1	moving	moving
2035	18099	67.8	65.4	125.7	62.2	moving	moving
18099	27006	5.9	5.4	131.1	65.3	moving	moving
27006	**18023**	10.0	15.9	147.0	37.8	moving	moving
18023	18023	0.0	5.0	5.0	0.0	stopped	stopped

Trip	O	D
1	**10**	**101015**
2	**101015**	**18023**

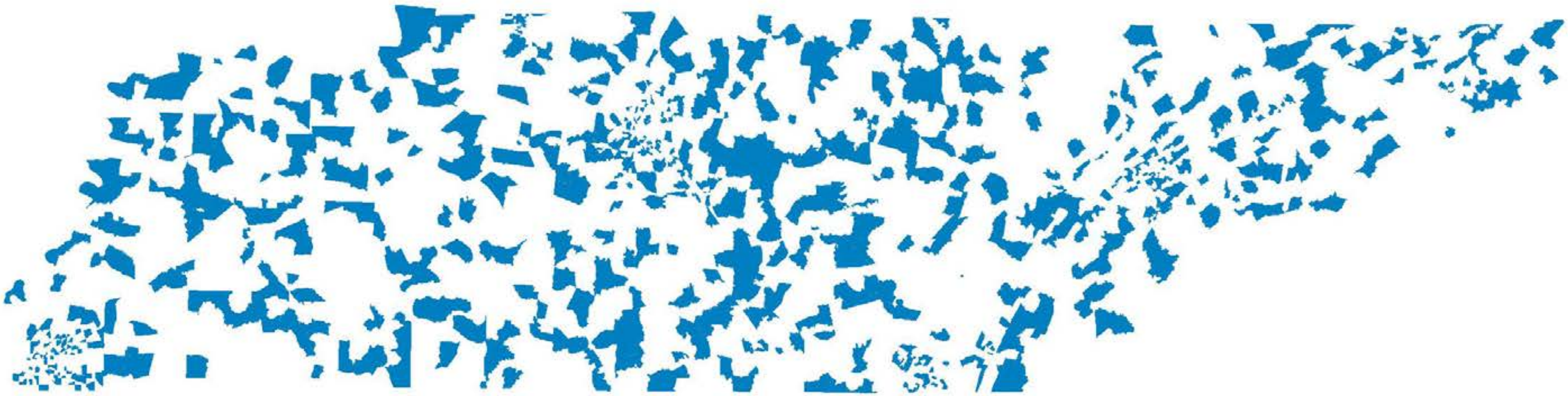
TDOT Cellular vs. Survey Data

- Combined household survey
  - NHTS + 2 MPOs
  - 10,344 households

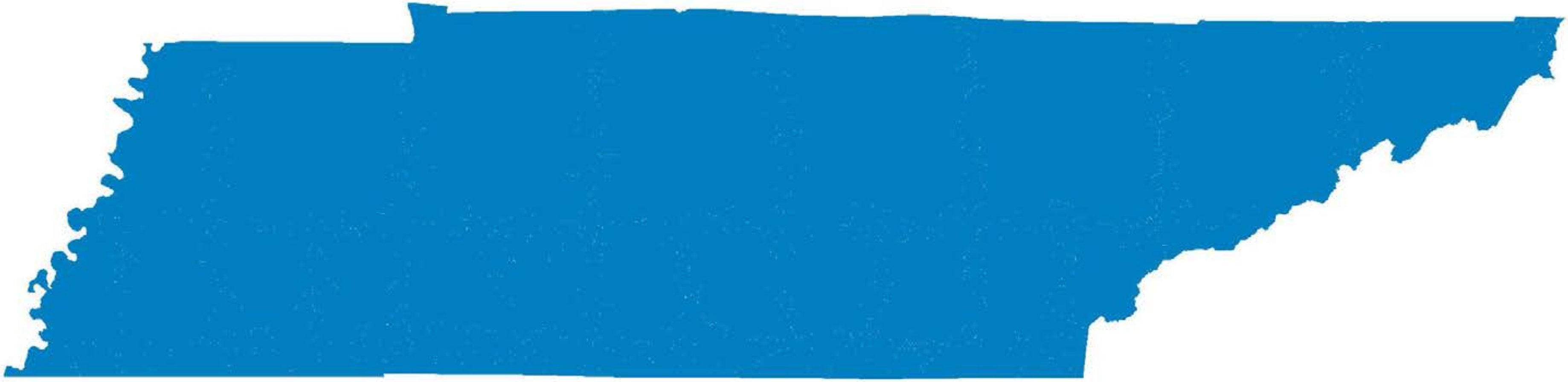
- Trip Table (OD pairs)
  - Total: 12,744,900
  - Survey: 39,782 **0.3%**
  - AirSage: 3,355,539 **26.3%**

Can you recognize the pattern based on a 0.3% sample?

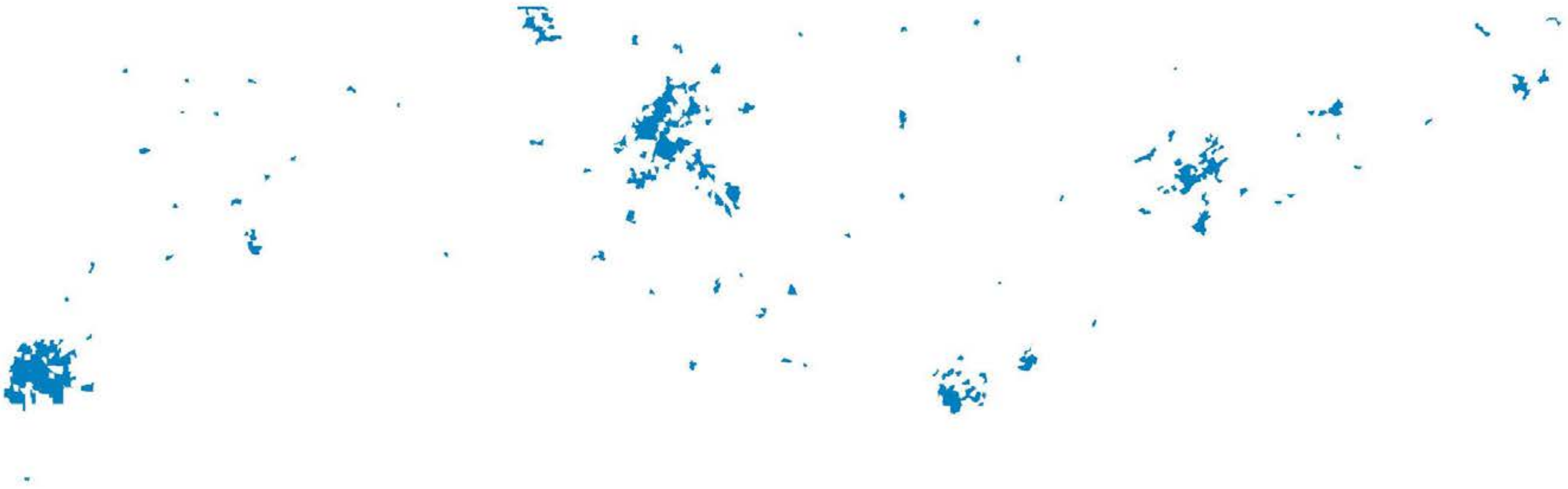
How about a 26.3% sample?



Big Data allows us to see the Big Picture

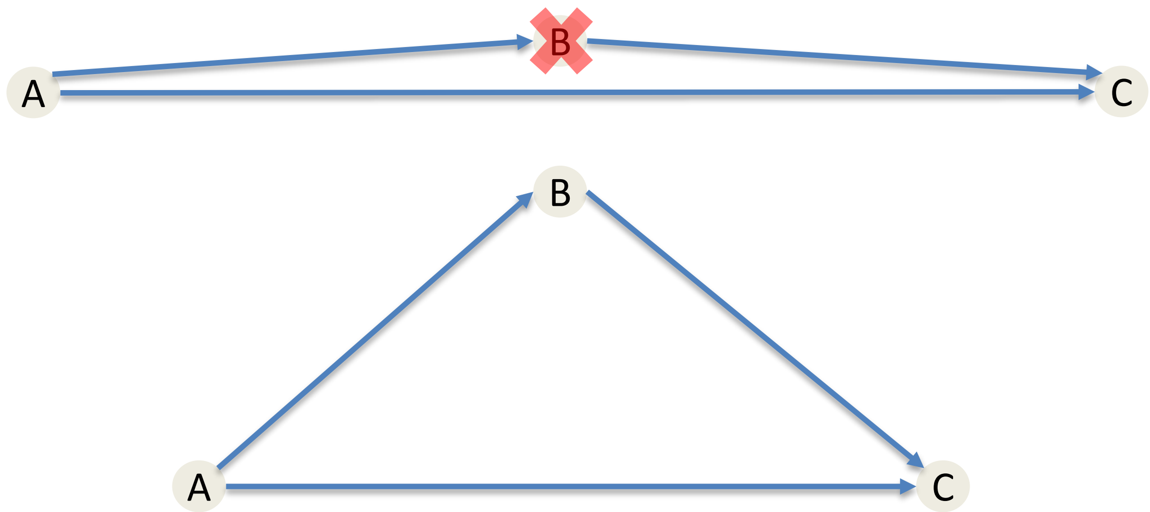


But what about this 26.3%?



Big Data Fusion – AirSage & ATRI

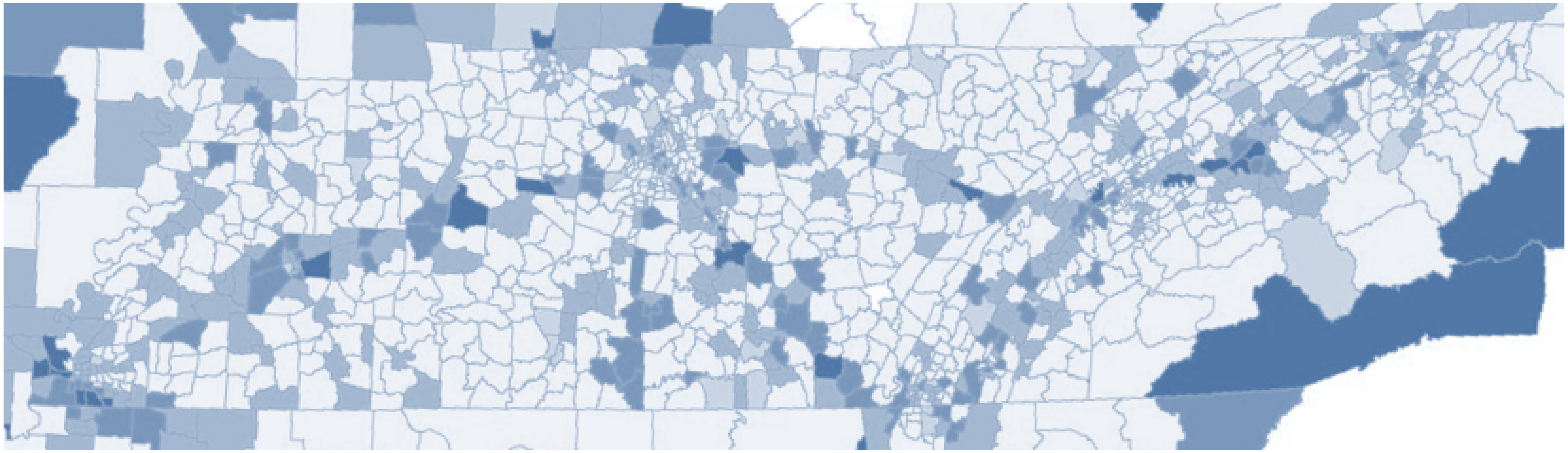
- Re-processed ATRI data to filter out "intermediate" stops on long distance truck trips
  - Meant to make ATRI trips comparable to both AirSage and commodity flows (FAF/Transearch)
  - Used similar but slightly different algorithm than AirSage – compared distances, if $AB + BC \approx AC$ then drop B



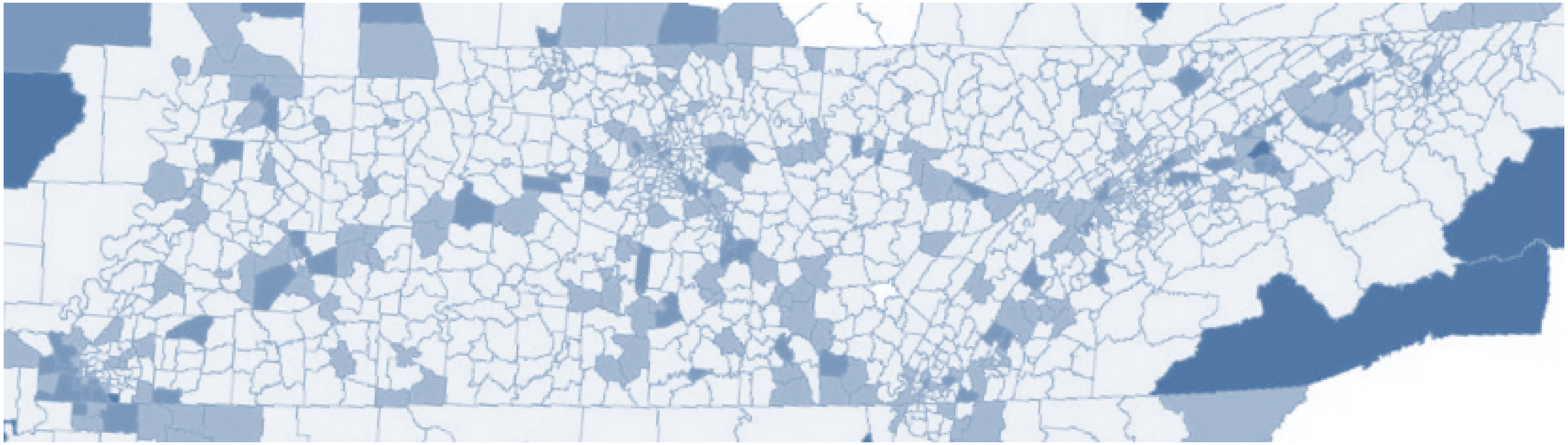
AirSage – ATRI

- Before ATRI filtering, 11% of AirSage cells and 0.20% of AirSage trips showed more truck trips than total trips
- After filtering ATRI, 1.3% of AirSage cells and 0.09% of AirSage trips showed more truck trips than total trips
- Still not perfect, but **filtering ATRI reduced conflicts by 87%**
- Remaining conflicts still indicate remaining issue with intermediate stops, or perhaps coverage drops along Interstates

AirSage – ATRI



AirSage – ATRI



Reconciling Passive OD Data and Traffic Counts

Overview of Expansion Methods

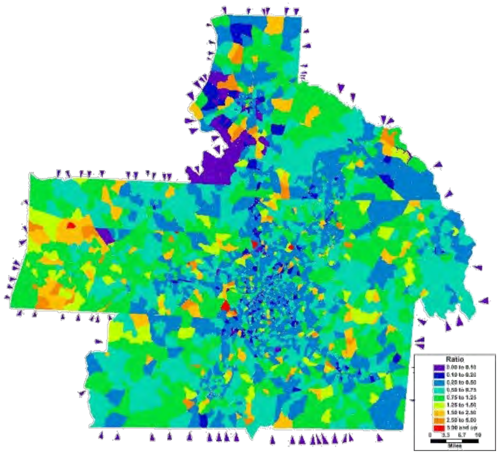
- Aware of 8 methods currently in use, and new methods being actively being researched
- Most robust expansion schemes combine several methods
  - SE data based and simple scaling to counts are among most commonly used and most commonly used alone
  - But these cannot correct for trip/activity duration biases
- Group methods first by type of control data used
  - Then subdivide count-based methods based on single/multiple factors, network based / not, parametric / non-parametric

Taxonomy of Expansion Methods

- SE Data Methods
  - **Market Penetration (Residence-based)**
  - **Trip Generation-Based**
- Traffic Count Methods
  - **Simple Scaling to Counts**
  - Multi-factor Scaling
    - Non-Assignment-Based
      - **Iterative Proportional Fitting to Counts (Frataring)**
      - **Iterative Screenline Fitting / Matrix Partitioning**
    - Network Assignment-Based
      - Nonparametric (ODME)
        - » **Direct ODME**
        - » **Indirect ODME**
      - **Parametric Scaling to Counts**
- *Trace Data Methods*

SE Data Methods

- Market Penetration-based
  - Requires device ID persistence to impute residence location
    - Not currently viable for GPS datasets
  - Compare resident devices per area to population to compute expansion factors by device residence areas
  - Good for addressing demographic biases, not for duration bias
- Trip Generation-based
  - Does not require residence imputation/ID persistence
  - Compares trips to/from zone to estimated trips to/from zone to estimate expansion factor
  - May be better for data validation than data expansion



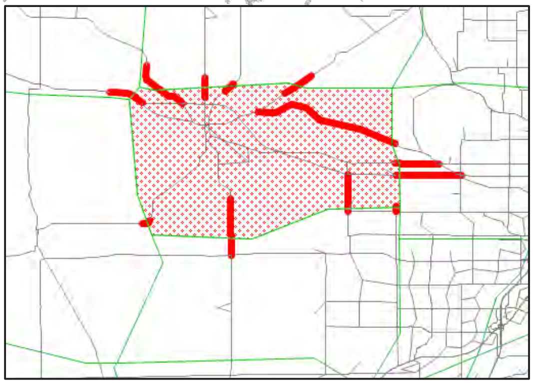
Simple Count-based Methods

- Simple Scaling to Counts
  - Use a single expansion factor to minimize average loading error
    - Usually done via assignment but can be done with map-matching for data with sufficient locational precision (GPS, some LBS)
  - Almost always used as part of / in combination with other more complex count-based methods
  - Sometimes explained in terms of vehicle occupancy but this is only one of several effects that can be captured/reflected
- Iterative Proportional Fitting to Counts (Fratar)
  - Requires counts into/out of zone
  - Commonly used for expanding external stations
  - Also sometimes for airports and other special generator zones

TMIP

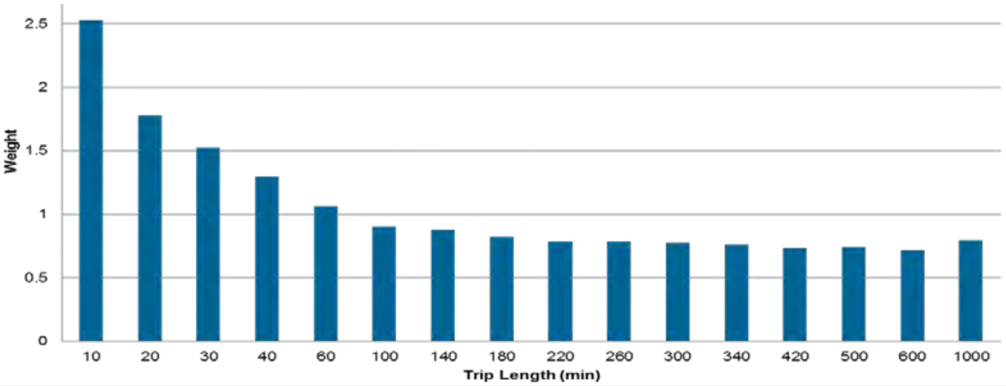
Iterative Screenline Fitting (ISF)

- Loop over screenlines
  - Uses screenlines which partition region into two sets of zones – which partition the OD matrix into quadrants
  - Diagonal quadrants receive factor of 1
  - Off-diagonal quadrants receive factor based on ratio of weighted total counts to aggregated OD trips
    - Weight based on number of screenlines each count is on, etc.
  - Average new factors from this screenline with prior expansion factors



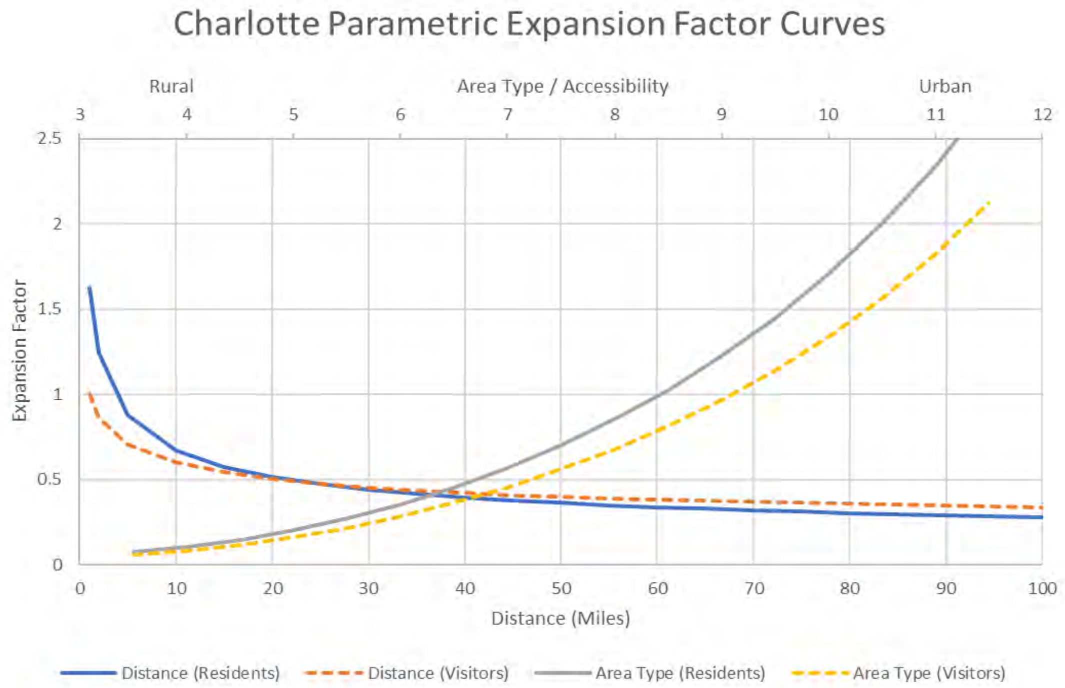
Non-Parametric Assignment-based Methods

- Direct ODME
  - OD/cell-specific expansion factors (lots)
  - Beware of over-fitting to counts!
    - Many different ODME methods, **important** to use one that either minimizes error with respect to both counts and the original ODs or that minimizes error with respect to counts but only within certain constraints (e.g., -50% and +200%) – easier if ODME done after other methods
    - Should measure difference / distance from original to output OD flows (e.g., MAE, MAPE), not just compare TLFDs
  - Relatively easy to do but difficult to interpret / understand
- Indirect ODME
  - Analyze results of ODME to create simpler set of expansion factors based on distance, regions, etc.



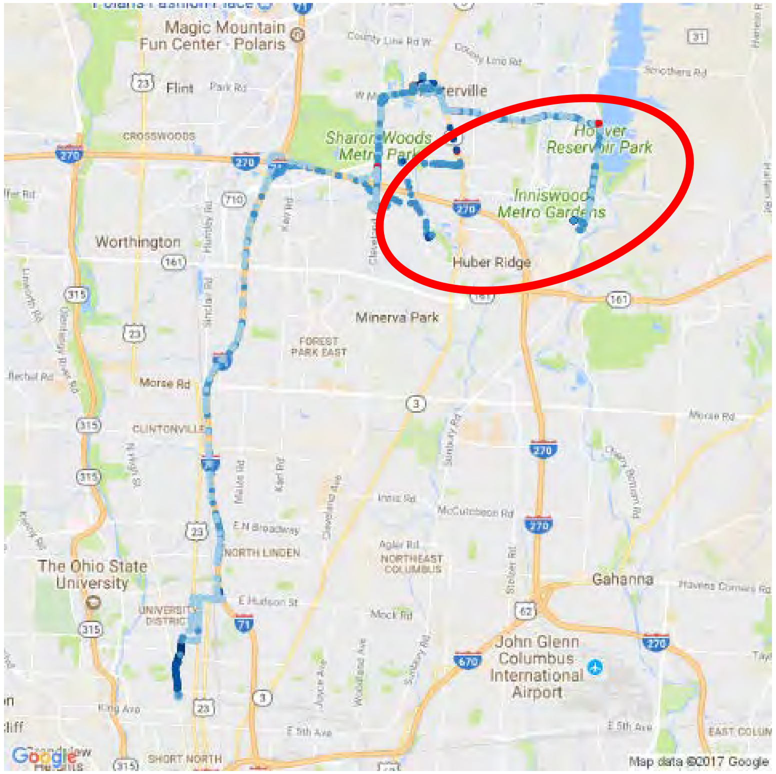
Parametric Scaling to Counts

- Uses assignment within a larger framework to estimate / calibrate parameters for an expansion factor function
- Terms often include
  - Distance
  - Area type or accessibility
  - Intradistrict / Intrazonal
  - Adjacency
- Estimation is NP-Hard
  - Mixed success with genetic algorithm
  - Mixed success with regression on ODME
  - Manual calibration

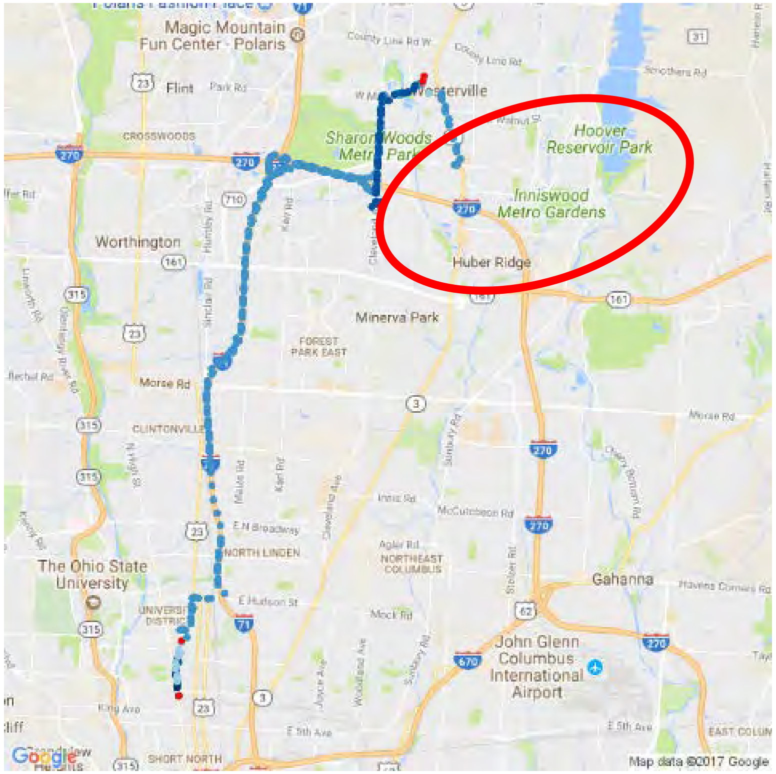


Disaggregate Trace Auditing

- Example of matched traces with short trips in rMove but missing in Cuebiq



rMove



Cuebiq

Comparison of Expansion Methods

#	Method	Fix Trip Length Bias	Fix Coverage Problems	Fix Demographic Bias	Independent of Network	Ease of Application	Holdout Count Sample	Transparency
1	Market Penetatraion-based	✘	✓	✓✓	✓	✓	✓	-
2	Trip-Generation-based	✘	✓	✓✓	✓	-	✓	✓
3	Single-factor Scaling	✘	✘	✘	✓	✓	-	✓
4	Frataring	✘	✓	✘	✓	✓	✘	✓
5	Iterative Screenlines	✓	✓	✓	✓	-	✓	✓
6	Direct ODME	✓	✓	✓	✘	✓	-	✘
7	Indirect ODME	✓	✓	✓	✘	✘	-	-
8	Parametric Scaling	✓✓	✓	✓	✘	✘	-	-
9	*Disaggregate Trace Auditing*	✓✓	✓	✓✓	✓	✘	✓✓	✓

- Ensemble methods best for now
- Count-based expansion necessary for now
- Disaggregate methods hold promise

Tennessee Data Fusion

AirSage Expansion Problem

- AirSage does preliminary expansion based on carrier market share by resident census tract, then analysts scale for "auto occupancy" – actually, vehicle trips/cell trips
- Process has previously worked reasonably well for both urban areas and intercity corridors
- Applying this standard practice to TN statewide data produced significant **urban under-loading and rural/intercity over-loading** (e.g., -10% vs. +15%)

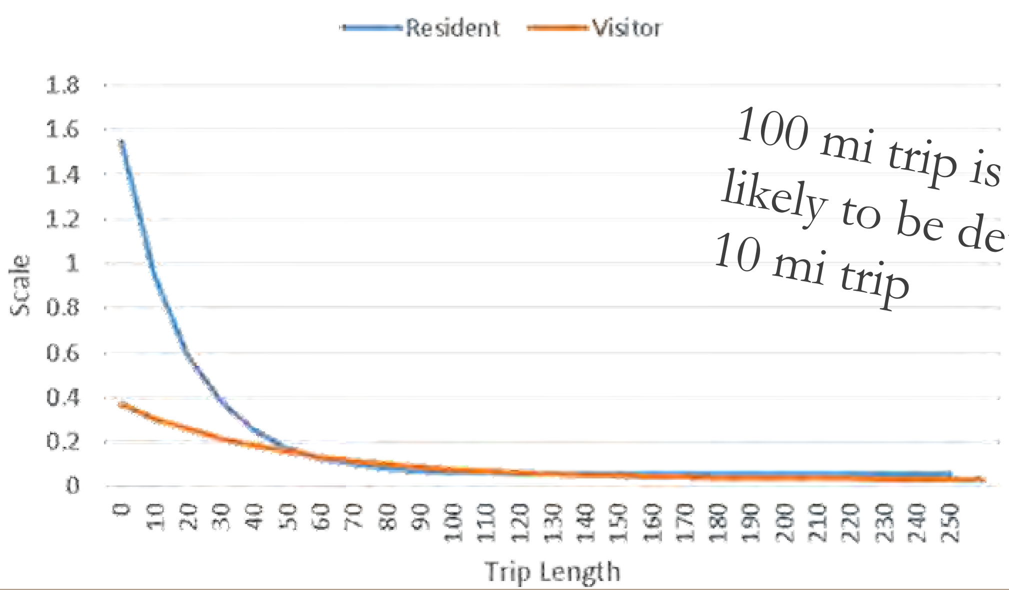
TDOT AirSage Expansion

- 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 Scaling

- *Resident Scale = 0.0612 + 1.6404*Exp(-0.05071*Length)*
- *Visitor Scale = 0.02920 + 0.3376*Exp(-0.01951*Length)*
- Visitors are already long distance travelers – may be more likely to have cell phones / higher auto occupancy

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



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

Non-Parametric Adjustment (ODME)

- **Controls**
  - Minimum factor 0.5
  - Maximum factor 5.0
  - Only 10 iterations
- **Results**
  - RMSE vs. counts from 55.5% to 36.6%
  - Modest additional increase in short trips

Iteration	Versus Traffic Counts			Versus AirSage		Versus ATRI	
	%Error	%RMSE	MAPE	MAE	MAPE	MAE	MAPE
0	-5.48	55.42	81.12	0.00	0.00	0.00	0.00
1	-0.20	46.92	69.53	2.01	0.65	0.01	0.43
2	-0.57	42.64	64.51	2.74	0.92	0.02	0.75
3	-0.92	40.43	61.60	3.20	1.09	0.02	0.93
…	…	…	…	…	…	…	…
10	**-1.90**	**36.11**	**55.74**	**4.47**	**1.54**	**0.02**	**1.41**

Data-Driven Traffic Forecasting and Modeling

How to Move from Base OD Data to Forecasts

- So, how do you use an expanded OD matrix to produce forecasts
  - Pivoting Point Methods
  - Fixed Factor / Constant Rich Methods

Data Driven / Pivot Point Approaches

- **Premise**
  - Know model can't replicate OD patterns, but hope it can predict how they change in response to things like network changes, tolls, and maybe land development
  - Assume things we don't know about - don't change (instead of don't matter)
- **Methods**
  - Additive, Multiplicative, and more sophisticated methods combining the two (8 case, ODOT, NCHRP 255)
- **Practice**
  - Common in Europe and Australia; required in UK
  - Used in some statewide models in US (FL, IN, TN, MI, etc.)
  - Growing practice in transit forecasting ("data driven approach")

TMIP

Fixed Factor / Constant Rich Approach

- **Premise**
  - Same as pivoting
- **Methods**
  - Absorb observed patterns / effects into the model by estimating constants (fixed factors / shadow prices) in the utility functions
  - Constants can be for zones, districts, or interactions between zones or districts
- **Practice**
  - Works for disaggregate (activity-based), not just aggregate models
  - Makes estimation harder
  - Can reduce specification bias in other parameters
  - Can lead to over-specification if careless

Reproducing OD Patterns, not just TLFDs

- Chattanooga 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 1	2	3	4	5	6	7	8	9	10	11	12	Grand Total
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%

Looking Forward

- Improving forecasts has more to do with using better data than more advanced models
  - Big data not solution for everything but its greatest strength addresses our models and survey data's greatest weaknesses
  - The "Volume" of big data allows us to see the big picture of where people are going – not just how far they go
  - The "Velocity" of big data has the potential to allow us to observe how travel behavior changes over the next decade
- But new data should result in new modeling approaches
  - Need to be humble enough to admit limitations of "pure" models and capitalize on the new opportunity