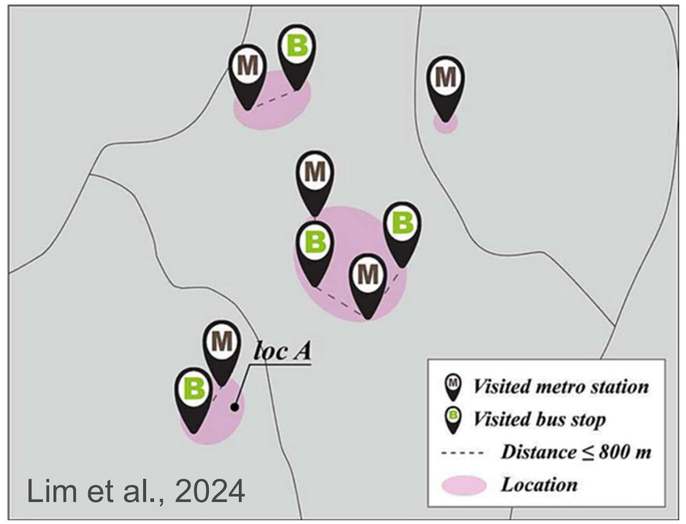
INRIX O-D Data Evaluation and Use for Spatial Downscaling of the NextGen O-D

Use of INRIX O-D data for model validation

RSG
the science of insight®

Background and Problem

- Widespread availability of passive mobility data
- New and innovative applications for this data
- Variations in quality and uniformity of this data
- Variations in understanding of the data
- Agencies with access to INRIX products derived from passive mobility data often want to use it as a validation source



Lim et al., 2024

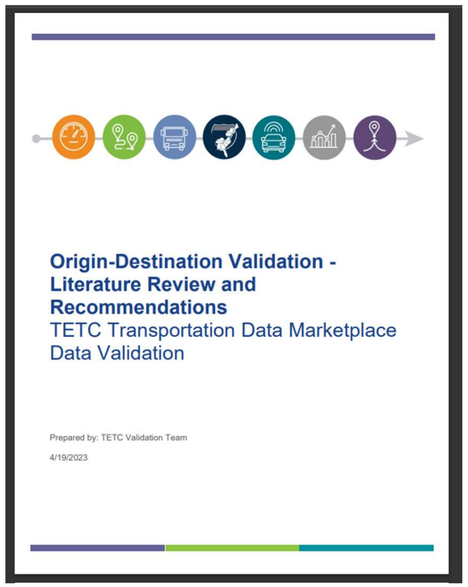
- Since many only have access to an unexpanded sample, a straightforward ample expansion is not feasible
- Phenomena like poor network connectivity and IP address rotation confound the uniformity and accuracy of the sample

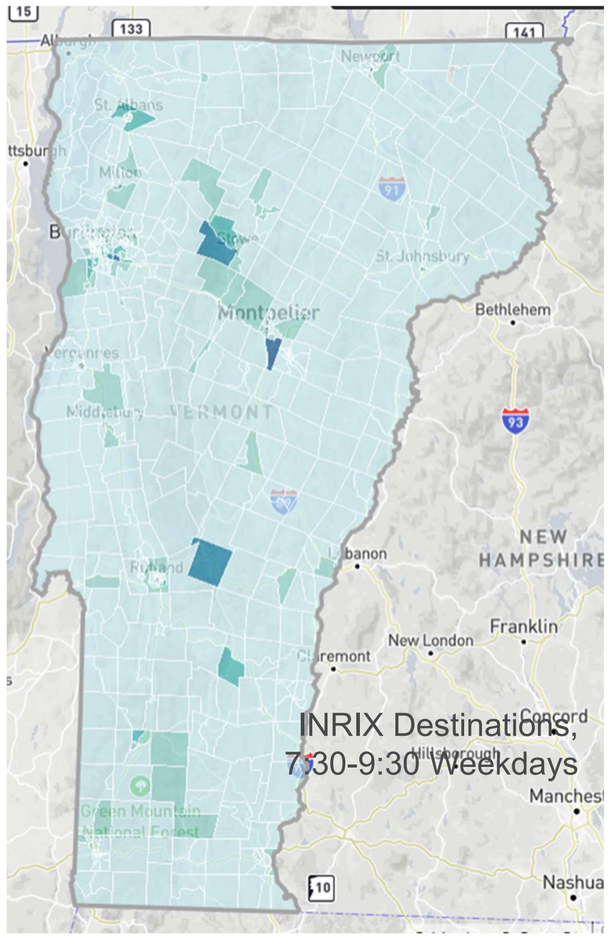
Description of Work

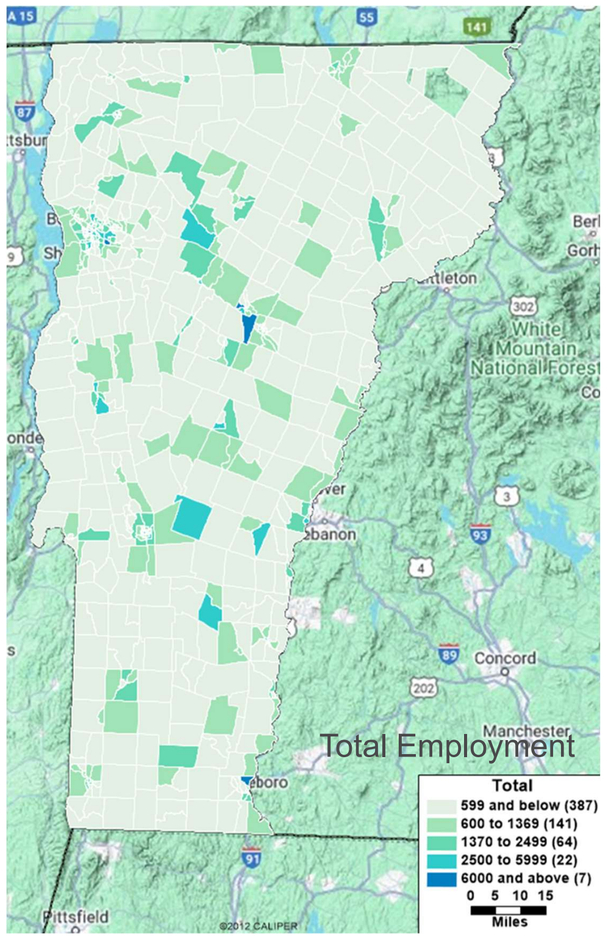
- Evaluate and filter the INRIX O-D sample to improve its accuracy and uniformity
  - *Data quality evaluation* to ensure the accuracy of the INRIX O-D sample
  - *Data comparison tests* to test/filter the uniformity of the sample
- Use it to *spatially downscale* the free National NextGen O-D to get a validation source for Vermont
  - 3 internal zones
  - 29 external zones

Data quality evaluation

- Reasonableness Checks
  - Unreasonably long trips
  - Time-of-day feasibility
  - Road accessibility
  - Matrix symmetry
  - Data coverage
  - Data biases

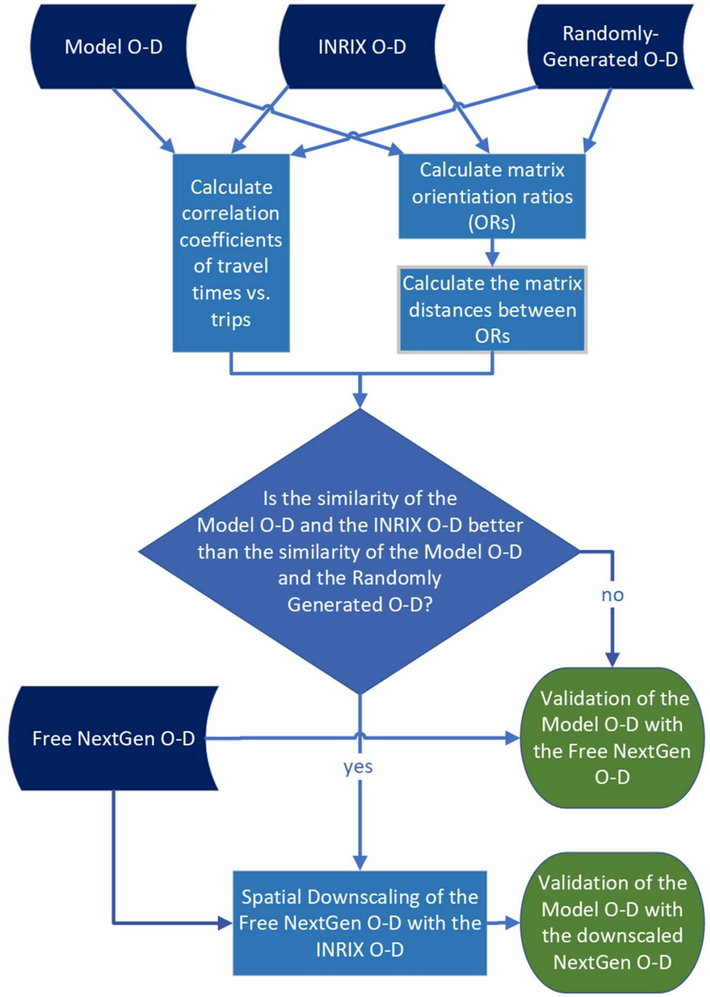






Data comparison tests

- Comparison of matrix structures
- Is the similarity of the Model O-D and the INRIX O-D better than the similarity of the Model O-D and a randomly-generated O-D?



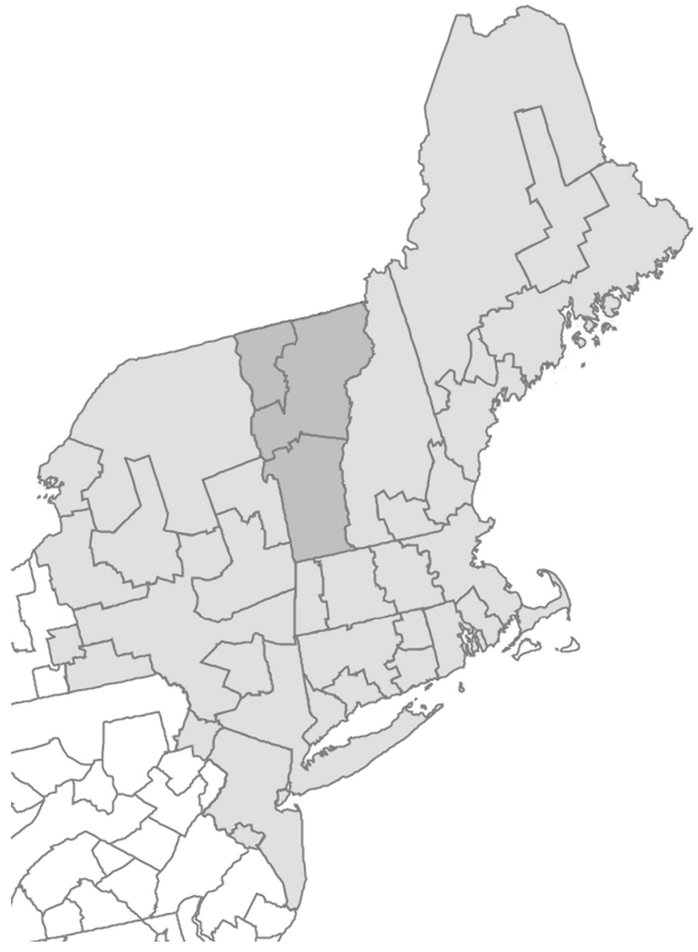
Data comparison tests

- Comparison of matrix structures
- Is the similarity of the Model O-D and the INRIX O-D better than the similarity of the Model O-D and a randomly-generated O-D?

Distance (d) between orientation ratios (OR) with…		A Model O-D	Model O-D
		B INRIX O-D	Random O-D
All cells included	d(A,B)	494,730	6,121
Excluding diagonals	d(A,B)	12,558	5,374
Excluding diagonals and zones with bad (no) mobile service (2022 Mobile Wireless Drive Test)	d(A,B)	3,965	3,633
	d(A,B) for pairs closer than 150 min	3,797	3,554
	d(A,B) for pairs closer than 120 min	3,086	3,522
	d(A,B) for pairs closer than 90 min	2,984	3,496
	d(A,B) for pairs closer than 60 min	2,881	3,465
	d(A,B) for pairs closer than 30 min	2,710	3,438

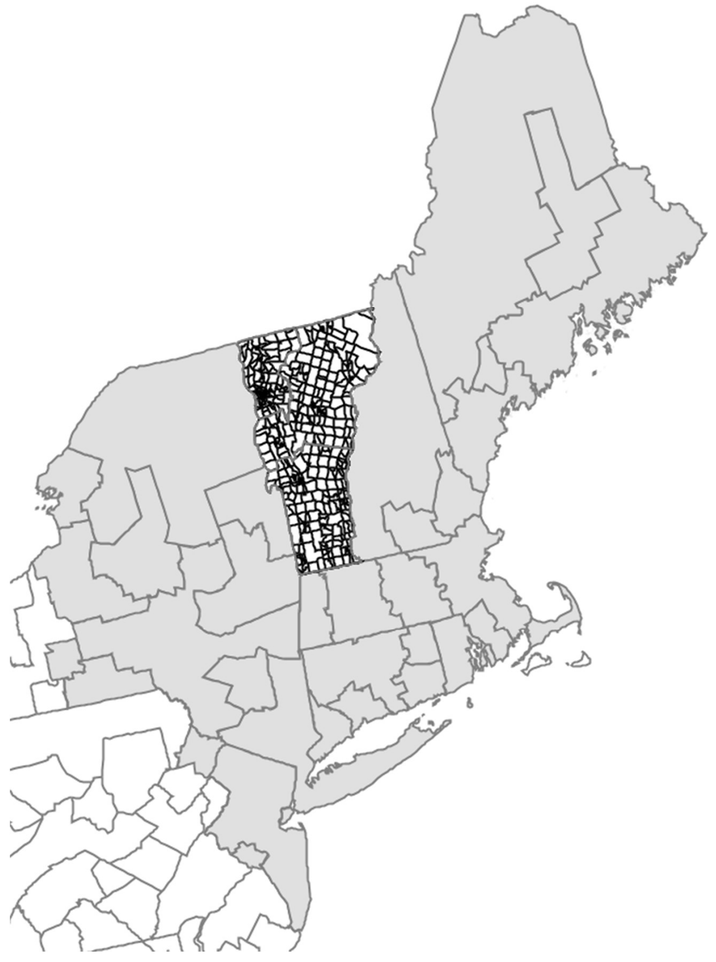
Spatial downscaling

- Calculated fractions of trips for each pair in the INRIX O-D
- Multiplied those fractions by the trips in the 3-zone representation of the NextGen NHTS National O-D
- The result is a version of the NextGen NHTS National O-D downscaled to the Vermont Model zone structure



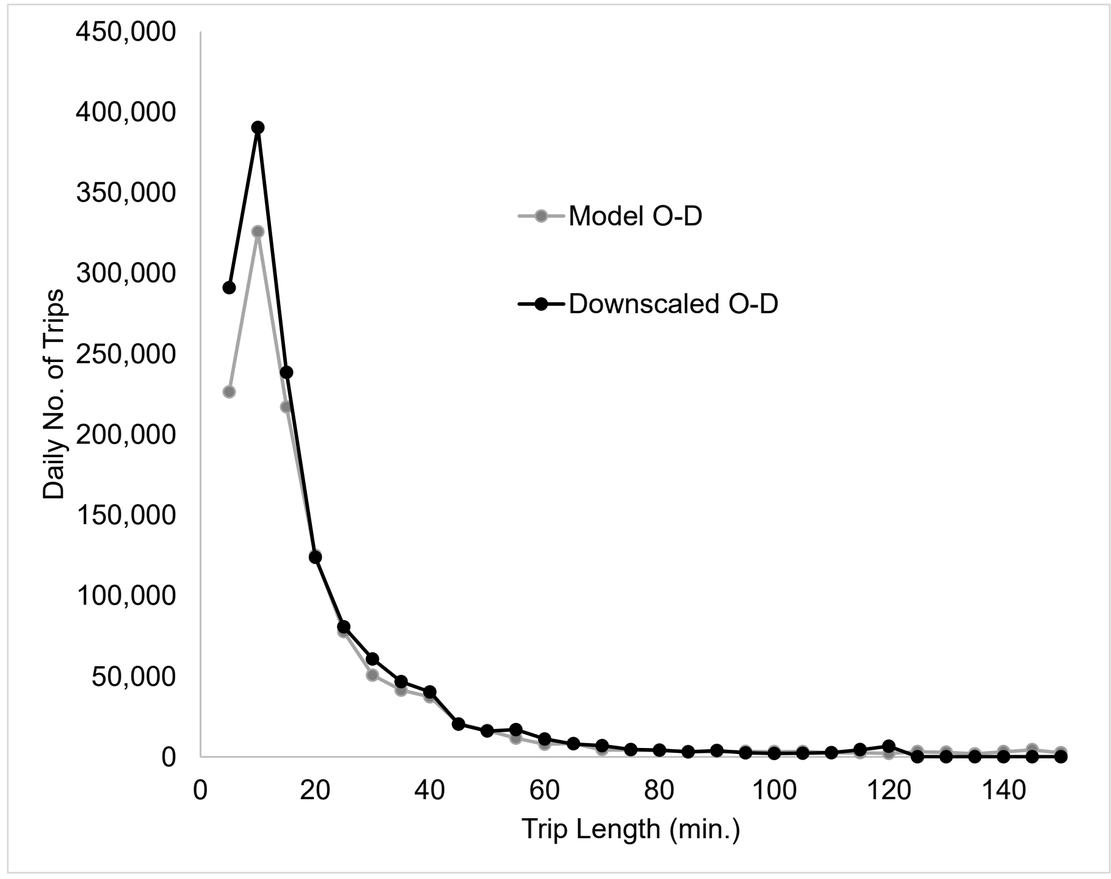
Spatial downscaling

- Calculated fractions of trips for each pair in the INRIX O-D
- Multiplied those fractions by the trips in the 3-zone representation of the NextGen NHTS National O-D
- The result is a version of the NextGen NHTS National O-D downscaled to the Vermont Model zone structure



Validation

- Trip distribution validation
- Noticeable IP address rotation every 30 minutes



Contributions

- Improved the scientific defensibility of the use of passive mobility O-D data:
  - Identified problem areas for passive mobility O-D data
  - Workflow for evaluation, testing, and expansion of passive mobility O-D sample data when typical data needed for expansion is not available
- Helped the client make use of this resource

Data Quality Evaluation

Data Quality Evaluation

Unreasonably long trips

1. Queried the data from the RITIS Trip Analytics tool
2. Created a network
3. Calculated shortest paths
4. Extracted all trips over 450 miles

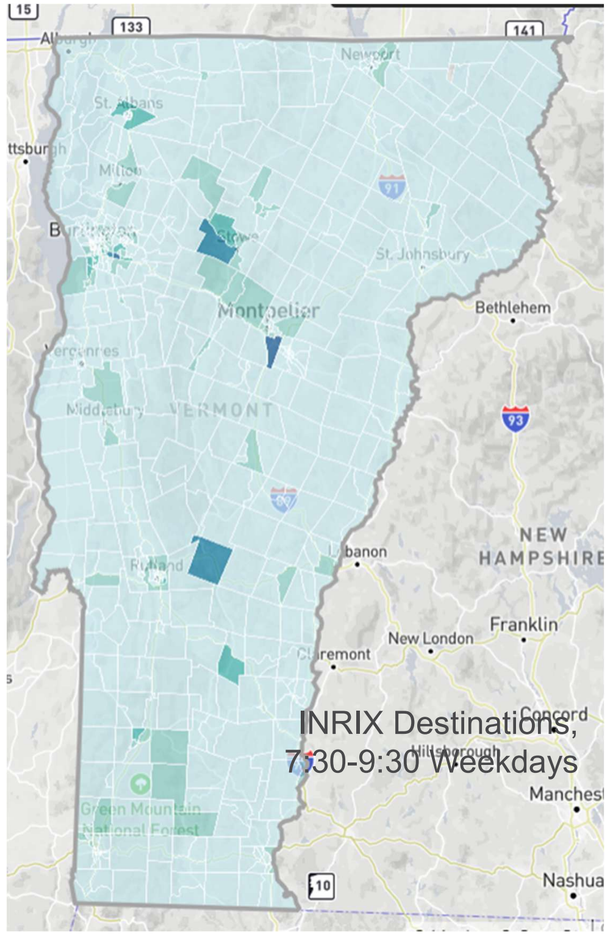
None were found to be unreasonable, or frequent enough to be an error

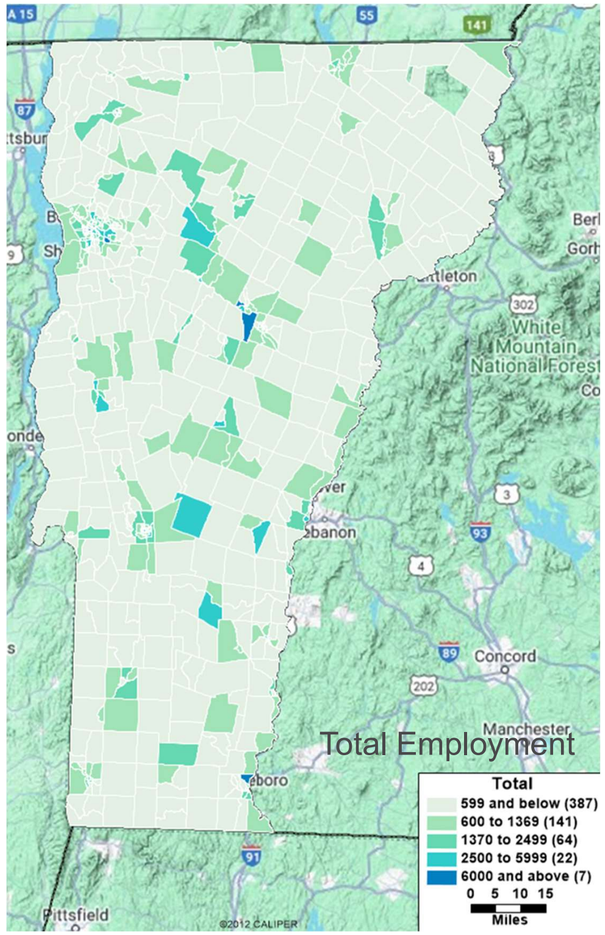
Origin State	Origin County	Destination VT Zone	Distance (mi.)
Virginia	Loudoun County	5000700000136	550
Maryland	Montgomery County	5001300201001	529
Maryland	Montgomery County	5000700000211	523
West Virginia	Berkeley County	5002709659003	490
West Virginia	Berkeley County	5002709656003	485
Maryland	Frederick County	5002309557002	477
Virginia	Shenandoah County	5000309702001	474
Maryland	Anne Arundel County	5002709654001	471
Virginia	Shenandoah County	5002509716002	461
Maryland	Frederick County	5000109605001	459
Pennsylvania	Adams County	5002309553001	458
Maryland	St. Mary's County	5002509715003	455
West Virginia	Berkeley County	5000109609003	451

Data quality evaluation

Time-of-day feasibility

- Morning/evening commuting
  - Compared to relative levels of employment
  - Significant differences for zones with
    - staggered work shifts
    - seasonal commuting fluctuations
    - high levels of work-from-home employment
    - high retail square footage
- Nighttime/overnight
  - Most frequent activity
    - ski resorts
    - large grocery supermarkets





Data Quality Evaluation

Matrix symmetry

- A highly non-symmetric OD matrix may indicate a problem
- Normalized root-mean-square error between each reciprocal pair of cells

$$NRMSE = \frac{1}{n}\sqrt{\frac{\sum(T_{A-B} - T_{B-A})^2}{\sum {T_{A-B}}^2}}$$

  - NRMSE = 4%
  - Maximum Error: 1285%
- These findings suggest that the INRIX O-D is *highly symmetric*, with the exception of certain pairs

Data Quality Evaluation

Data coverage

- Percentage of values in the INRIX O-D that are NOT “empty” (0 or null)
- Percentage of cells in the Model O-D that have more than 1 annual trip
- External zones that are outside the highway-shed of the Model were excluded

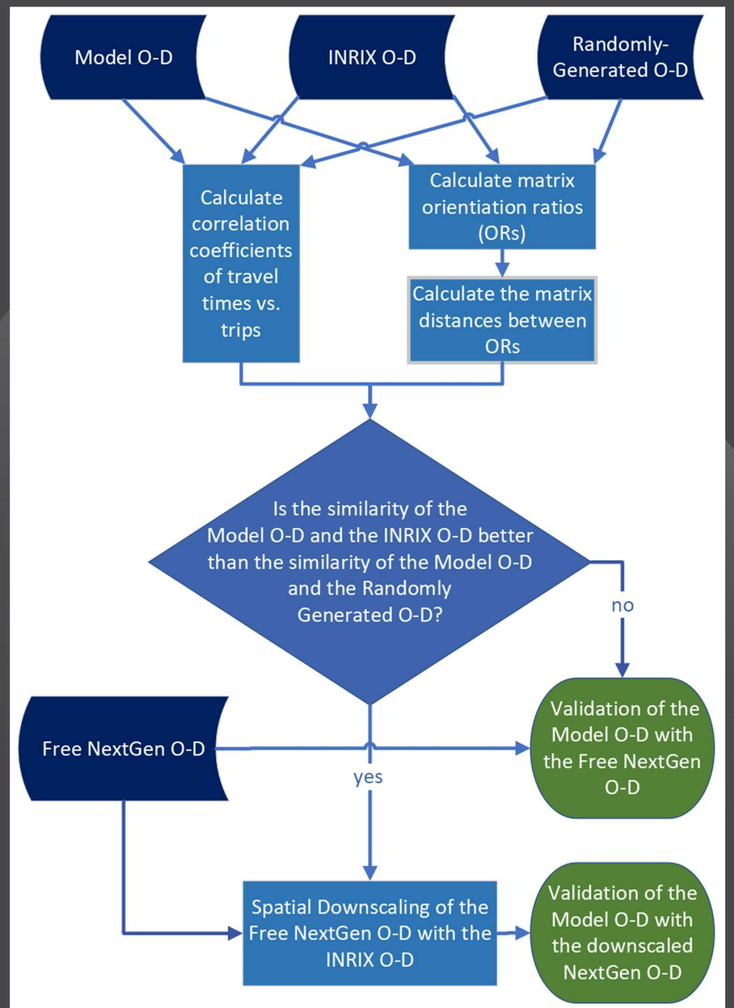
		INRIX O-D	Model O-D
Internal - Internal	No. of “empty” cells	191,125	83,926
	Total no. of cells	389,376	797,449
	Coverage %	51%	89%
Internal - External	No. of "empty" cells	78,807	39,745
	Total no. of cells	107,328	82,156
	Coverage %	27%	52%
External - Internal	No. of "empty" cells	78,774	71,001
	Total no. of cells	102,336	82,156
	Coverage %	23%	14%
Overall	No. of "empty" cells	348,706	194,672
	Total no. of cells	599,040	961,761
	Coverage %	42%	80%

Data Quality Evaluation

Data biases

- Data biases determine the extent to which sample data represents the overall population uniformly throughout the study area.
- Phenomena like poor network connectivity and IP address rotation confound the uniform and accurate identification of trips from passive mobility data
- Total light-duty vehicle (LDV) trip counts can give use an idea of the relative level of sample representation:
  - Sum of LDV trips in the INRIX O-D: 40,386,517
  - Sum of LDV trips in the Model O-D: 522,053,892
  - Sum of LDV trips in the Free National LDV O-D: 516,016,093
- Sample representativeness of around 7.7%.
- Potential biases will be investigated with a series of *Data Comparison Tests*

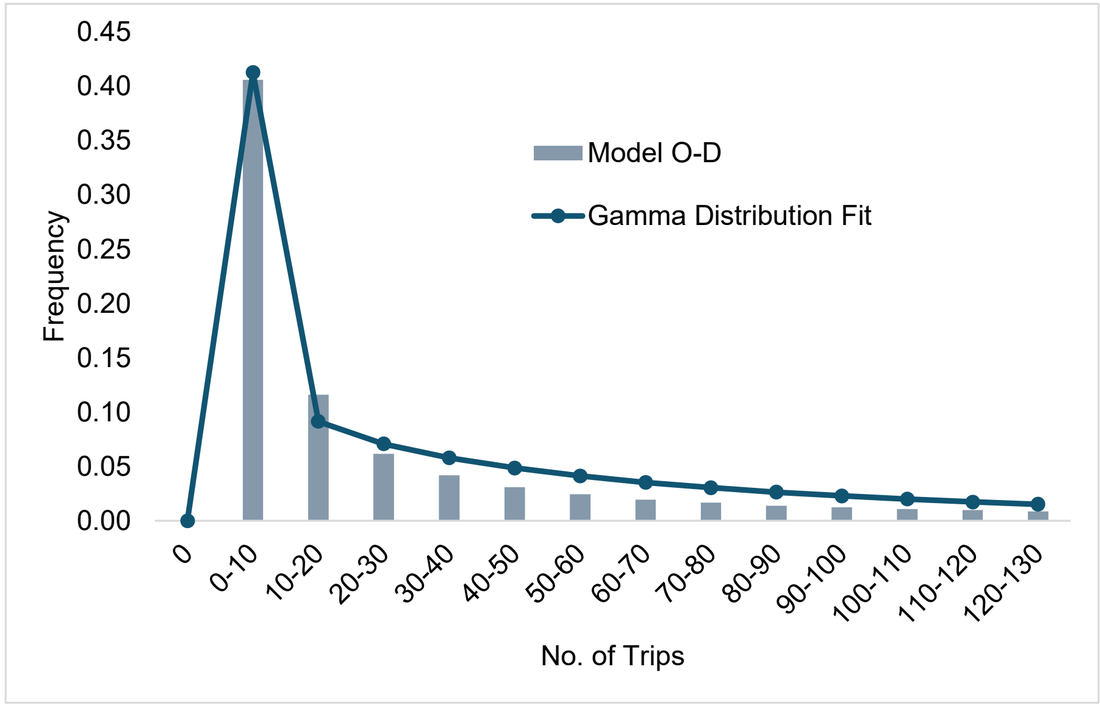
Data Comparison Tests



Data Comparison Tests

Creation of the randomly-generated O-D

- Creation of the randomly-generated matrix
  - Analyzed the frequency distribution of trip lengths
  - Fit to a Gamma probability distribution function
    - α = 0.8
    - β = 85
  - Generated a new set of random values to fit this distribution
  - Made it symmetrical



Data Comparison Tests

Spatial comparison test

- Measured the correlation of O-D travel times and trip counts from each matrix

Correlation between…	…and O-D travel times
Model O-D	-0.17
Random O-D	-0.09
INRIX O-D	-0.11

Data Comparison Tests

Spatial comparison test

- To test for IP address rotation, the correlations were calculated again after clipping the longest trips from the datasets

	…and O-D travel times less than…				
Correlation between…	∞	2 hrs	90 min.	1 hr	30 min.
Model O-D	-0.17	-0.26	-0.30	-0.35	-0.37
Random O-D	-0.09	-0.12	-0.15	-0.23	-0.22
INRIX O-D	-0.11	-0.16	-0.18	-0.21	-0.21

Data Comparison Tests

Spatial comparison test

- To test for IP address rotation, the correlations were calculated again after clipping the longest trips from the datasets

	…and O-D travel times less than…					
Correlation between…	∞	∞ (with diagonals removed)	2 hrs	90 min.	1 hr	30 min.
Model O-D	-0.17	-0.18	-0.26	-0.30	-0.35	-0.37
Random O-D	-0.09	-0.09	-0.12	-0.15	-0.23	-0.22
INRIX O-D	-0.11	-0.16	-0.16	-0.18	-0.21	-0.21

Data Comparison Tests

Comparison of matrix orientation

- Calculate the orientation ratio:

$$OR_{i,j} = \frac{a_{i,j} \Big/ \sum_{i=1}^{n} a_{i,j}}{\sum_{j=1}^{n} a_{i,j} \Big/ \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j}}$$

Standardizes the values and measures the tendency to be weighted toward origins

- Calculate the matrix distance (d) between OR matrices:

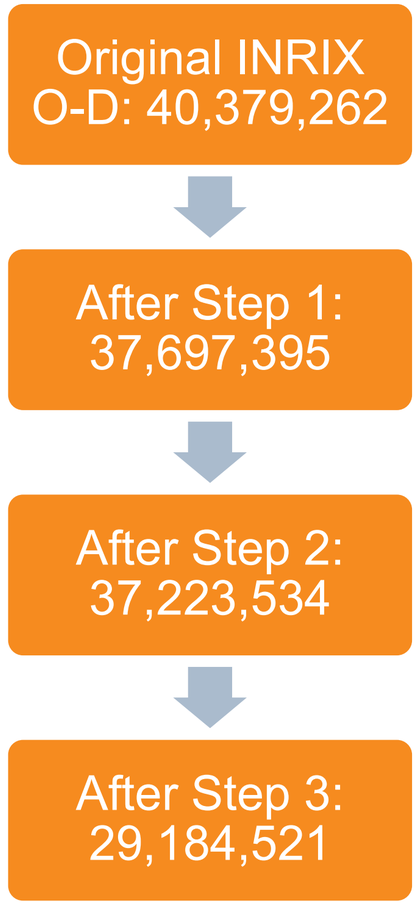
$$d_2(\mathbf{A}, \mathbf{B}) = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} (a_{ij} - b_{ij})^2}$$

Rules for the use of the INRIX O-D for validation

1. Zones with little or no mobile service are excluded

2. O-D pairs with travel times greater than 120 minutes are excluded

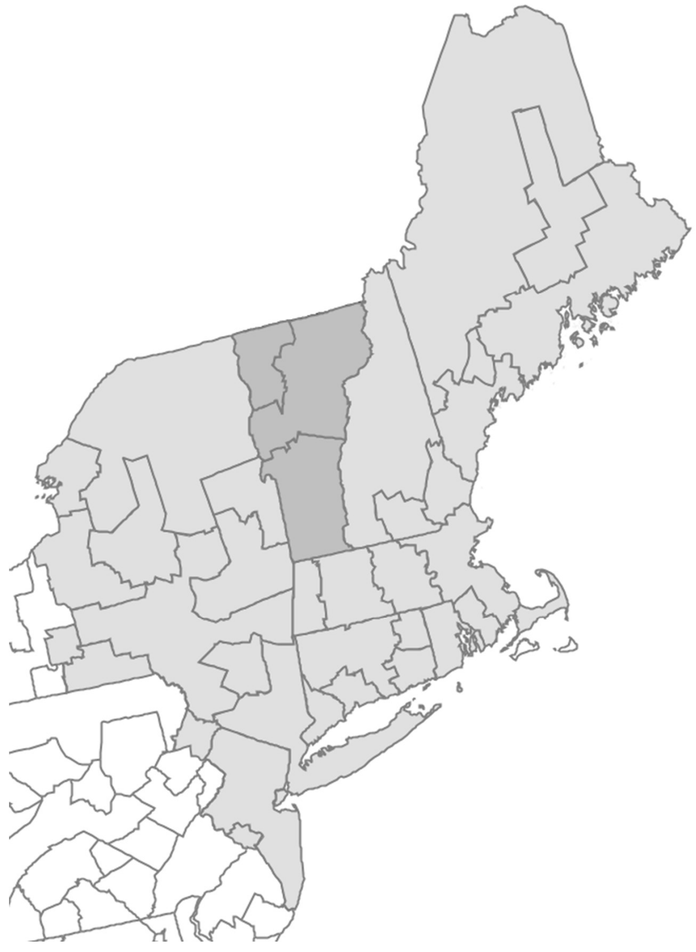
3. Diagonals are excluded



Spatial Downscaling

Spatial downscaling

- Calculated fractions of trips for each pair in the INRIX O-D
- Multiplied those fractions by the trips in the 3-zone representation of the NextGen NHTS National O-D
- The result is a version of the NextGen NHTS National O-D downscaled to the Vermont Model zone structure



Contributions

Improved the scientific defensibility of the use of passive mobility O-D data:

- **>Identified problem areas for passive mobility O-D data**
- **>Workflow for evaluation, testing, and expansion of passive mobility O-D sample data when typical data needed for expansion is not available**

Helped the client make use of this resource

RSG®

Contacts

JAMES SULLIVAN
SENIOR MODELING CONSULTANT

james.sullivan@rsginc.com

+1 802 861 0520

www.rsginc.com

Road accessibility

- Need to locate O-D pairs with a single limited-access route between them
- This has been difficult due to the coarseness of the zone geography

Data Quality Evaluation

Matrix symmetry

- Not clear why this imbalance exists for certain pairs
- Possible explanations
  - Trips are made over several days, or routinely include an overnight stay
  - IP address is rotated in one direction, but not the other

Normalized RMSE	From - To	From – To No. of Trips	To – From No. of Trips
1285%	Milton – Swanton	27	374
900%	Essex Jctn – Northfield	13	130
879%	Milton – Georgia	126	1233
841%	Essex Jctn – Morristown	17	160
771%	Windsor – Brattleboro	14	122
700%	Bradford – Guilford	11	88
700%	S. Burlington – Bethel	12	96
674%	Essex – Barre Town	19	147
669%	Brome-Missisquoi, QC – Georgia	13	100
667%	Burlington ONE –Burlington Oak Ledge & Redstone Parks	18	138
658%	S. Burlington – Northfield	24	182
650%	Guilford – Suffolk County, NY	12	90
650%	Fairlee – Guilford	14	105

Data Quality Evaluation

Matrix symmetry

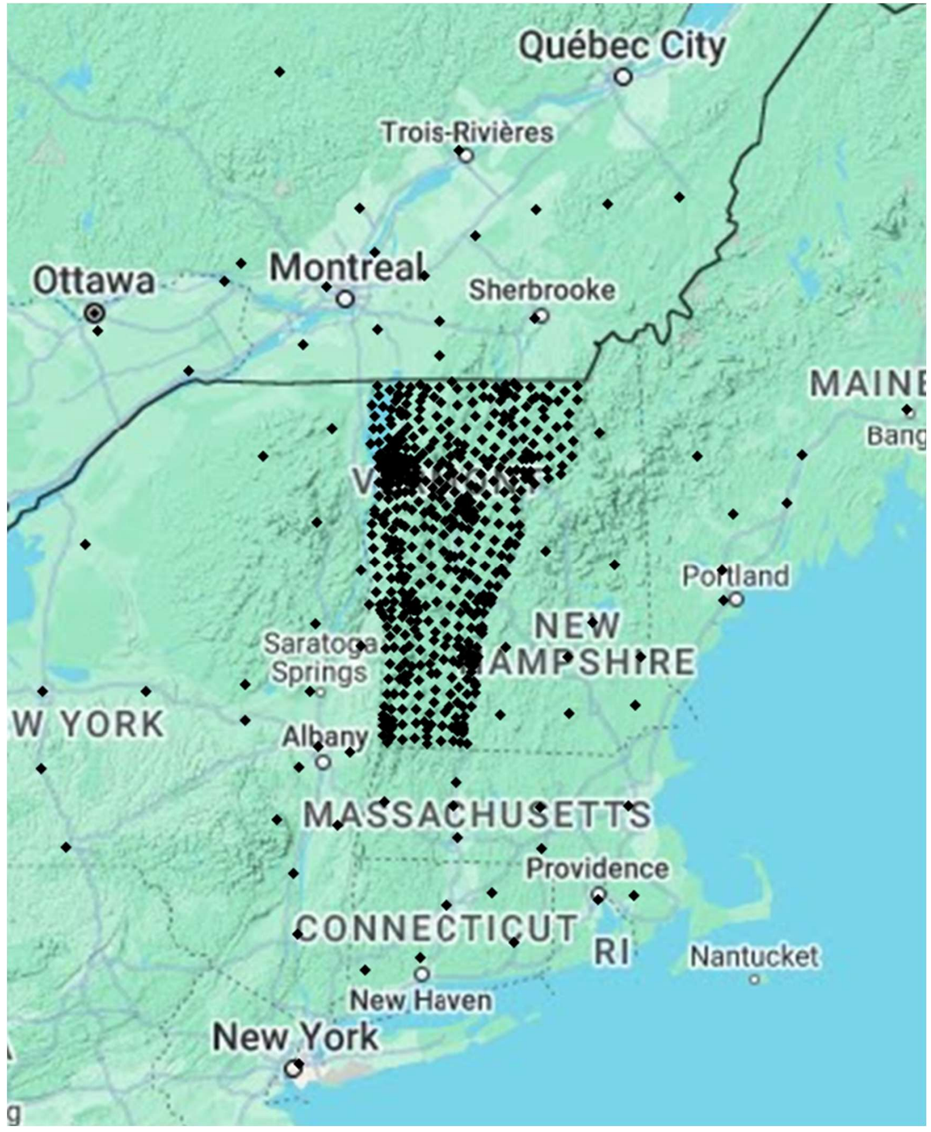
- Further evidence for IP address rotation can be found in the inconsistencies between external–to-internal trips

Origins	INRIX O-D	Model	Free National O-D
Albany--Schenectady, NY	59,708	1,302,058	1,924,314
New York--Newark, NY--NJ--CT	29,785	1,418,130	2,011,432

Data Comparison Tests

Spatial comparison test

- Grand Matrix Geography
  - 942 INRIX zones
  - 985 Model zones
  - Collapsed into 690 “Grand” zones
    - Matched
    - Aggregated to INRIX zone
    - Aggregated to Model zone



References

- Yabe, T., Luca, M., Tsubouchi, K. et al. Enhancing human mobility research with open and standardized datasets. Nat Comput Sci 4, 469–472 (2024).
- Lim, Sungho, Haesung Ahn, Seungchul Shin, Dongmin Lee, Yong Hoon Kim, 2024. Investigating night shift workers' commuting patterns using passive mobility data. Transportation Research Part A: Policy and Practice, Volume 181, March 2024.