Oregon Department of Transportation

ODOT Research

Methods for Monitoring Nonmotorized Transportation - A Proof of Concept in Bend, OR

6/17/2020

Background

Josh’s Background

- Travel, land use, air quality, and GHG modeling
- Traffic count program development
- Crash safety analysis
- Public health analysis

ODOT Role

- Active and Sustainable Transportation Research Coordinator
- Coordinate and conduct research

RESEARCH

Oregon Department of Transportation

Agenda

Agenda
Background/Objectives
Why Count Nonmotorists?
Count Program
Data Fusion Modeling
Next Steps
Discussion & Questions

Research Objectives

Initial Objectives

- Assist Bend MPO in setting up multimodal traffic data collection system
  - Measure project success
  - Plan for the future
  - Prioritize maintenance activities and operations
    - Improve safety analysis
- Measure crash risk for all modes

High Level Objectives

- Develop data collection system with ability to scale easily to other urban areas
- Make it simple and automated as possible
- Provide usable data for high end uses (planning modeling, KPM, health analysis)

Why Count Nonmotorized Traffic?

Invisible Traffic

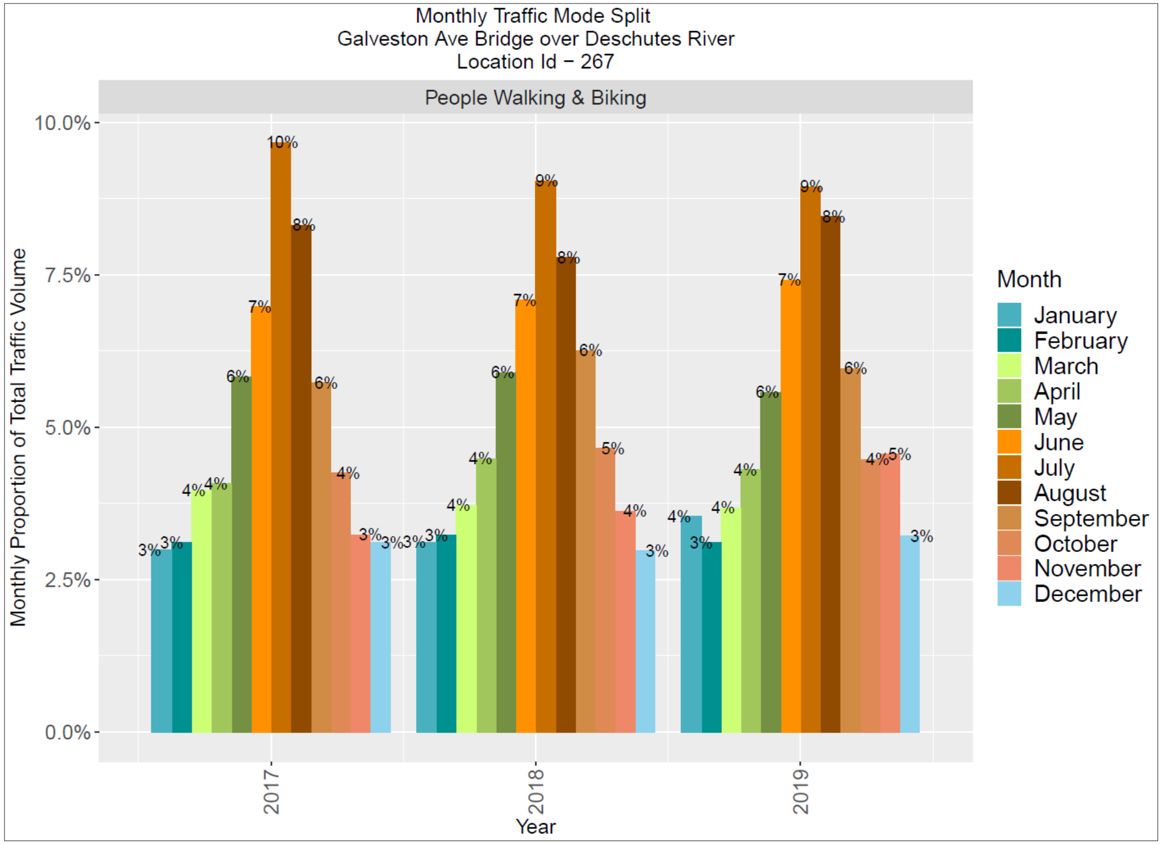
- What's not counted doesn't count
- Short term counts not the whole picture

Highlighting Invisible Traffic

- 405K Vehicle Traffic (July)
- 41K Bike & Pedestrian Traffic

Modal Comparisons

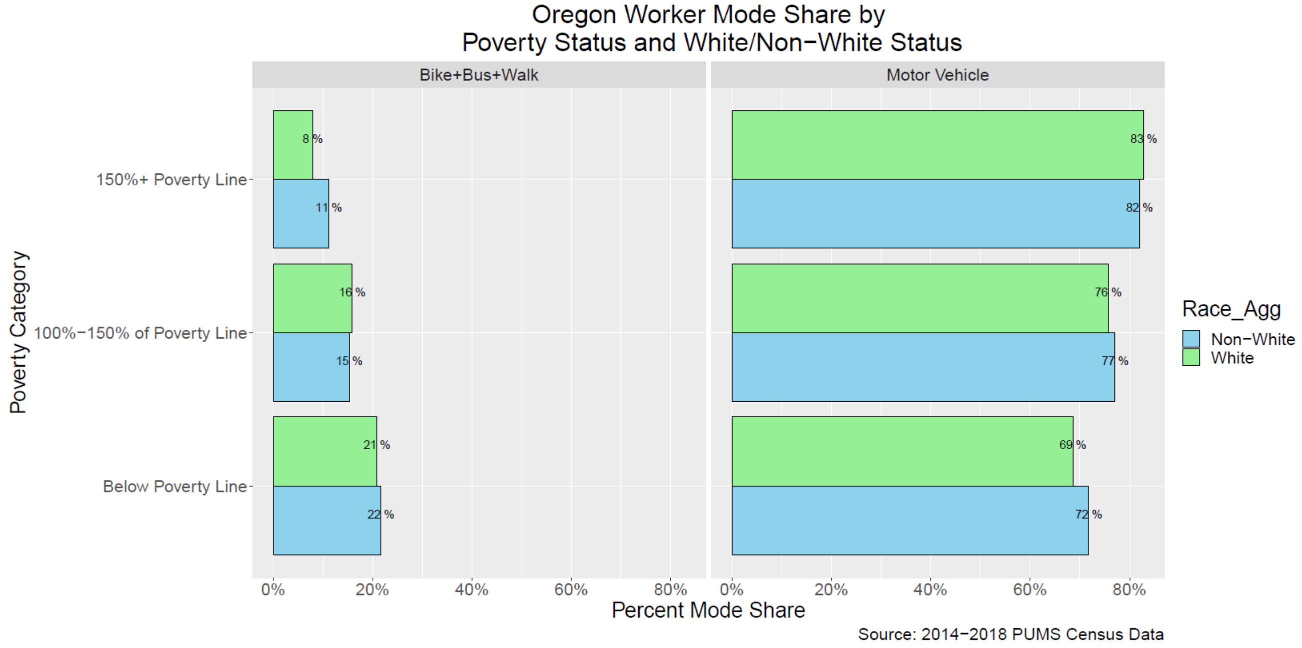
- Segment mode share not a static property



Why Count Nonmotorized Traffic?

Social Equity

- Social justice implications for now accounting for nonmotorized traffic activity



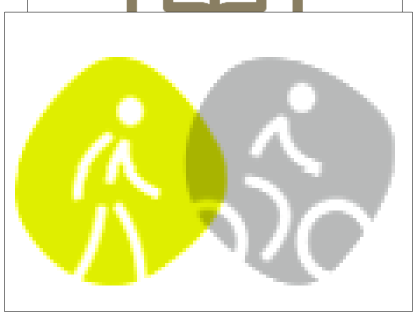
Count Program Overview

Field Staff Data Entry

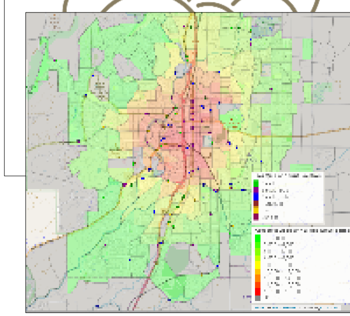
Cloud Based Data Storage

Counts Processor in R (CPiR)

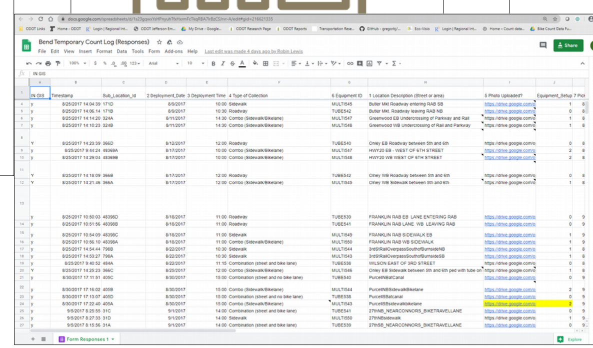
Data Visualization



Eco Counter Counts Database



Count Station Spatial Data



Google Sheet Repository

Deployment Picture

Count Program Overview

Cou
(CPi

Step 1 Raw Data Retrieval

- Pull data
- Assign temporal and spatial info

Step 2 Data Preprocess

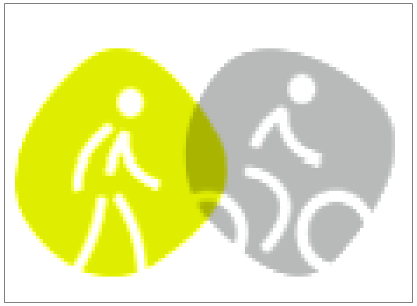
- Assign detailed spatial data
- Process portable sites
- Split user counts

Step 3 QA/QC Check

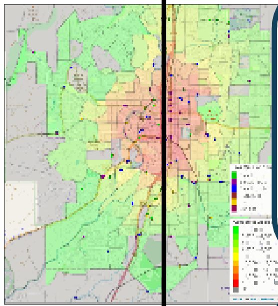
- Assign error flags

Step 4 Annual Estimation

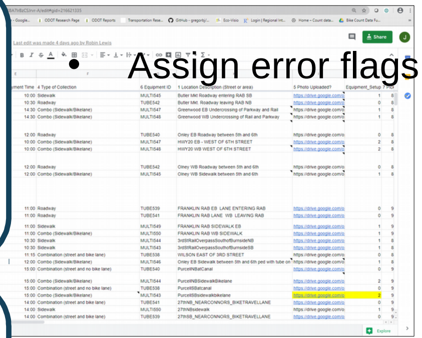
- Daily count imputation
- Annual estimates applying Doy factors



Eco Counter Counts Database



Count Static
Spatial Dat



oogle Sheet
Repository

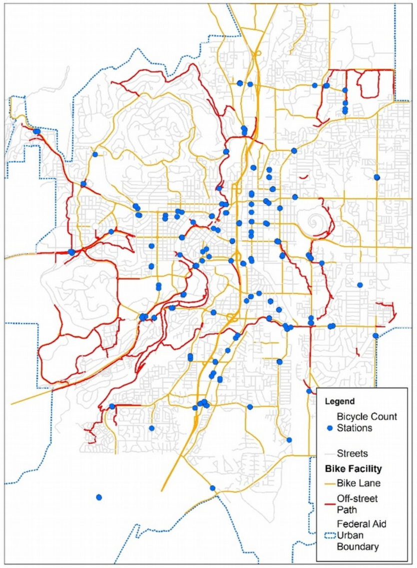
Deployment
Picture

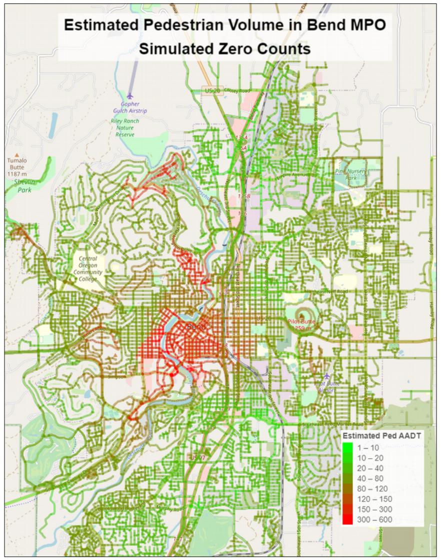
We Have Counts Data….Now What?

Goal
- Estimate activity across the network

Issue: Limited Spatial Resolution
- 56 - 94 sites

Solution: Model traffic
- Use relationships between land use, accessibility and network features and counts
- Parametric vs. machine learning approaches

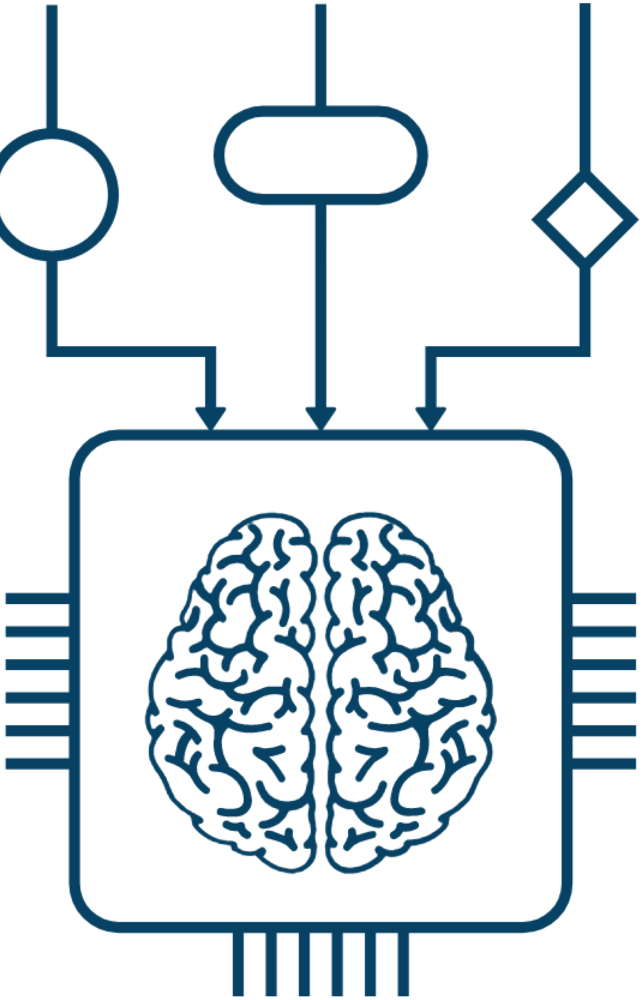




Data Fusion with Machine Learning

What is Machine Learning?

- Algorithms that find and apply patterns in data (MIT Technology Review)
- Many different types for different purposes
- Classification vs. Regression
- Supervised vs. Unsupervised



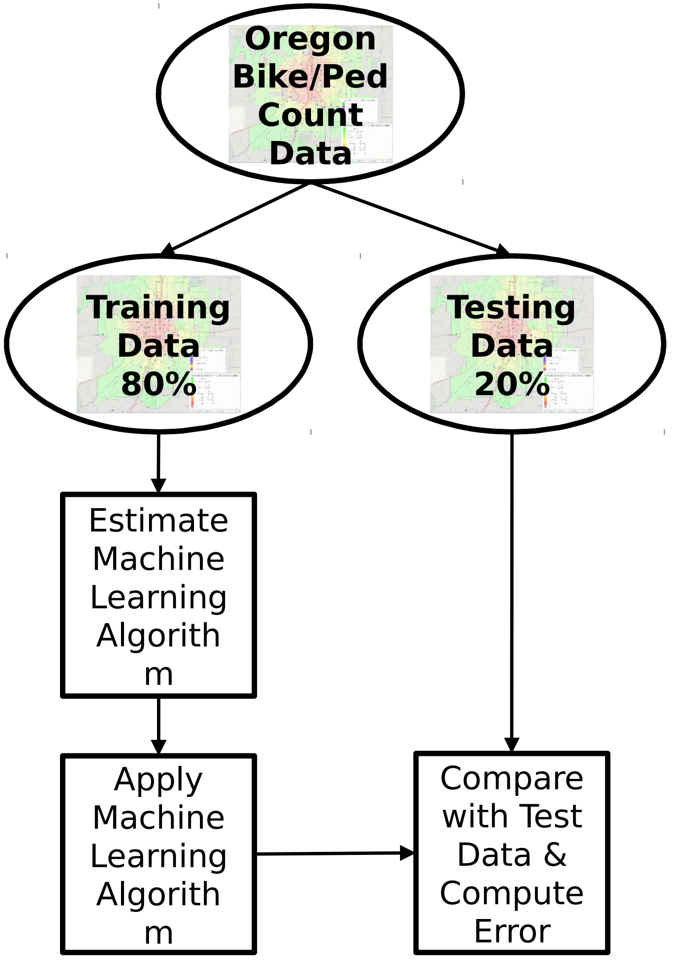
Data Fusion with Machine Learning

Typical Uses
- Marketing, genetic research, physics, social media, and transportation!

Selected Methods
- Negative Binomial Regression
- Decision Tree
- Random Forest

Cross-Validation

Cross Validation
- Divide data into training and testing sets
- Training data for estimating model
- Testing data for determining accuracy of model
- Performed many times to ensure model stability



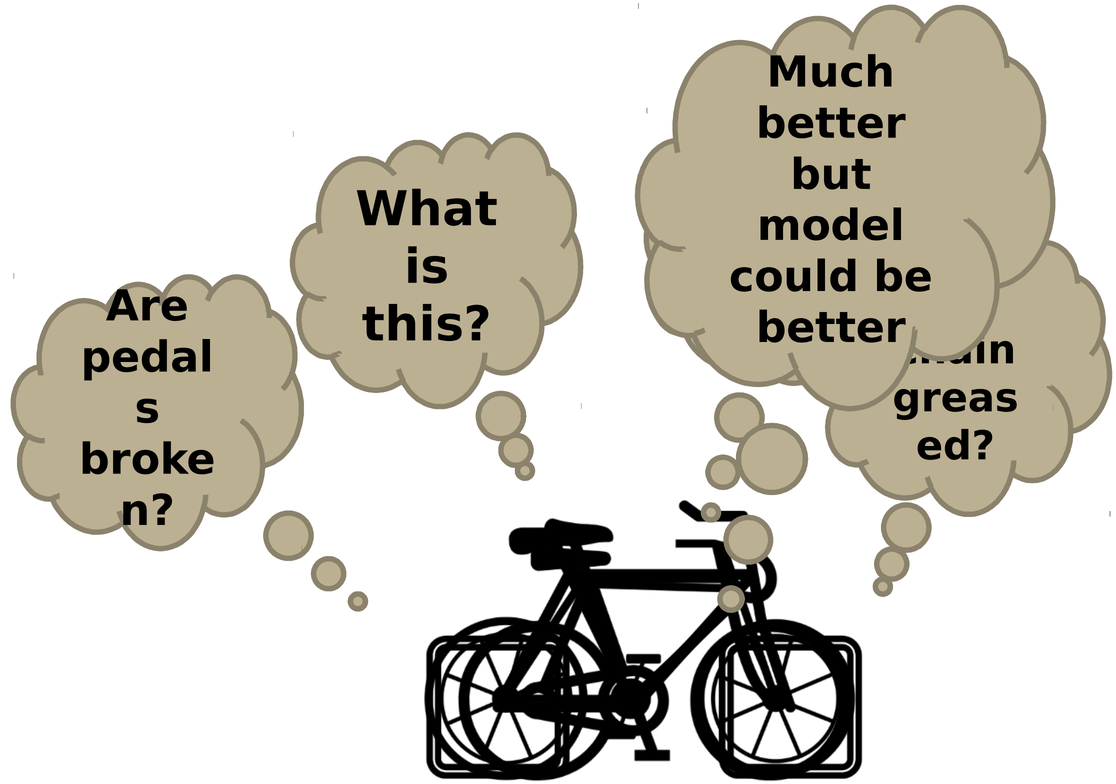
Network Modeling – Data Fusion

What is a model?

- Representation of a thing or phenomenon useful for understanding and decision making
- Performance of a model depends on uses and decisions being made
- “All models are wrong, some are useful”
- Data driven models allow us to put our assumptions on the table

Travel models poor tools for nonmotorized transportation

- Travel surveys collect limited information on nonmotorized
- Assignment procedures make oversimplified assumptions
- No bike/ped counts to calibrate to anyway
- TDMs been a little tyrannical



Network Modeling – Data Fusion

Objective

- Activity estimates for entire network

Uses

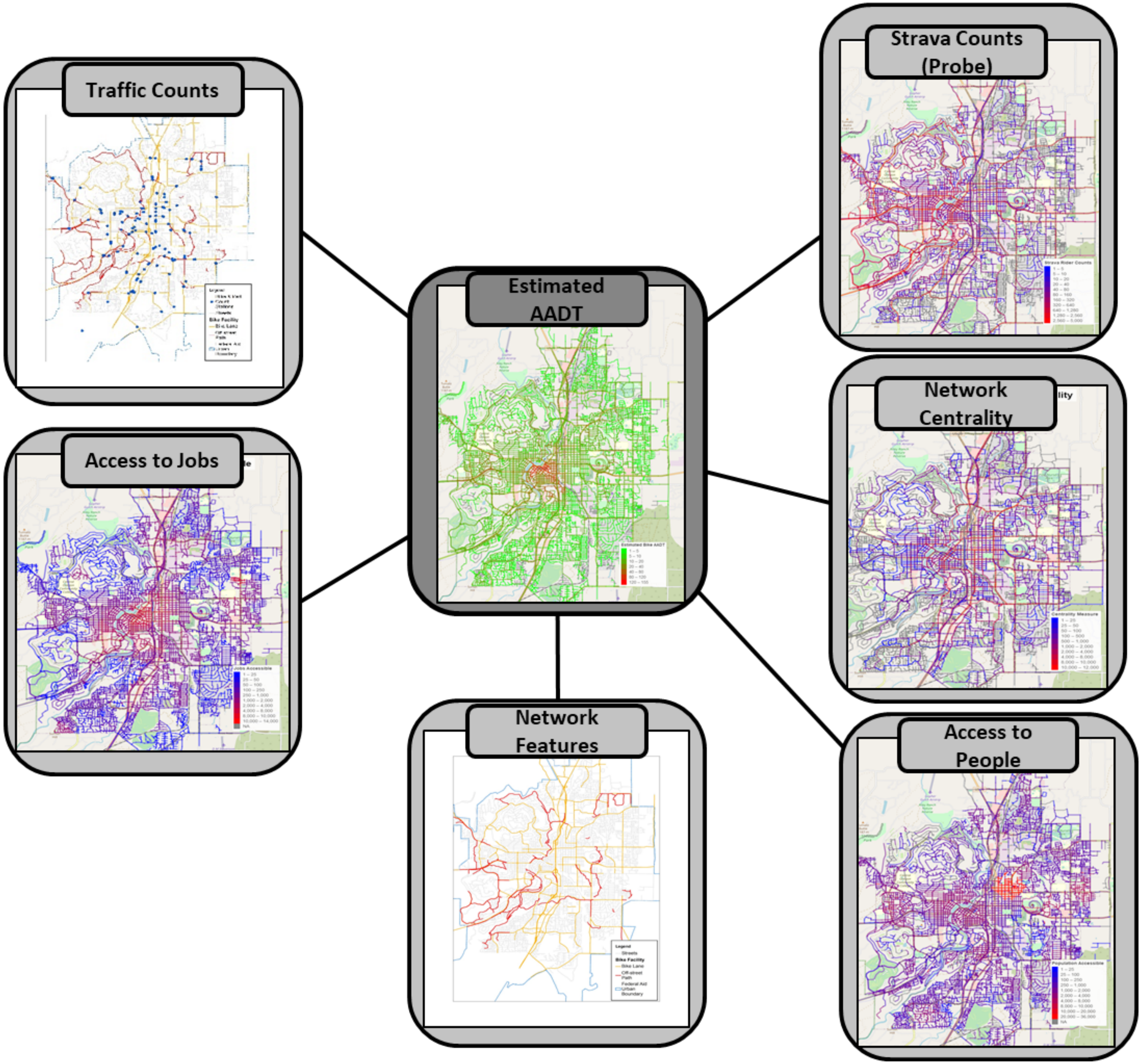
- Planning, monitoring, crash analysis

Methods

- Merges data from multiple features and apply machine learning or statistical model

Output

- Quantifying total network activity
- Crash analysis input
- Health analysis input



Network Modeling – Data Fusion

User Types

- Vehicle
- Bicycle
- Pedestrian

Data

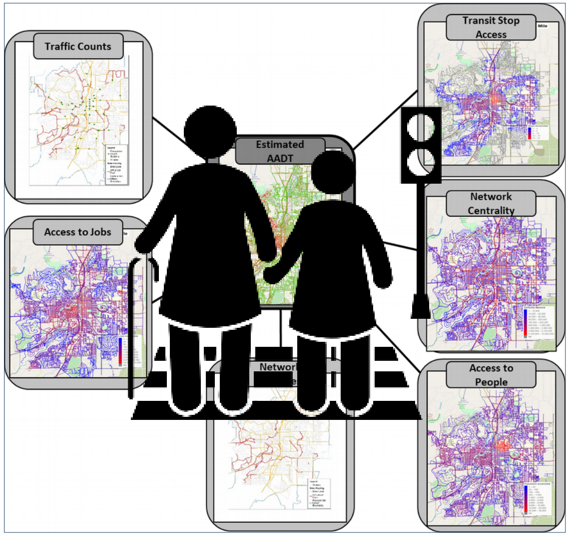
- Network characteristics
- Accessibility
- Centrality
- Probes

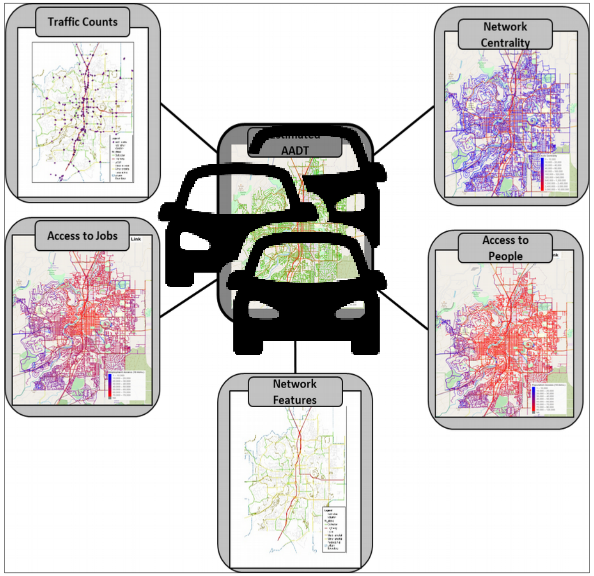
Methods

- Random forest and XgBoost

Output

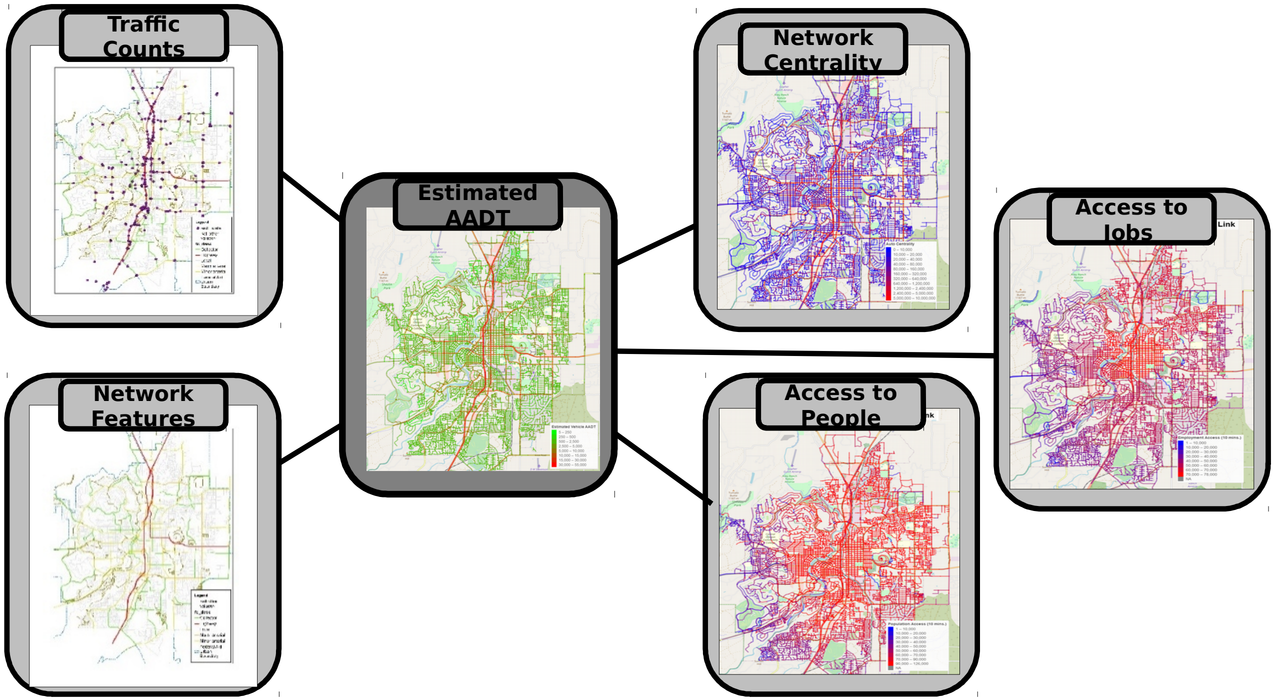
- Quantifying total network activity
- Crash analysis input
- Health analysis input





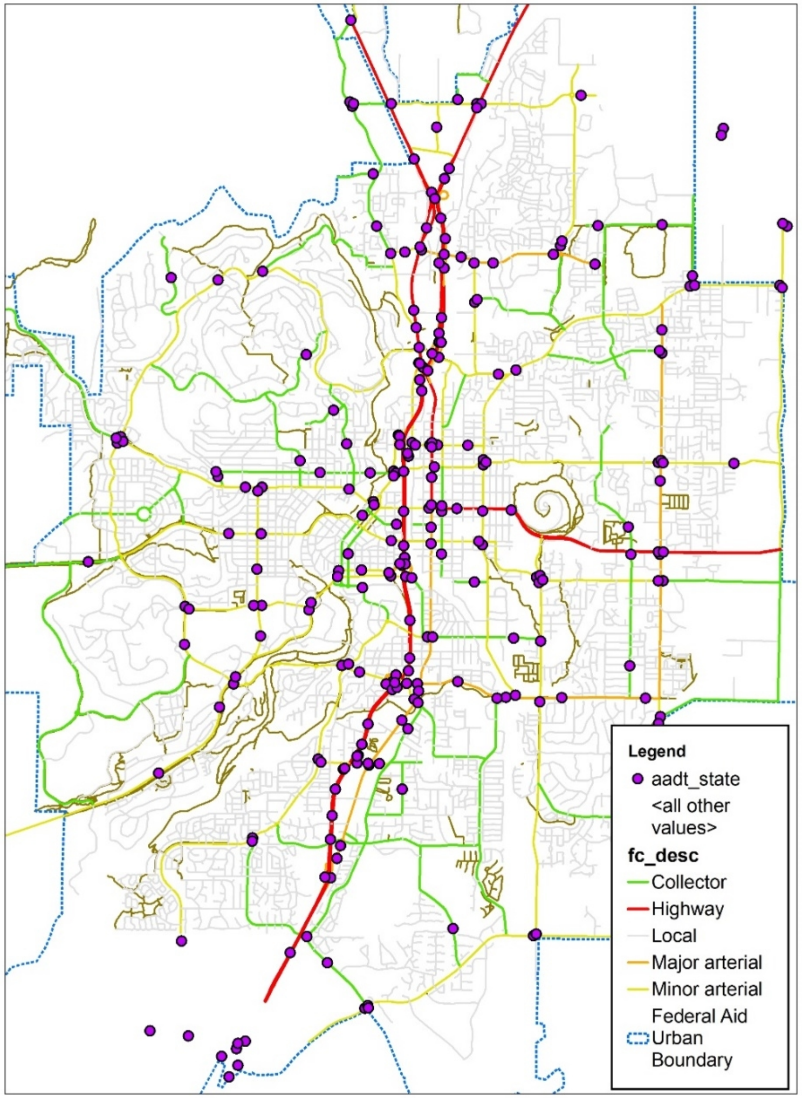
Vehicle AADT Data Fusion Scheme

- Vehicle Model Objectives
  - Validate data fusion approach
  - Provide network wide estimates of vehicle traffic
- Data and Models Used
  - Up to 433 data features in some specs
  - XgbBoost & Random Forest
  - Census, TAZ, properly attributed routable network
- Validation
  - Internal 10-fold cross validations (random partitions)
  - External 10-fold (stratified partition)
  - Leave-one-out validation
  - Comparison with Federally reported data (HPMS)



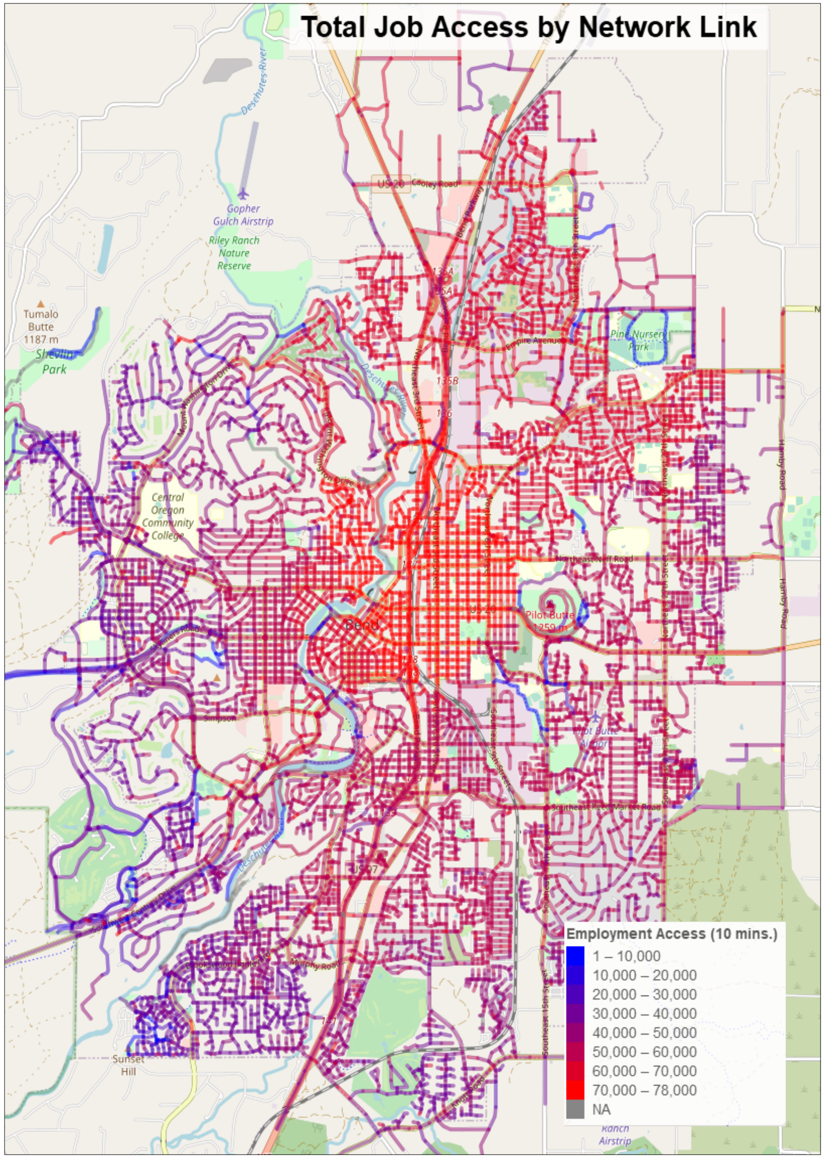
Vehicle AADT Model Data

- Vehicle Model Data
  - Traffic Counts
    - 2018 & 2019
    - N = 255
  - Network Features
    - Functional classification
    - Posted speed limit



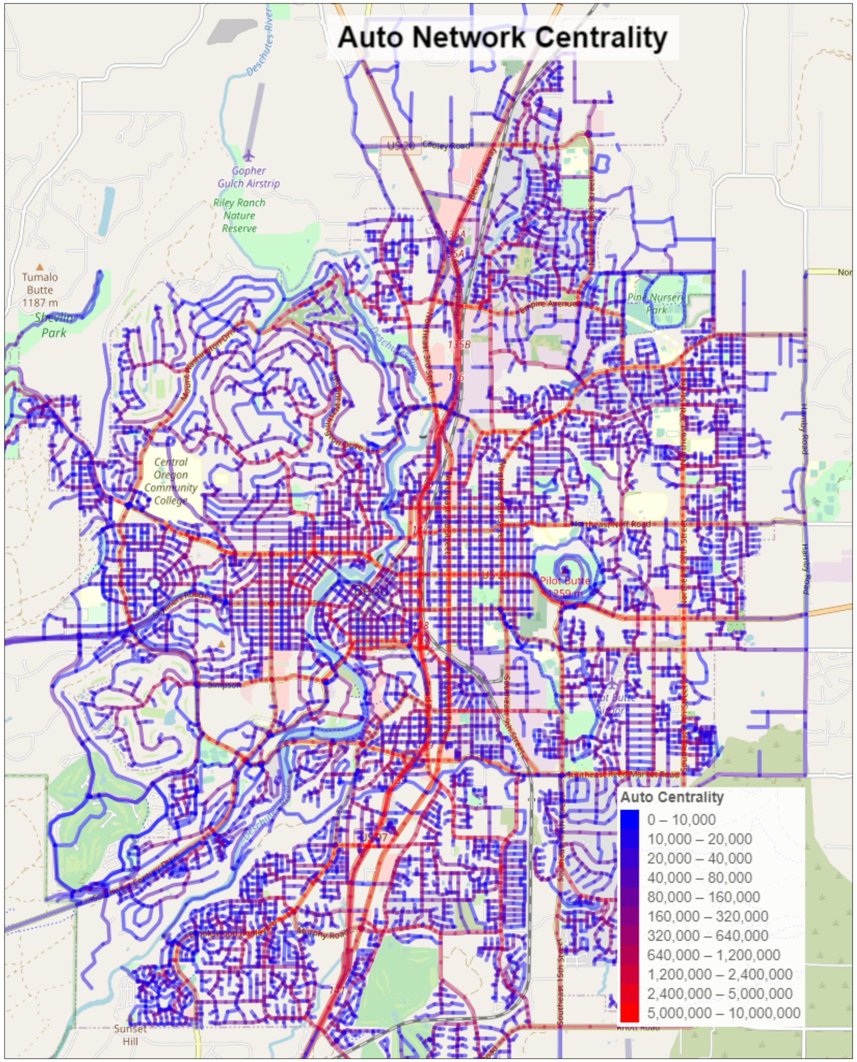
Vehicle AADT Model Data

- Vehicle Model Data
  - Traffic Counts
    - 2018 & 2019
    - N = 255
  - Network Features
    - Functional classification
    - Posted speed limit
  - Accessibility (drive time)
    - Jobs
    - People



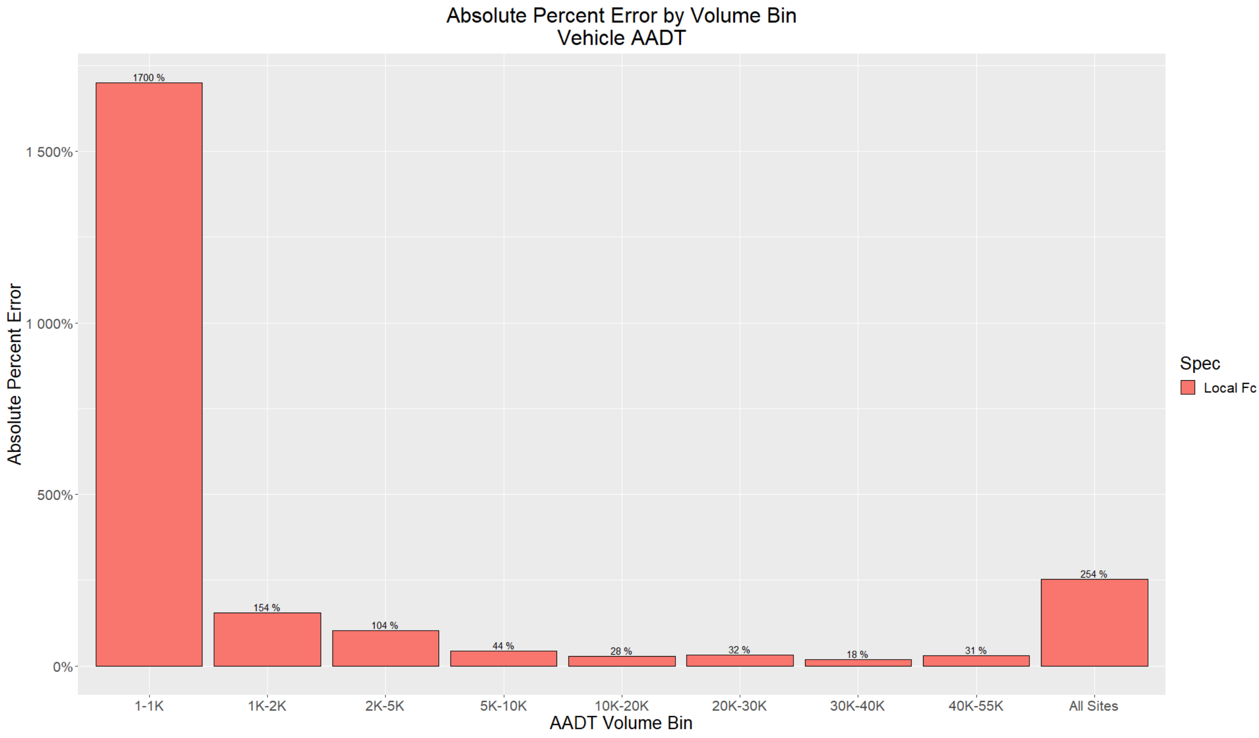
Vehicle AADT Model Data

- Vehicle Model Data
  - Traffic Counts
    - 2018 & 2019
    - N = 255
  - Network Features
    - Functional classification
    - Posted speed limit
  - Accessibility
    - Jobs
    - People
  - Centrality
    - Measures link importance



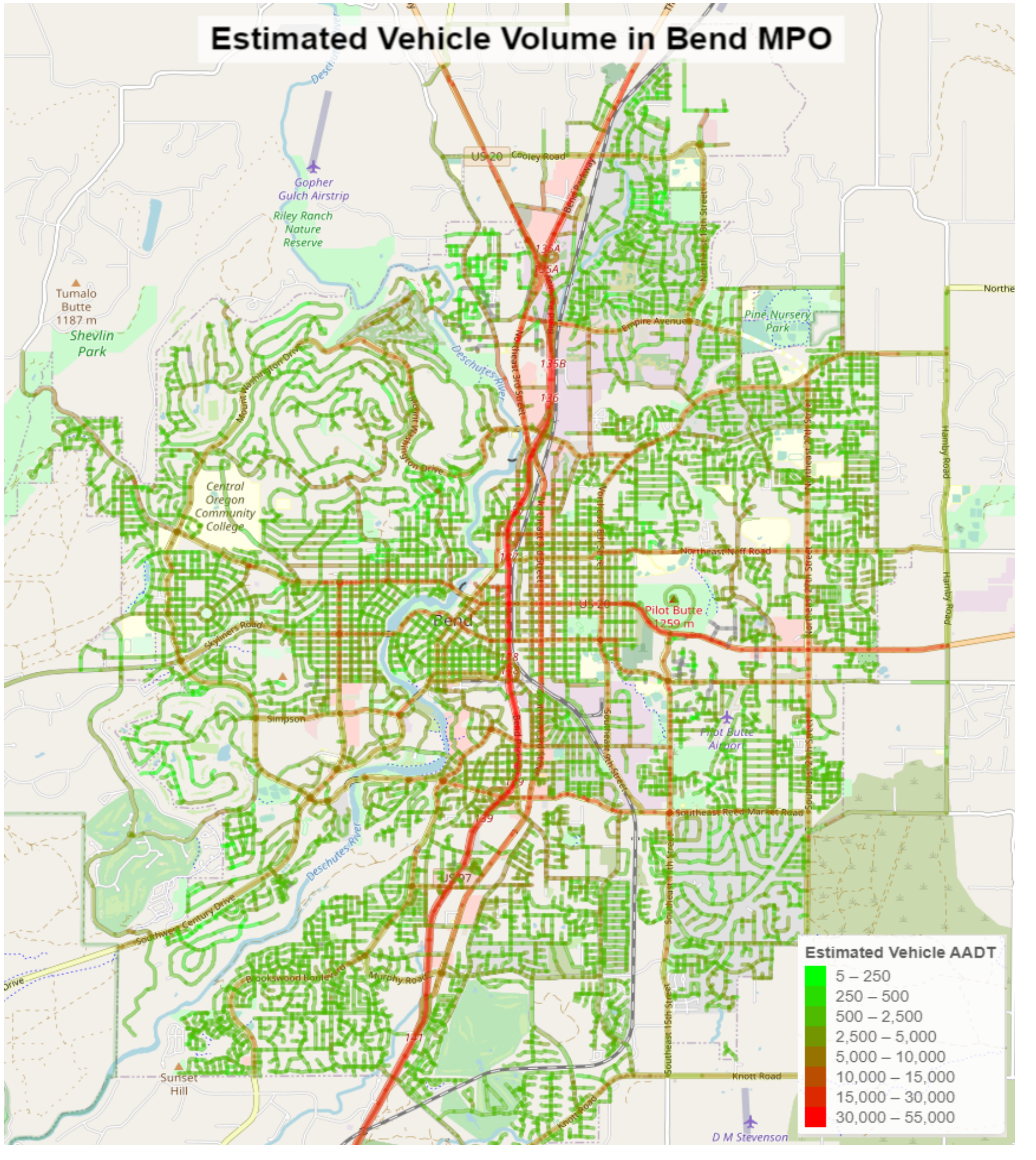
Vehicle AADT Model Validation

- 10-fold Cross-validation
  - Multiple specifications tried – local and federal fc
  - Prediction error varies by volume bin
  - Overall 254% error
  - 25% median error for volume bins 5K and greater



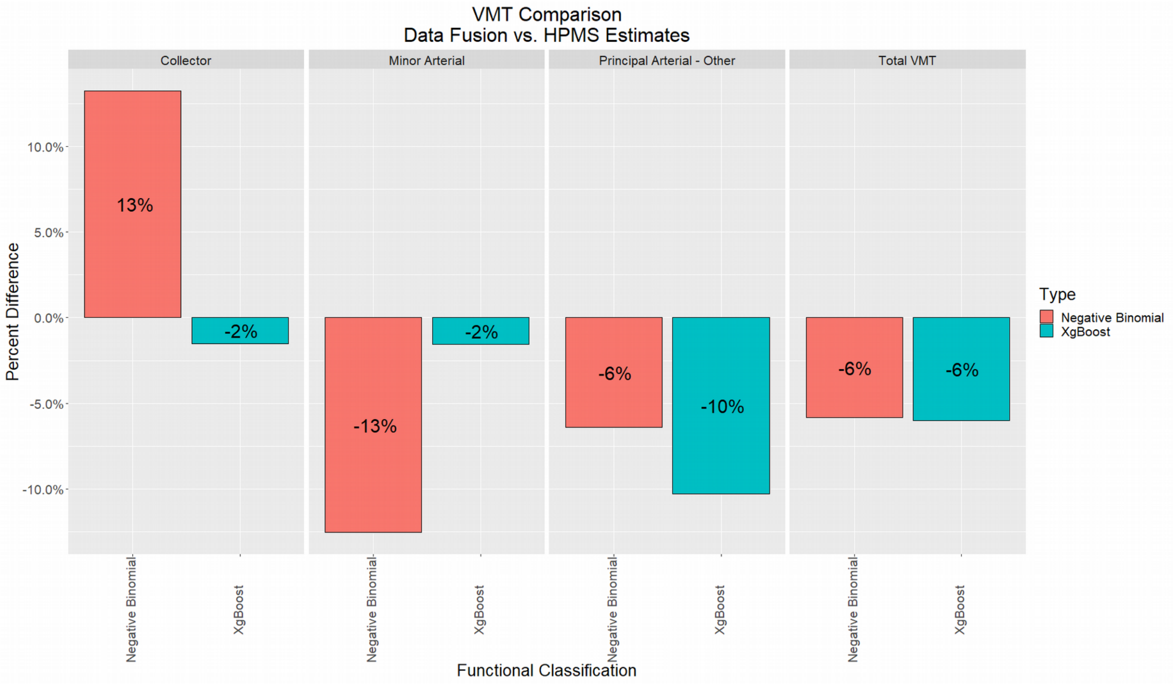
Vehicle AADT Model Results

- Network wide estimates
  - High volume roads appear reasonable
  - Low volume local streets appear overestimated



Vehicle AADT Model Results Comparison

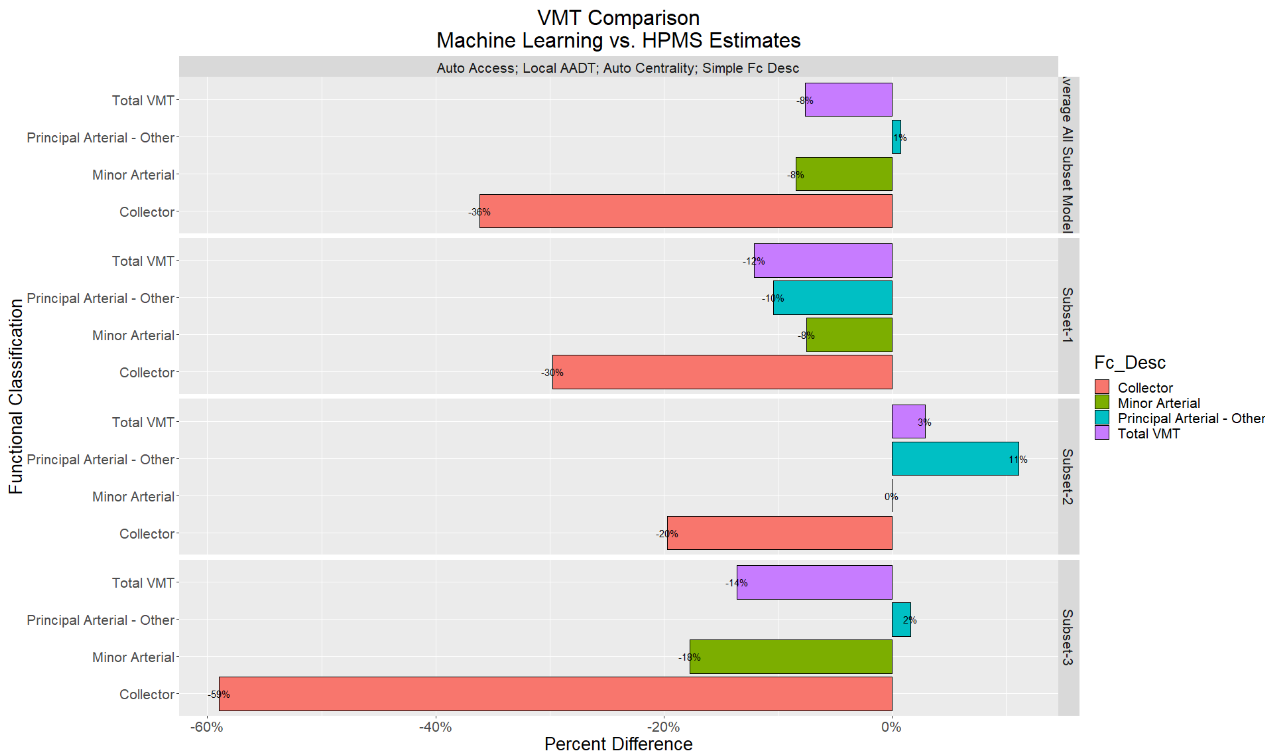
- Comparison with HPMS
  - Overall VMT estimate within 6% (model over estimates)
  - Model approaches provide reasonable system level estimates
  - Principal arterial highest error at 10% for ML
  - Collector & Min. Art. Highest error for Neg. Bin



Vehicle AADT Model Results Comparison

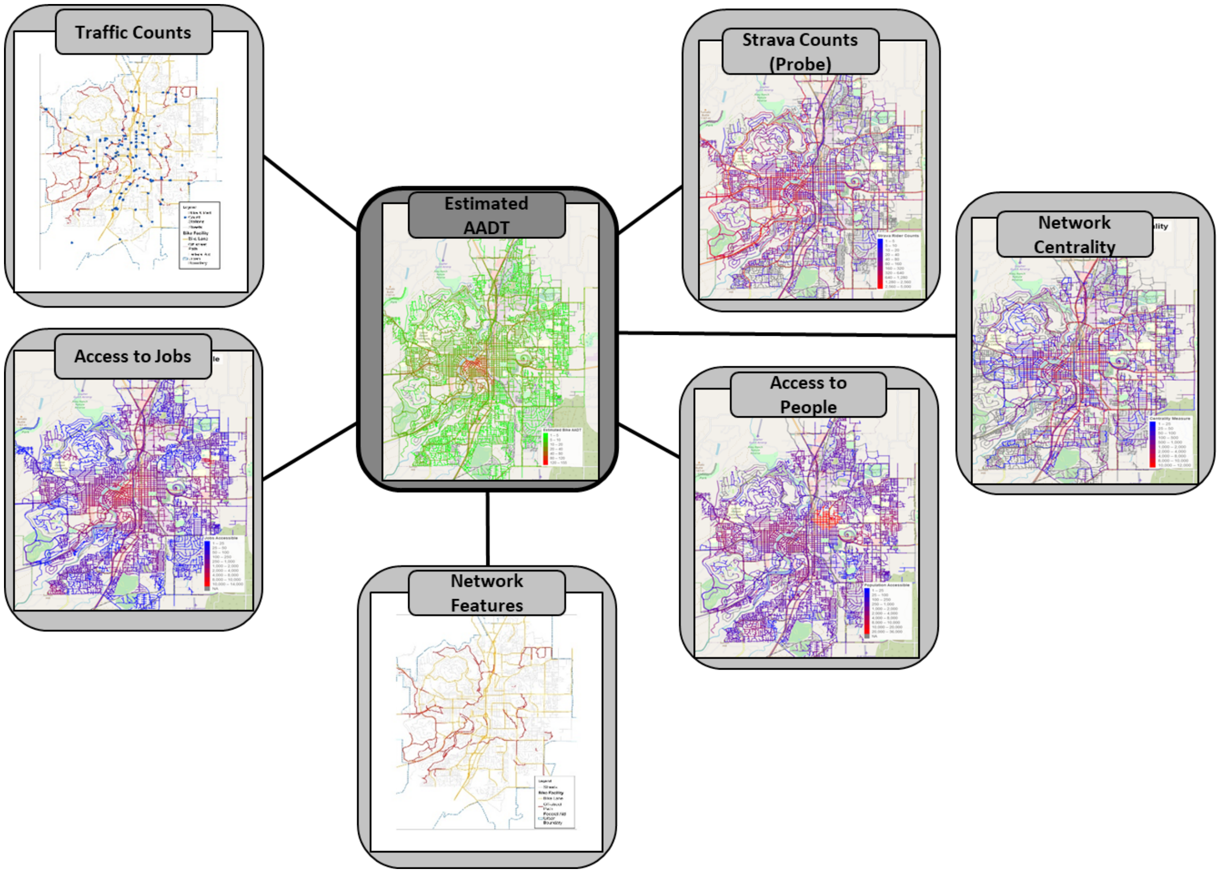
- Comparison with HPMS
  - Subset models are randomly partitioned into 3 datasets
  - Models within 3% to 14% compared to HPMS
  - Collectors perform poorly, likely due to small number of observations in training data

- Vehicle Model Conclusions
  - Approach performs well for aggregate and slightly disaggregate
  - Subset models improve confidence in
  - Disaggregate level useful in planning applications (& crash analysis?)
  - Results for each year available
  - Probe data will vastly improve approach (coming?)



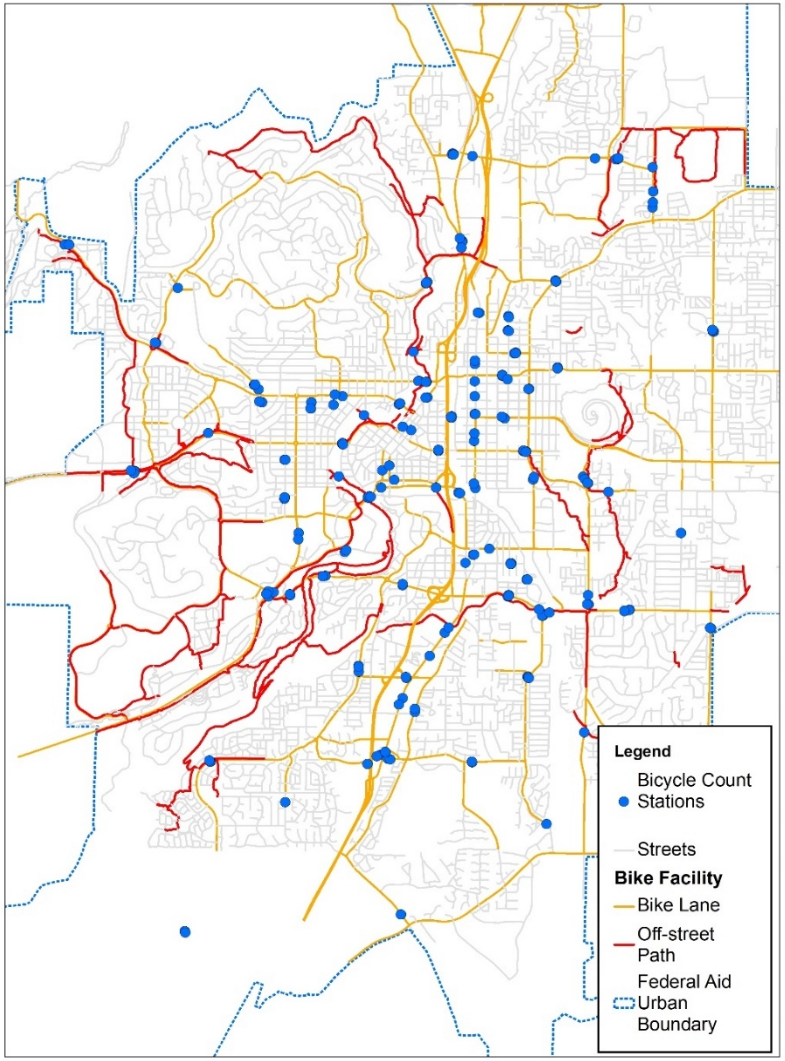
Bicycle AADT Data Fusion Scheme

- Bicycle Model Objectives
  - Provide network wide estimates of bicycle traffic
- Data and Models Used
  - Up to 516 data features in some specs
  - XgbBoost & Random Forest
  - Census, TAZ, properly attributed routable network, and probe data
- Validation
  - Internal 10-fold cross validations (random partitions)
  - External 10-fold (stratified partition)
  - Leave-one-out validation



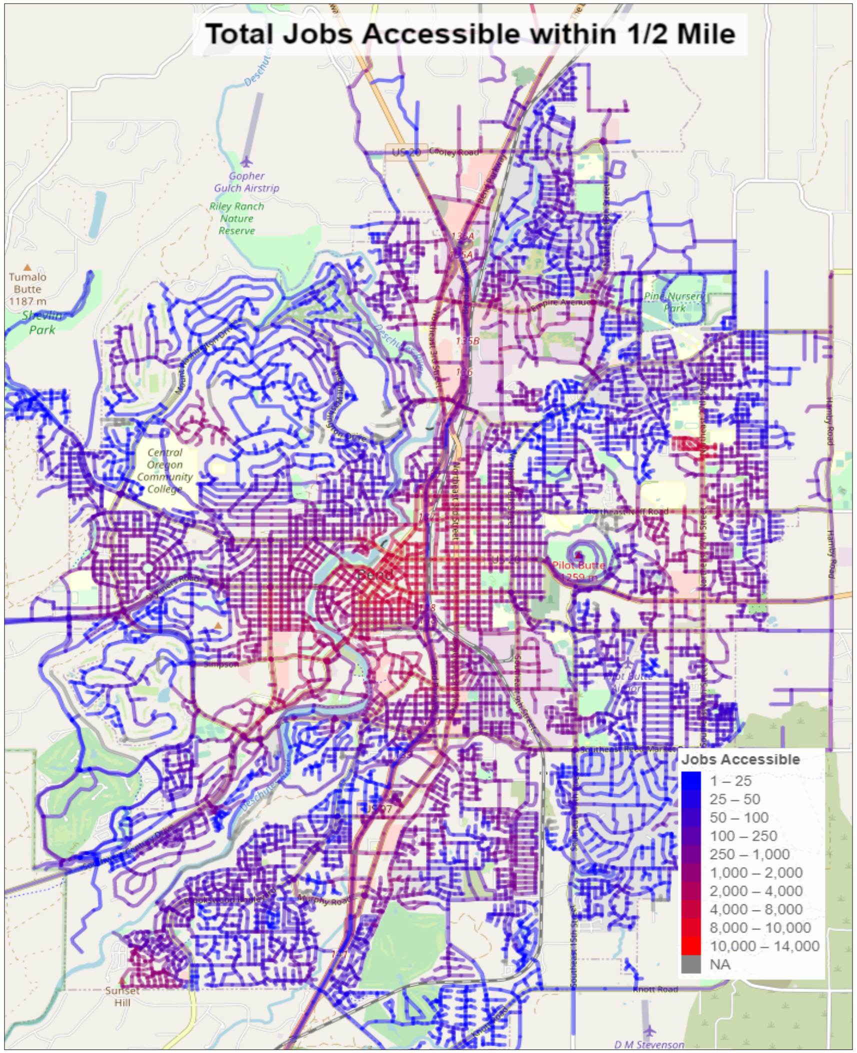
Bicycle AADT Model Data

- Bicycle Model Data
  - Traffic Counts
    - 2017, 2018 & 2019
    - N = 94
  - Network Features
    - Functional classification
    - Posted speed limit
    - Bicycle facility type



Bicycle AADT Model Data

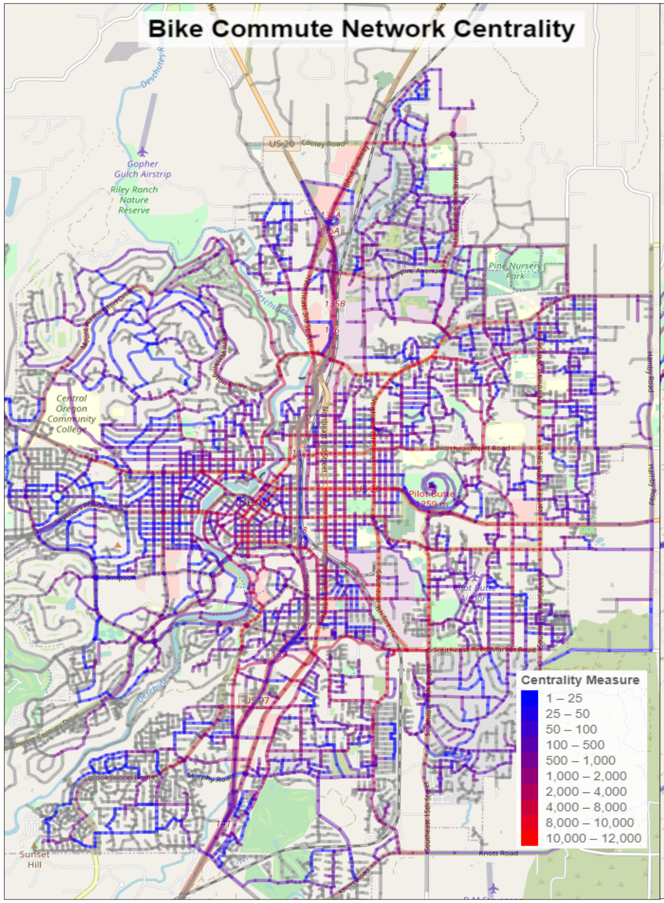
- Bicycle Model Data
  - Traffic Counts
    - 2017, 2018 & 2019
    - N = 94
  - Network Features
    - Functional classification
    - Posted speed limit
    - Bicycle facility type
  - Accessibility (distance)
    - Jobs
    - People



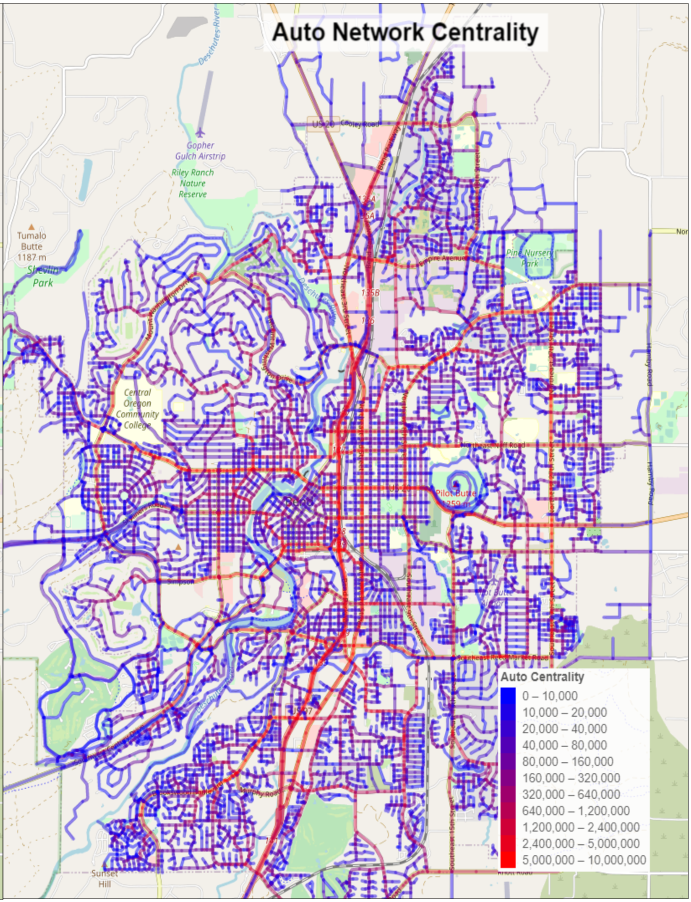
Bicycle AADT Model Data

- Bicycle Model Data
  - Traffic Counts
    - 2017, 2018 & 2019
    - N = 94
  - Network Features
    - Functional classification
    - Posted speed limit
    - Bicycle facility type
  - Centrality
    - Commute
    - Recreational
  - Accessibility (distance)
    - Jobs
    - People

Bicycle Centrality



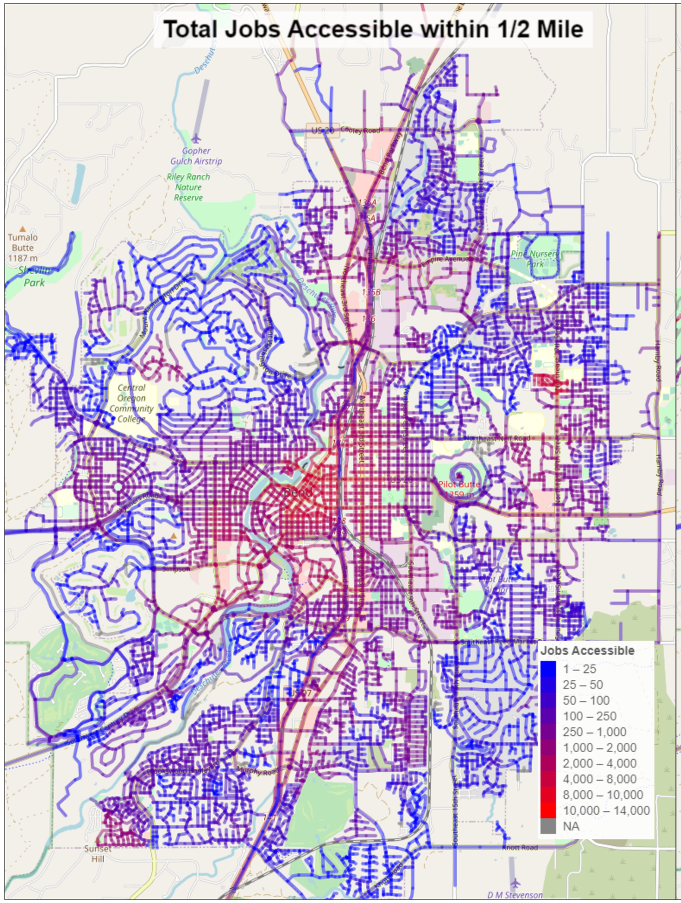
Vehicle Centrality



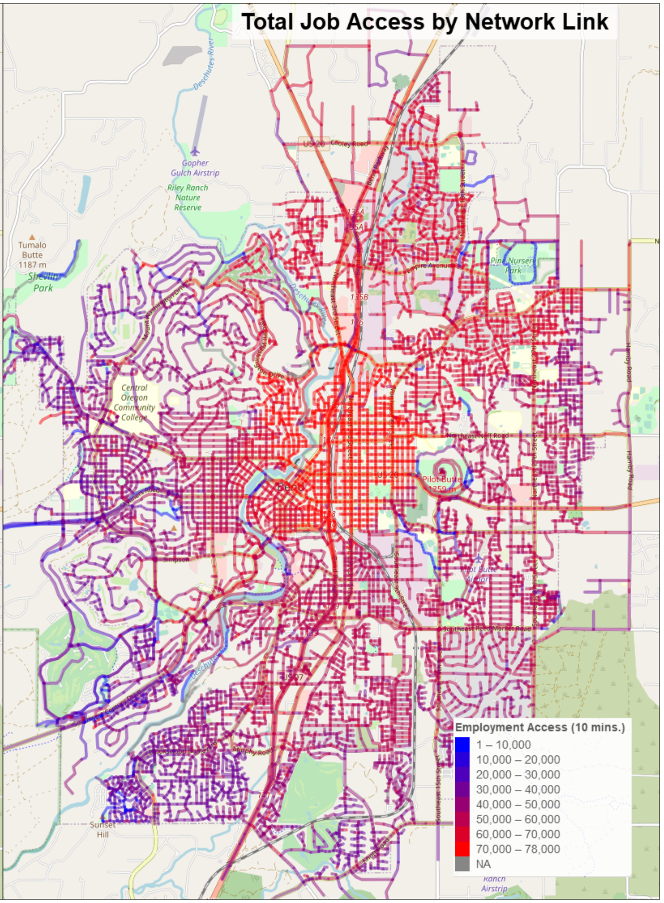
Bicycle AADT Model Data

- Bicycle Model Data
  - Traffic Counts
    - 2017, 2018 & 2019
    - N = 94
  - Network Features
    - Functional classification
    - Posted speed limit
    - Bicycle facility type
  - Accessibility (distance)
    - Jobs
    - People

Bicycle Access

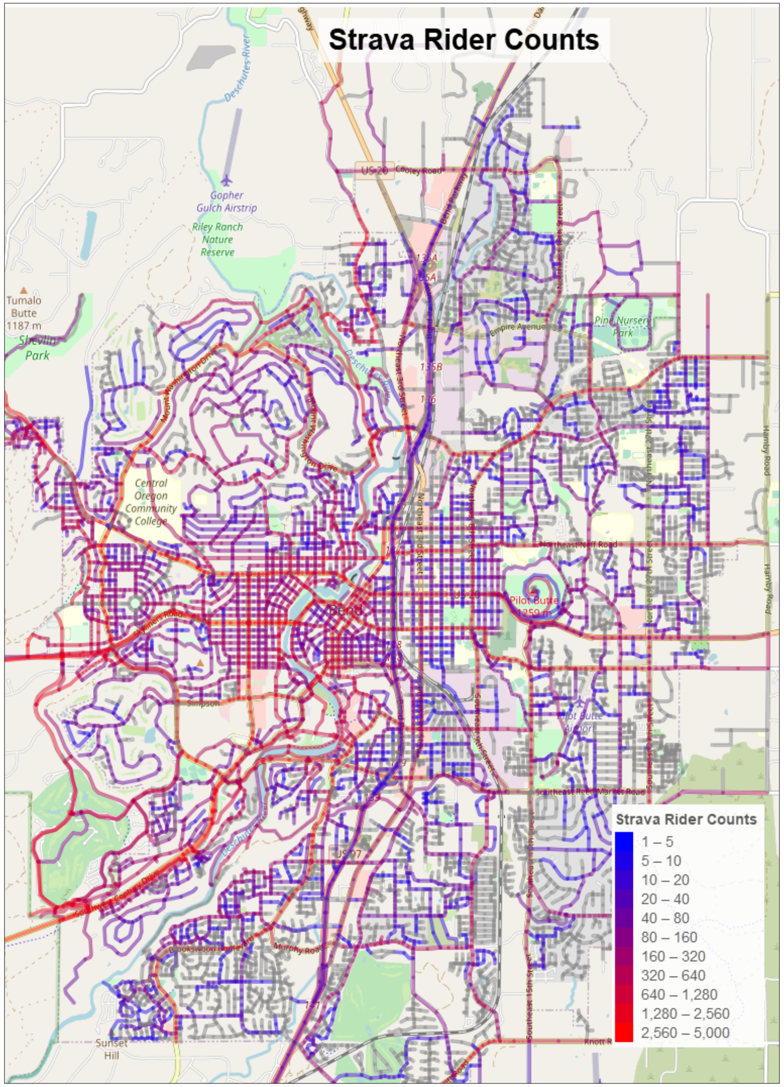


Vehicle Access



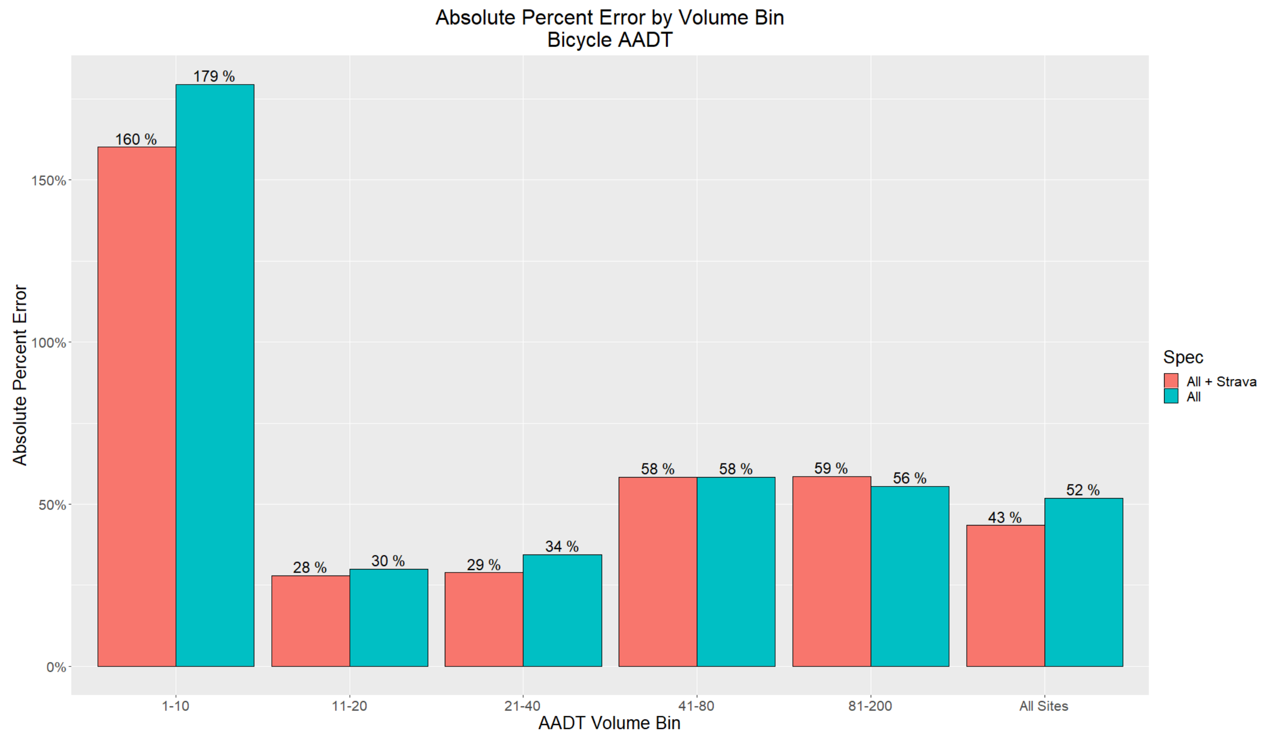
Bicycle AADT Model Data

- Bicycle Model Data
  - Traffic Counts
    - 2017, 2018 & 2019
    - N = 94
  - Network Features
    - Functional classification
    - Posted speed limit
    - Bicycle facility type
  - Centrality
    - Commute
    - Recreational
  - Accessibility (distance)
    - Jobs
    - People
  - Probe Data
    - Strava
    - 2017-2019 data



Bicycle AADT Model Validation

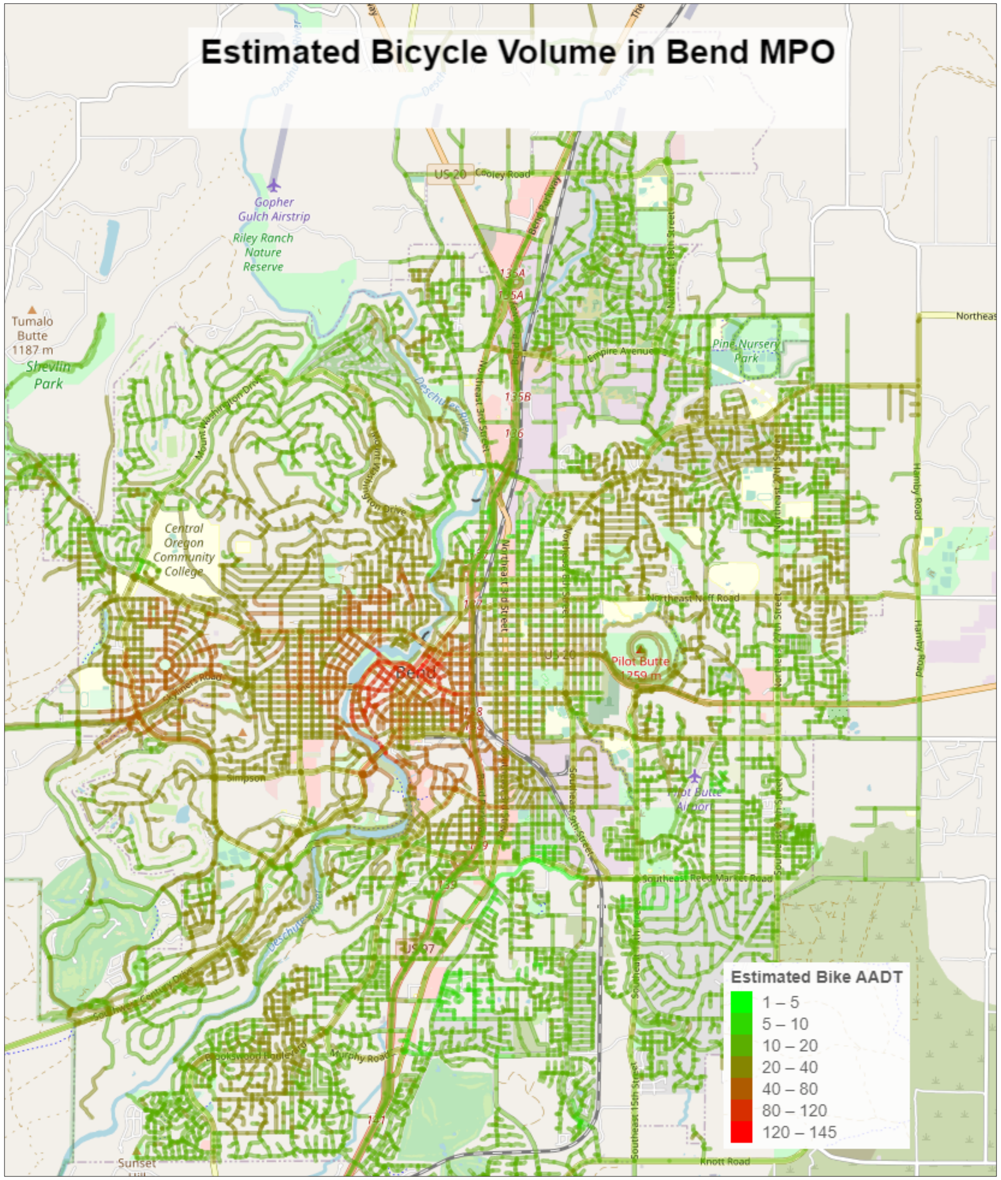
- 10-fold Cross-validation
  - Multiple specifications tried – without Strava and with
  - Overall 43% error (All + Strava model)
  - Prediction error varies by volume bin
  - Low Volumes makes modeling a challenge
  - Probe data helps in accuracy (but even more in application)



Bicycle AADT Model Results

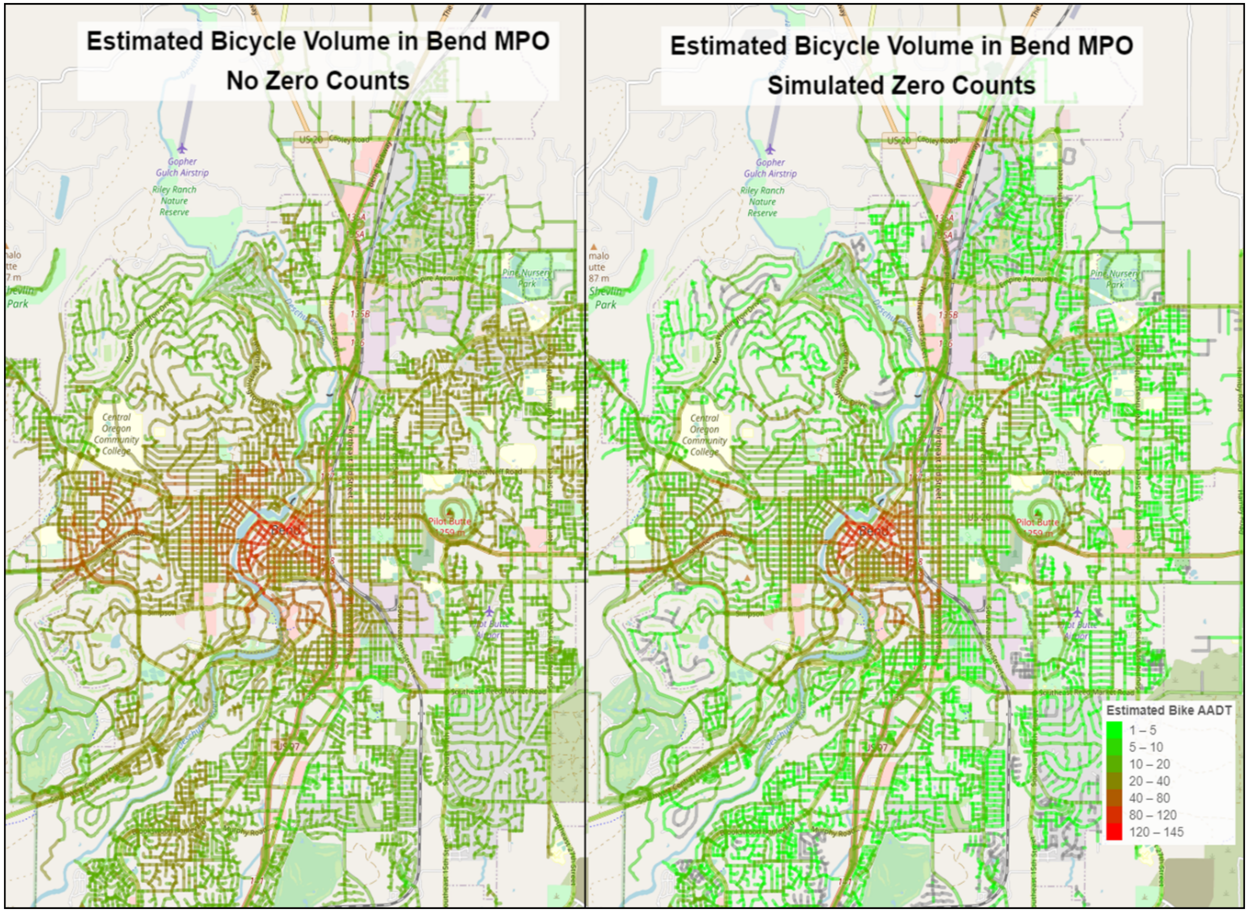
- Network wide estimates
  - Looks reasonable, but how to tell?
  - Activity concentrated near employment centers
  - Appears to estimate too much bike activity in low density residential areas

- Handling Lack of Zero Counts
  - Random selection of streets high likelihood of zero bike traffic
  - Criteria: local street; low population; density; low centrality, no Strava, no bike facility



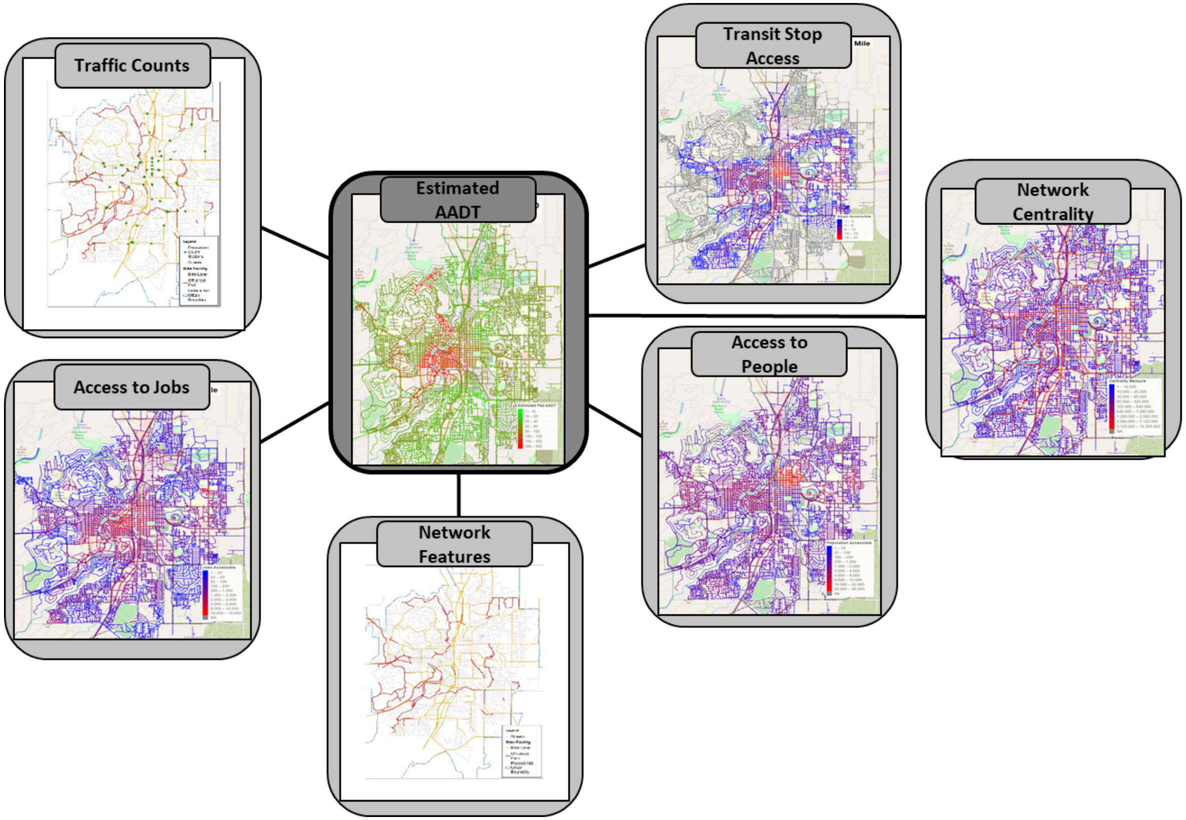
Bicycle AADT Model Results

- Handling Lack of Zero Counts
  - Random selection of streets high likelihood of zero bike traffic
  - Criteria: local street; low population; density; low centrality, no Strava, no bike facility
- Results
  - Moderates volume well in expected areas
  - Significantly decreases overall BMT



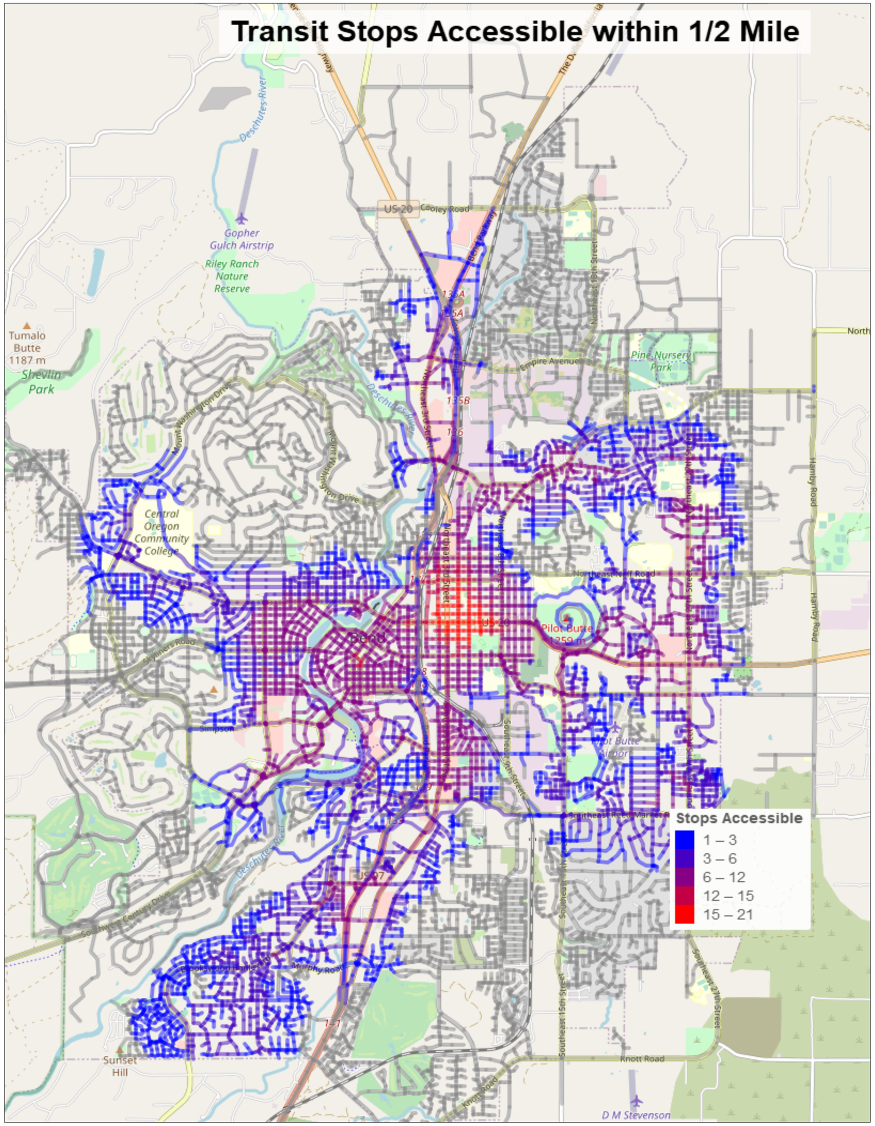
Pedestrian AADT Data Fusion Scheme

- Pedestrian Model Objectives
  - Provide network wide estimates of pedestrian traffic
- Data and Models Used
  - Up to 512 data features in some specs
  - XgbBoost & Random Forest
  - Census, TAZ, properly attributed routable network, and transit data
- Validation
  - Internal 10-fold cross validations (random partitions)
  - External 10-fold (stratified partition)
  - Leave-one-out validation



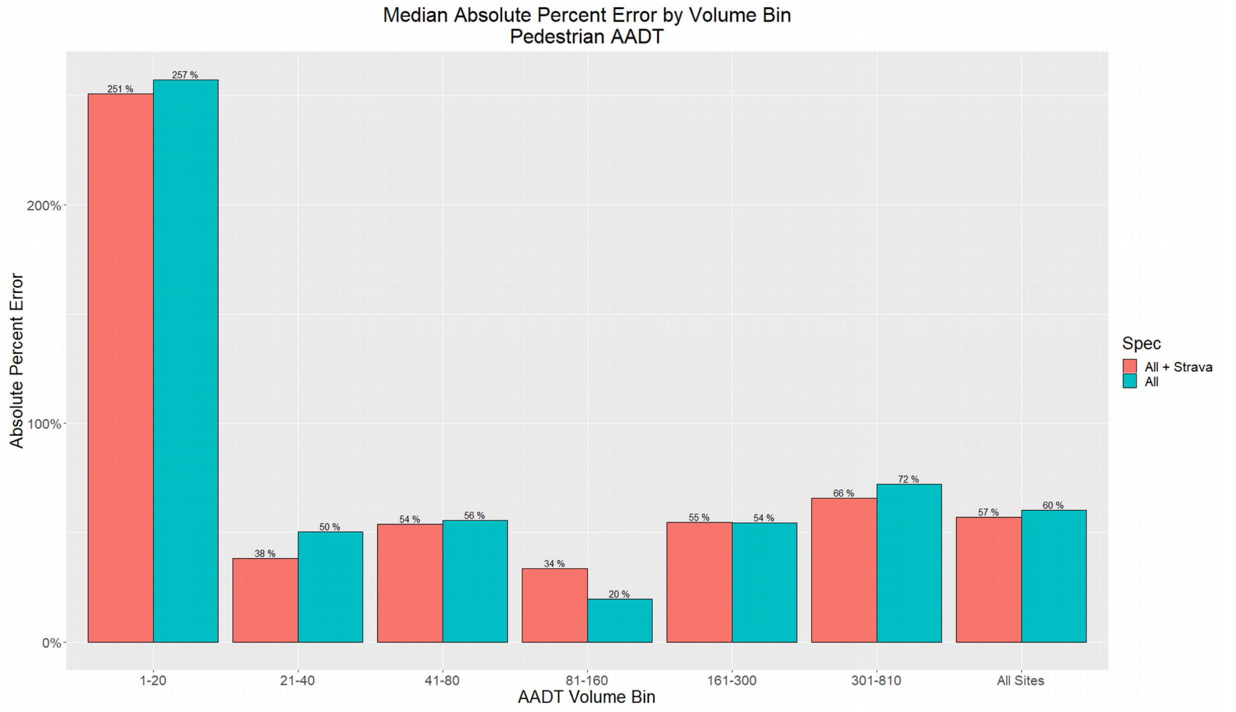
Pedestrian AADT Model Data

- Bicycle Model Data
  - Traffic Counts
    - 2017, 2018 & 2019
    - N = 56
  - Network Features
    - Functional classification
    - Posted speed limit
    - Off street system
  - Centrality
    - Commute
    - Recreational
    - Shortest
  - Accessibility (distance)
    - Jobs
    - People
  - Transit Stop Access
    - Ridership would be bette



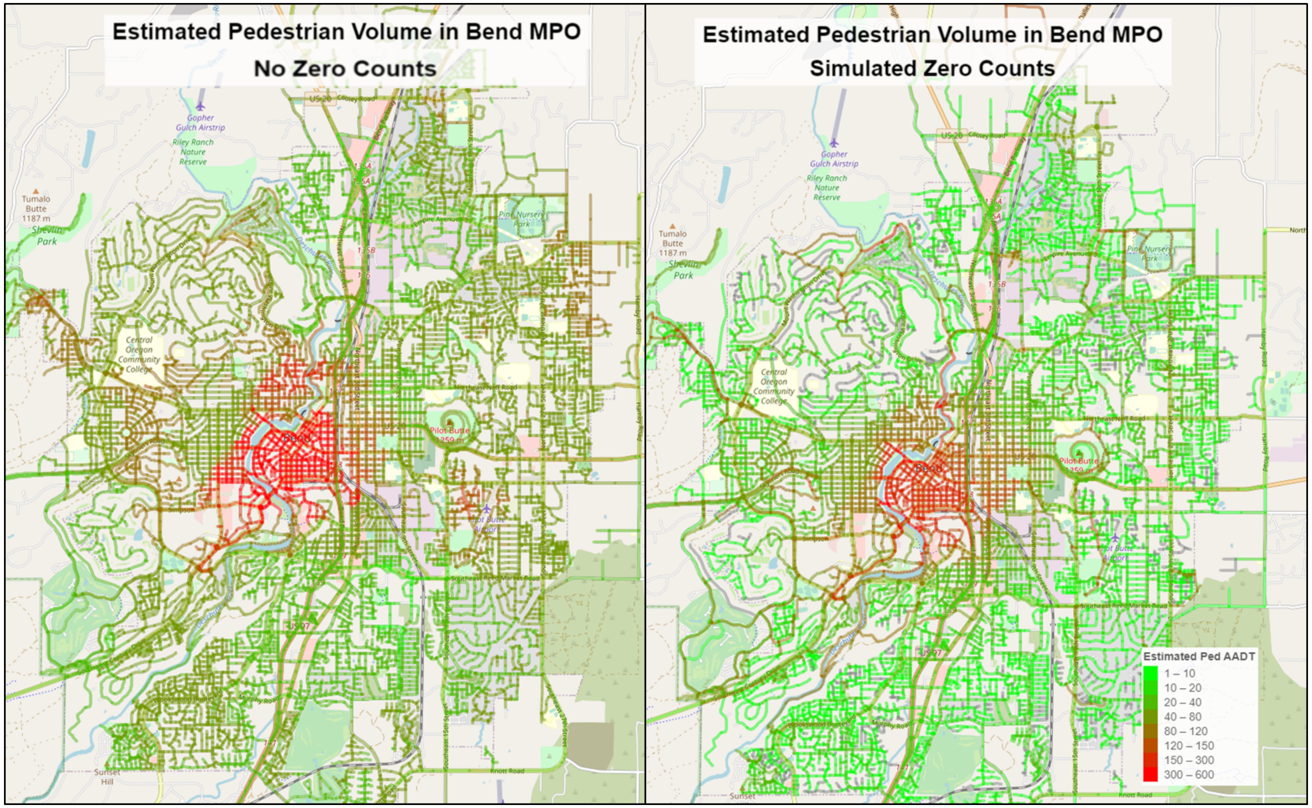
Pedestrian AADT Model Validation

- 10-fold Cross-validation
  - Multiple specifications tried – without Strava and with
  - Overall 57% error
  - Prediction error varies by volume bin
  - Low volumes makes modeling a challenge
  - Probe data helps (surprisingly)



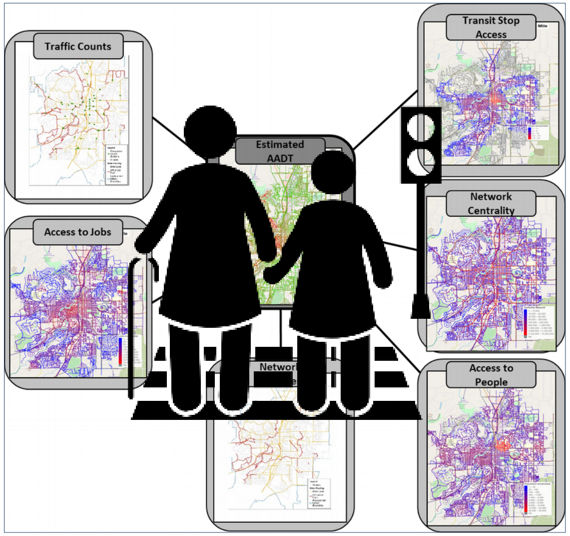
Pedestrian AADT Model Results

- Network wide estimates
  - Looks reasonable, but how to tell?
  - Activity concentrated near employment centers
  - Appears to estimate too much bike activity in low density residential areas



- Handling Lack of Zero Counts
  - Random selection of streets high likelihood of zero bike traffic
  - Criteria: local street; low population; density; low centrality, no Strava, no bike facility

Data Fusion Wrap-up

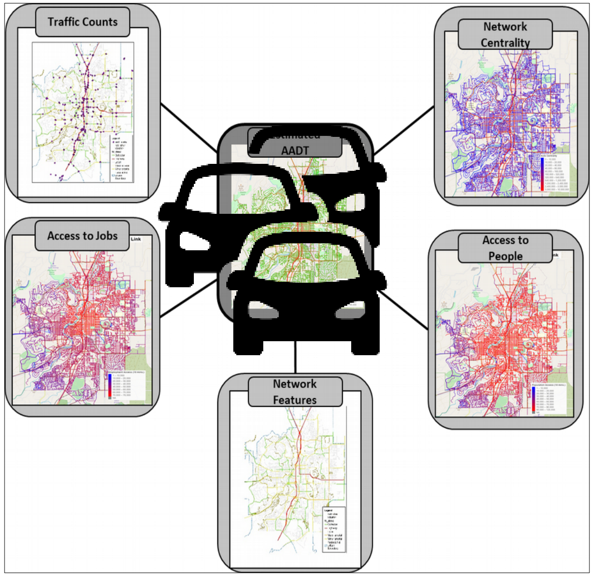


Limitations

- Need more counts
- Input features not all concurrent with counts (population & employment)
- No probe data for vehicles (or ped specific)
- Feature space could be reduced

Conclusions

- Information from models can inform multiple purposes
- More counts will improve the model
- Future discussions needed to determine further applications



Research Project Next Steps

Short term

- Crash data analysis – TAC meeting # 4
- Transfer data processing and related knowledge to Bend area staff
- Provide ongoing tech support for CPiR
- Develop useful data visualizations and data access
- NITC Pooled Fund December 2020

Longer term

- Statewide data support (centralized repository, QAQC) – many pathways to statewide program
- Institutionalize data fusion models for monitoring planning (incorporate NITC results)
- Pilot in another Oregon urban area
- UMD and I-95 Corridor Coalition (RITIS?)
- Better prepare for third-party platform offers – more evaluations of products (e.g. Streetlight Data evaluation)

Final Report (June/July 2020)

Deliverable	Description	Intended Audience
Data Collection	Describes the equipment and data collection strategy employed in this research	**Data Program Managers; Data Collection Staff and Contractors**
Annual Traffic Estimation	Develops and applies a new method for creating annual estimates of bicycle counts from daily counts	**Data Program Managers; Safety Analysts**
Total Bicycle Activity Estimation	Application of statistical models using annual bicycle counts and various infrastructure, accessibility and connectivity variables to estimate total Bicycle/Pedestrian miles traveled (BMT)	**Transportation Analysts; Modelers; Planners**
Crash Analysis	Employs bicycle miles traveled in crash analysis to assess risk and develop safety performance functions and (SPF)crash modification factors (CMF)	**Safety Analysts; Engineers; Planners**

National Institute of Transportation & Communities Pooled Fund

Objective

- Develop acceptance criteria for 3rd party data
- Activity estimates for entire network (just bikes)

Partners

- Oregon (Bend MPO, Central Lane MPO, PBOT, ODOT)
- Colorado DOT
- Virginia DOT
- Utah DOT
- DCDOT

Exploring Data Fusion Techniques to Derive Bicycle Volumes on a Network

Sirisha Kothuri
Joe Broach
Nathan McNeil
Portland State University

Kate Hyun
Steve Mattingly
University of Texas Arlington

Krista Nordback
The University of North Carolina at Chapel Hill

Frank Proulx
Toole Design

Questions

Questions?

Josh Roll Active and Sustainable Transportation Research Coordinator

Josh.F.Roll@ODOT.state.or.us

RESEARCH

Oregon Department of Transportation

Back up

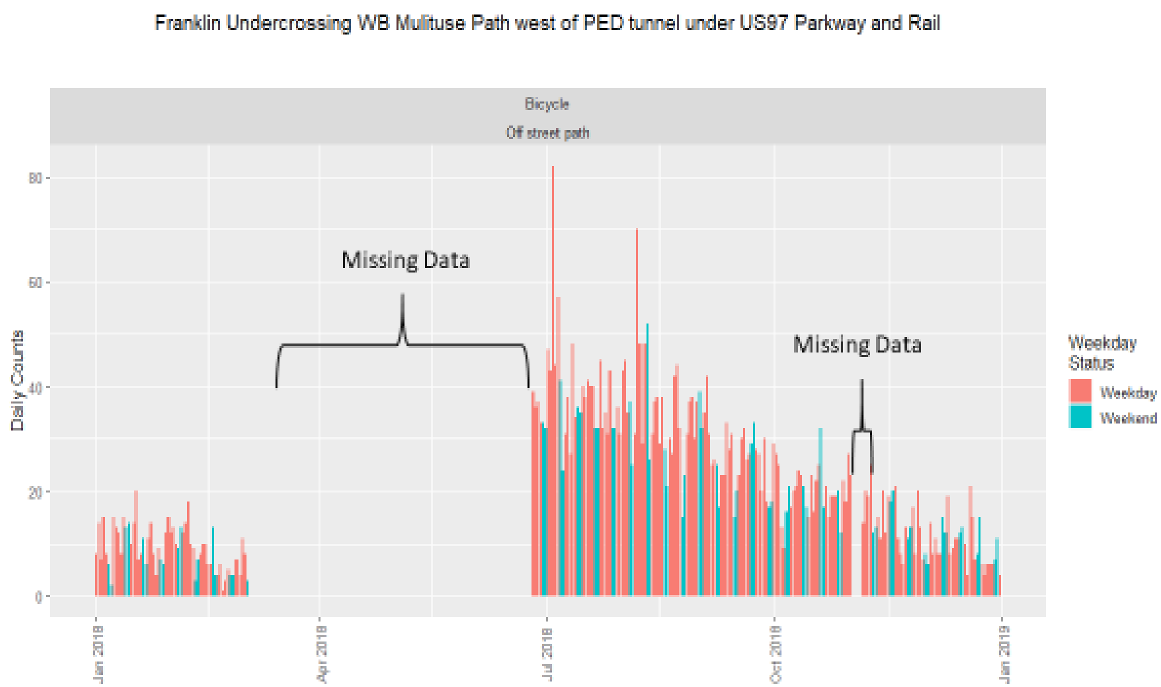
Traffic Data Imputation - What’s the Problem?

Why Missing Data?

- Equipment failure at permanent sites (bugs!)
- Data Transfer Issues

Solution:

- Traffic variation highly dependent on weather and day of week factors
- (Hanson and Hanson 1977; Niemeier 1996; Nankervis 1999;Richardson 2000; Brandenburg 2007; Rose et al. 2011; Tin Tin et al. 2012, Thomas, Jaarsma, and Tutert 2009; Lewis 2011;Gallop, Tse, and Zhao 2012; Miranda-Moreno and Nosal 2011; Nosal and Miranda-Moreno 2012; Schmiedeskamp and Zhao 2016).



Daily Imputation and Annual Estimation

Results by Months Used

- More months of data equals better results
- Likely scenario is 3 months or less of missing data
- 2-10% error when 9 months of data used

Limitations

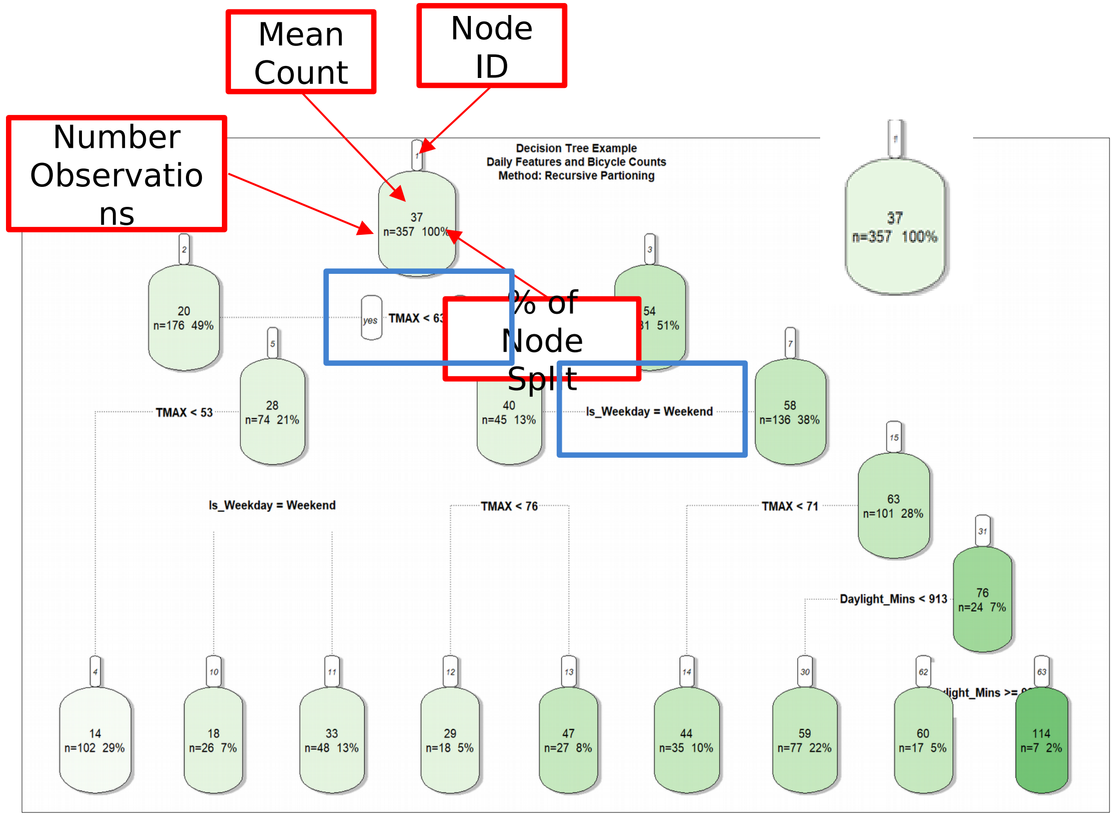
- Only using 1 year of data but results would be better if multiple years of data are used
- Negative Binomial does poorly when data poor

	Bicycle				Pedestrian			
	Negative Binomial		Random Forest		Negative Binomial		Random Forest	
Number of Months Used in Training	95th Pct.	Median	95th Pct.	Median	95th Pct.	Median	95th Pct.	Median
1	26,288%	38%	84%	34%	244%	18%	68%	20%
2	131%	14%	53%	11%	53%	9%	40%	10%
3	50%	10%	34%	7%	28%	6%	27%	7%
4	35%	7%	23%	5%	19%	5%	20%	6%
5	28%	5%	18%	4%	14%	4%	17%	4%
6	22%	4%	15%	3%	11%	3%	14%	4%
7	18%	3%	12%	2%	9%	2%	12%	3%
8	15%	3%	10%	2%	7%	2%	10%	3%
9	12%	2%	8%	1%	6%	2%	8%	2%
10	9%	2%	6%	1%	4%	1%	6%	2%
11	6%	1%	4%	1%	3%	1%	4%	1%

Wait what is Machine Learning Aga

Tree Split Stopping Rules/Criteria

- Guided by rules of impurity reduction with an aim of creating daughter nodes more pure than parent nodes
- Impurity quantified by GINI Index or Shannon Entropy
- Given a minimum # of observations left in node



Traffic Count Imputation Example

- TMAX – most important
- Weekday variable – also important
- Minutes of daylight – *also* important

Ensembles

- Example is single tree
- Multiple trees estimated
- Combined to create a forest!

Daily Imputation and Annual Estimation

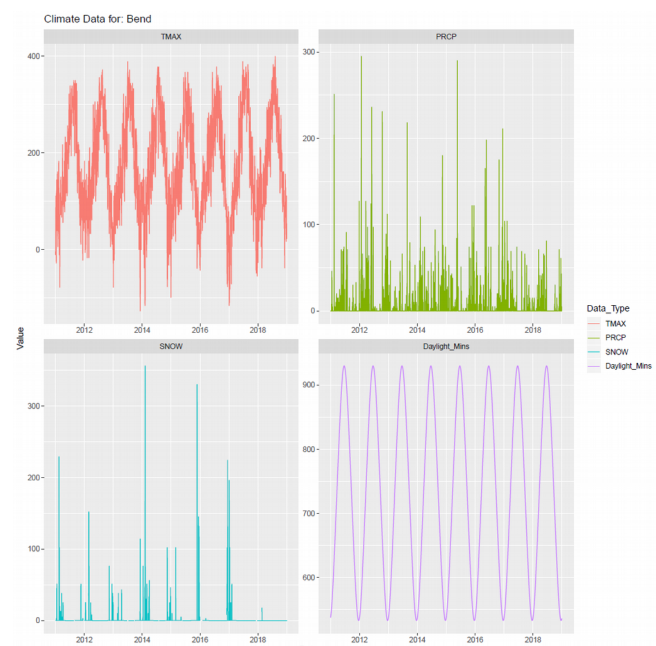
Data
- 21 unique locations from statewide data
- All sites have at least 98% of annual data

Imputation
- Machine learning (random forest, conditional inference, recursive partitioning)
- Negative binomial regression

Test setup
- Use permanent counters from around the state
- Hold out all possible combinations of month

City	User Type	Daily Counts Summary				Number Locations	
		Mean	Median	Std. Dev.	Records	Unique	Year/Location*
Bend	Bicycle	56	43	55	2,167	5	6
Bend	Pedestrian	148	99	150	2,907	7	8
Eugene	Bicycle	340	275	240	1,095	3	3
Eugene	Pedestrian	491	303	450	1,824	5	5
Portland	Bicycle	1,957	1,720	1,402	728	1	2
Salem	Bicycle	38	32	32	365	1	1
Springfie	Bicycle	185	125	182	1,460	4	4
Springfie	Pedestrian	103	97	42	365	1	1
Total	Total	327	116	627	10,911	21	30

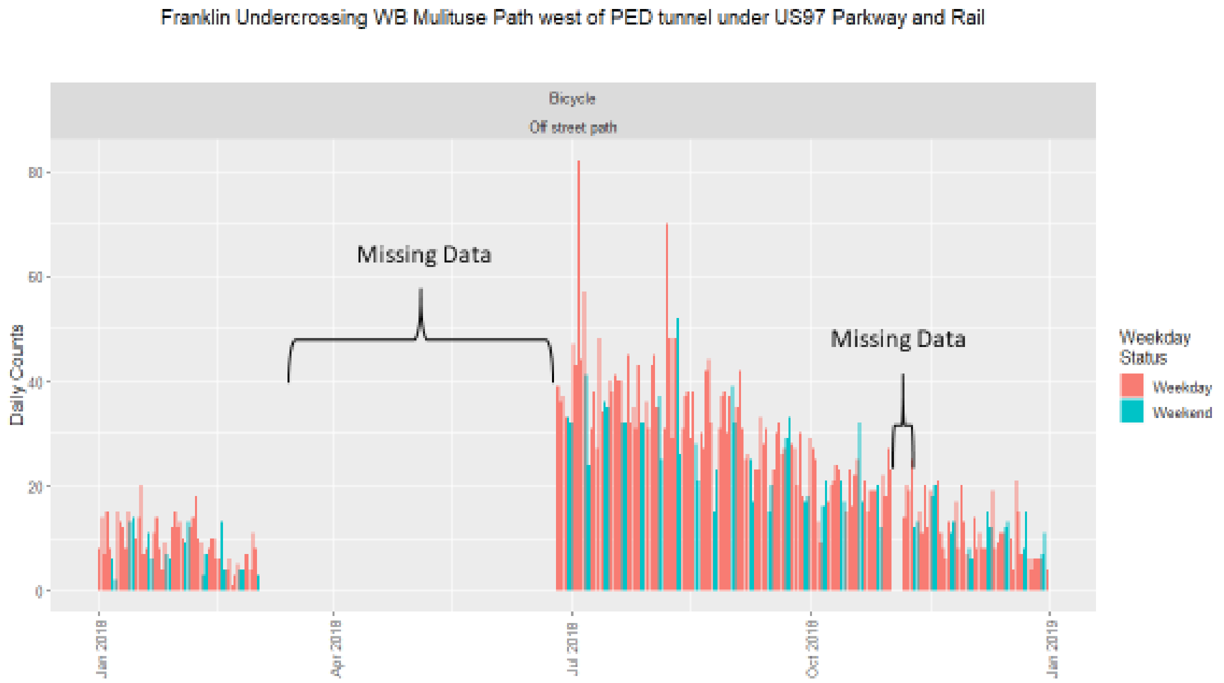
NOAA Data



Why Machine Learning?

Why Machine Learning?

- Negative Binomial Regression used previously (SARM) Roll and Proulx 2017
- Shown to predict annual traffic within 5% with just 3 weeks of counts
- But how to select best model?
- Interaction effects better captured in ML



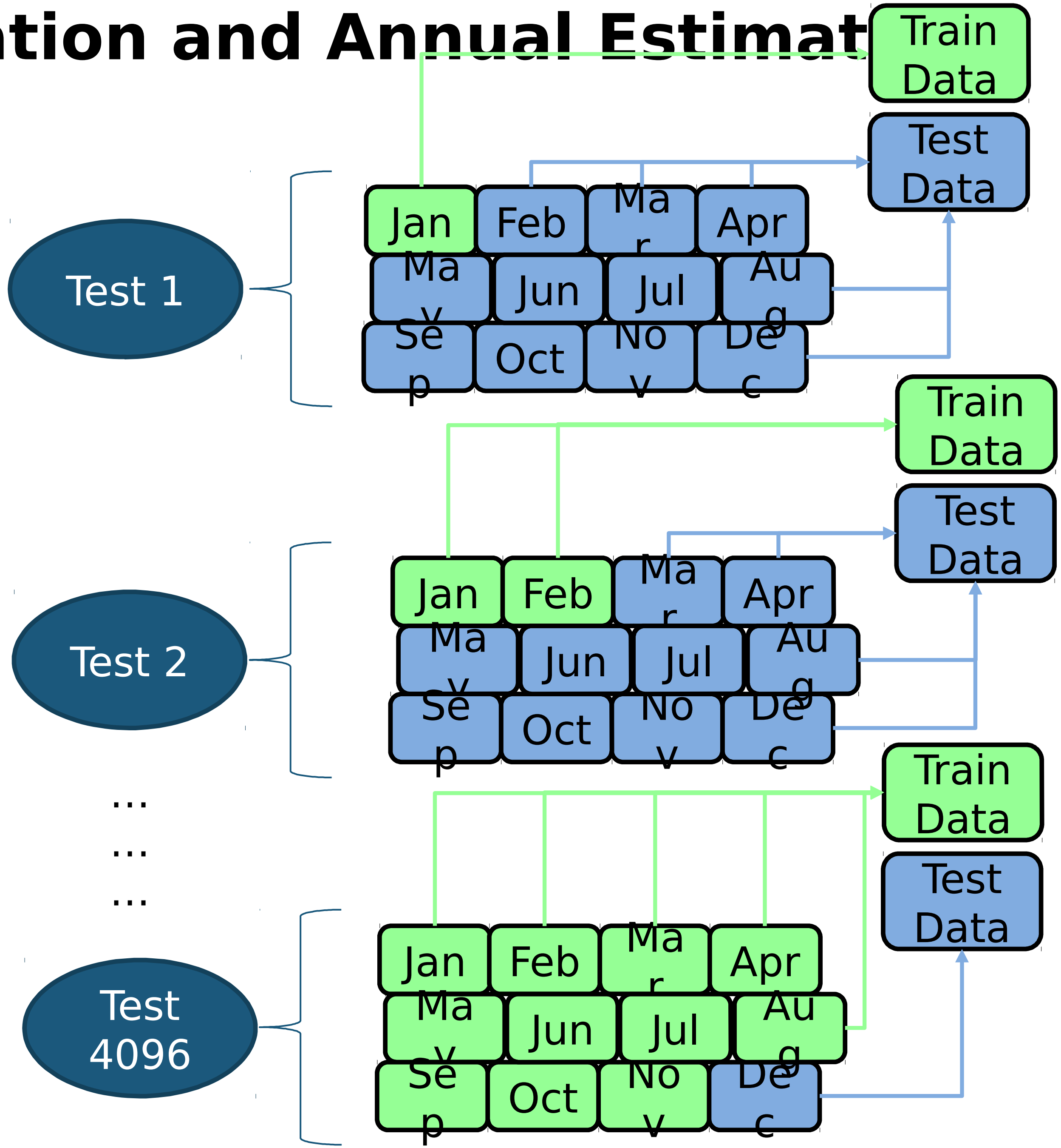
Daily Imputation and Annual Estimat

Imputation

- Machine learning (using recursive partitioning regression trees, random forests, conditional inference)
- Negative binomial regression

Test setup

- Use permanent counters from around the state
- Estimate **daily** traffic counts
- Hold out all possible (4,096) combinations of month
- Measure monthly and annual error



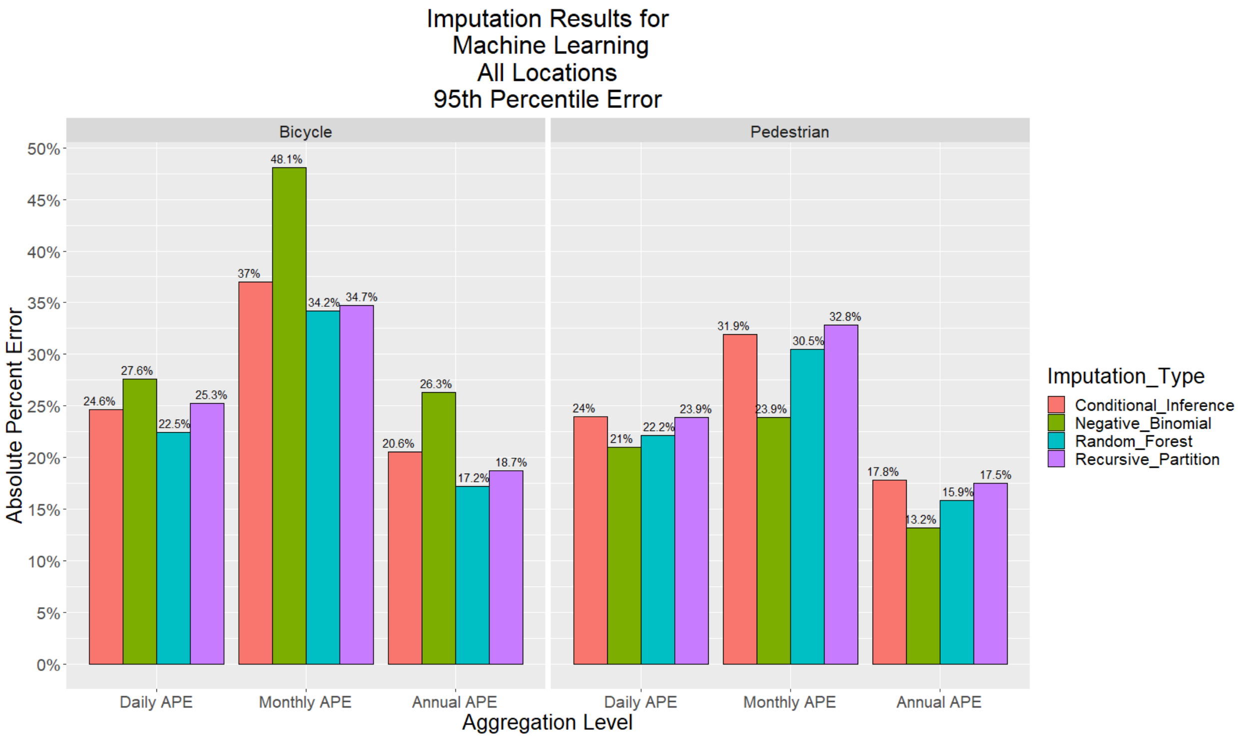
Daily Imputation and Annual Estimation

Results

- 3 levels of estimation
- Bikes – Random Forest works best
- Peds – Close tie between negative binomial and random forest

Test setup

- Use permanent counters from around the state
- Hold out all possible combinations of month



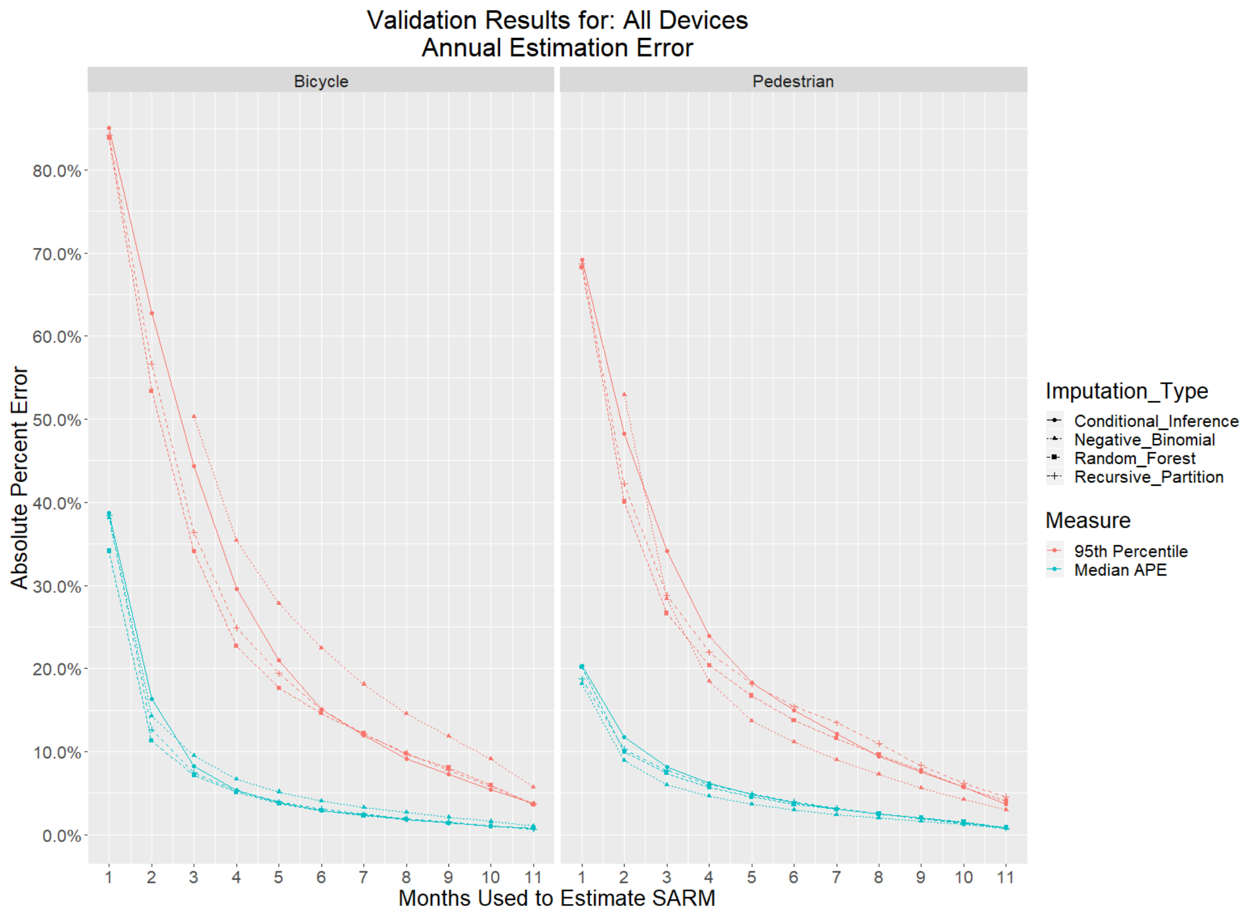
Daily Imputation and Annual Estimation

Results by Months Used

- More months of data equals better results
- Likely scenario is 3 months or less of missing data
- 2-10% error when 9 months of data used

Limitations

- Only using 1 year of data but results would be better if multiple years of data are used
- Negative Binomial does poorly when data poor



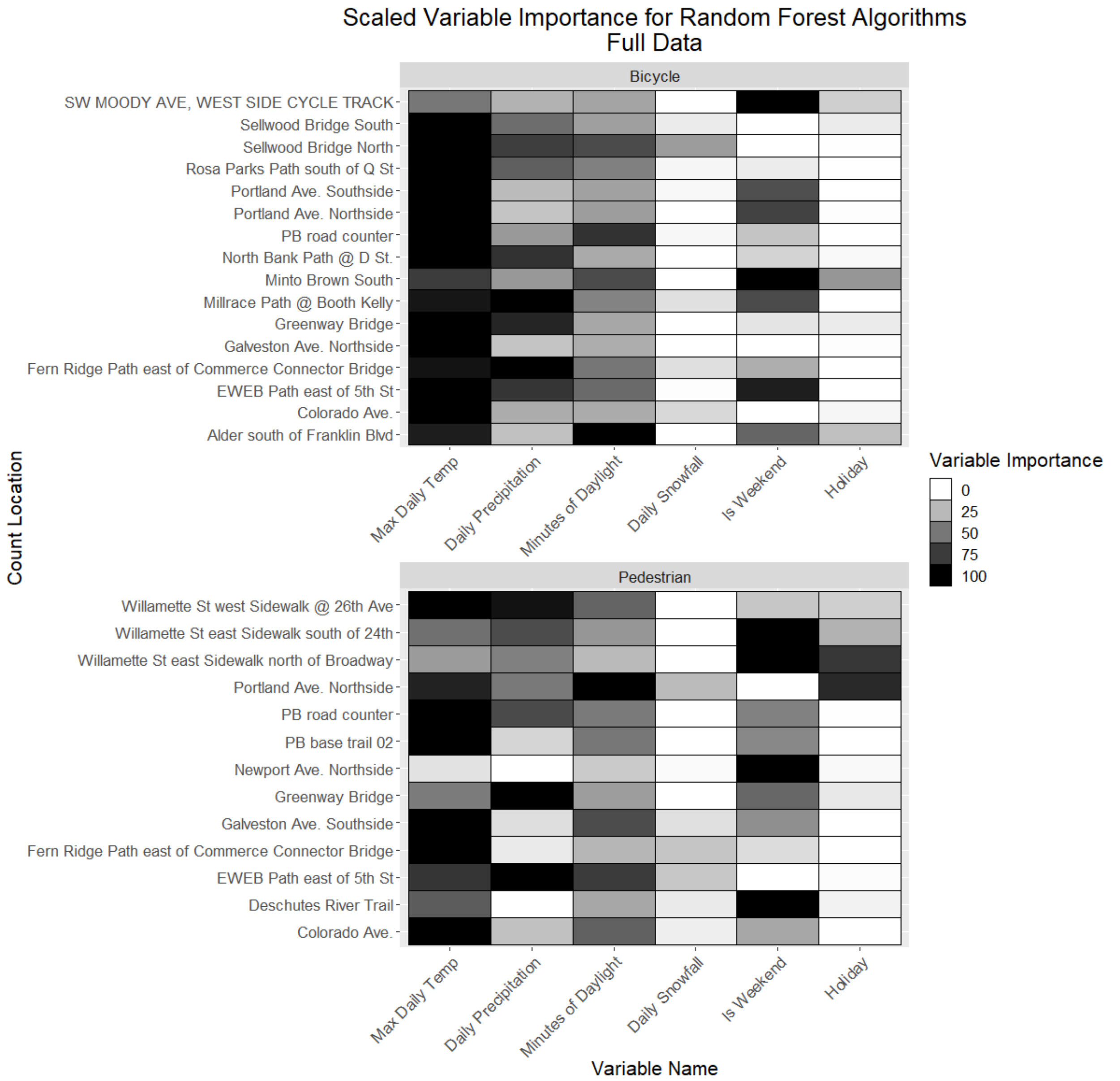
Variable Importance

What Variables Are Important in ML Algorithm?

- Inference generally a limitation of ML
- But variable importance can be calculated (at computational cost)
- Measure of node purity

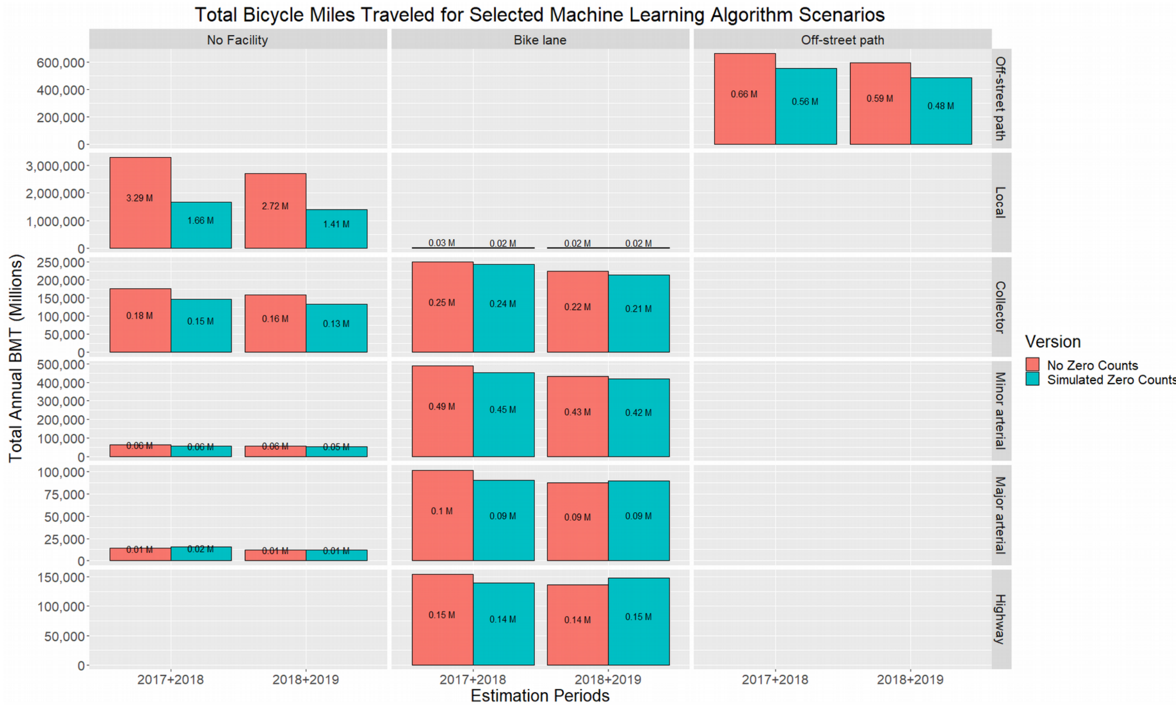
Variable Importance Results

- Temperature importance in all models
- Precipitation and daylight next most important



Bicycle AADT Model Results

- Handling Lack of Zero Counts
  - Random selection of streets high likelihood of zero bike traffic
  - Criteria: local street; low population; density; low centrality, no Strava, no bike facility
- Results
  - Moderates volume well in expected areas
  - Decreases overall BMT by about 1/3



Estimation Periods	Total Annual Bicycle Miles Traveled		Percent Difference
	No Zero Counts	Simulated Zero Counts	
2017+2018	5,225,730	3,385,390	65%
2018+2019	4,444,592	2,985,239	67%