Implausible Ignorability:

An Analysis of Factors Influencing Probe Vehicle Data Completeness

KRISTIAN HENRICKSON AND YINHAI WANG

UNIVERSITY OF WASHINGTON

NATMEC, MAY 2016

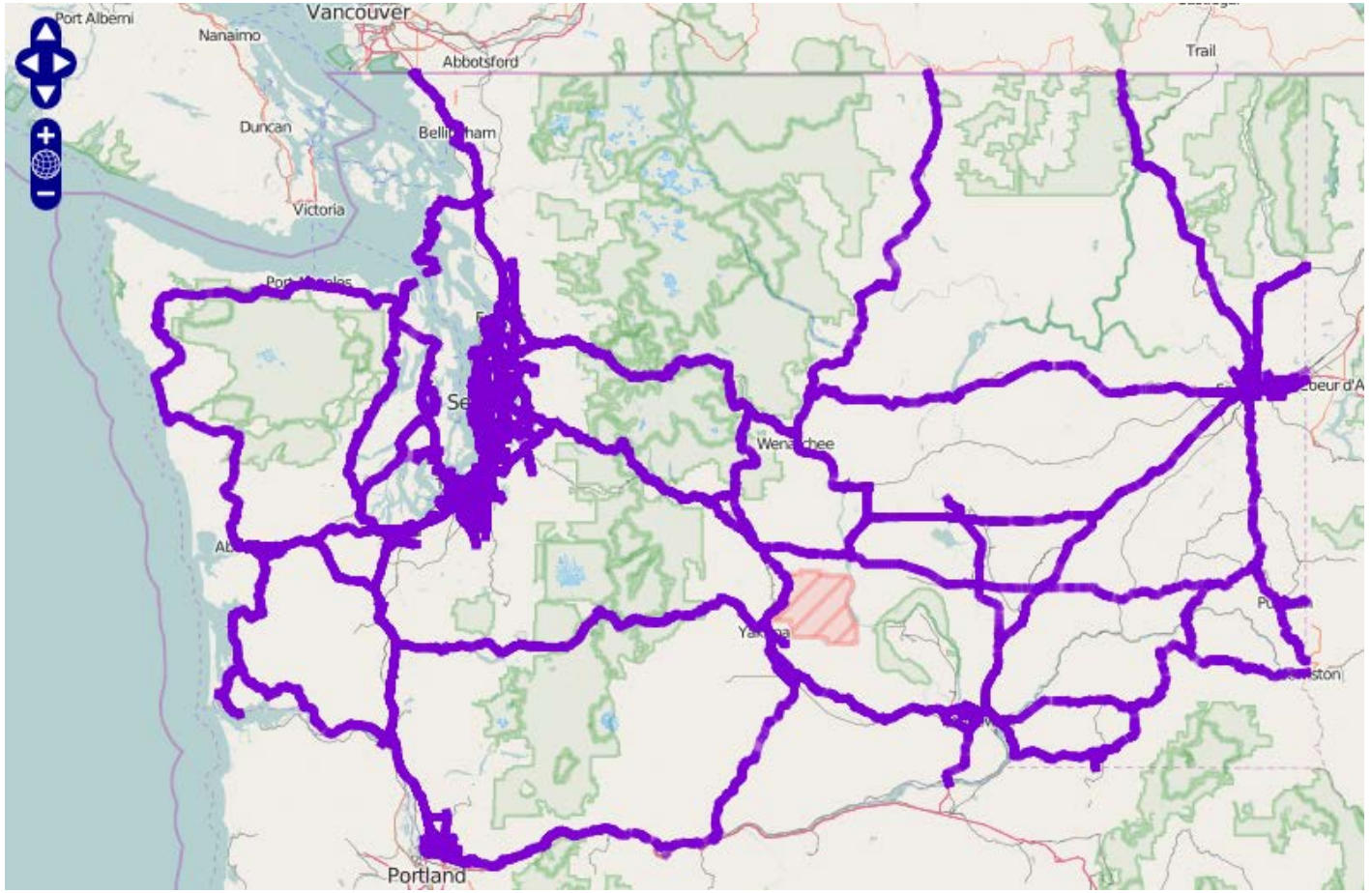
Background: Probe Vehicle Data

Road Link Level Speed and travel time from mobile GPS

Examples: INRIX, NPMRDS (HERE)

NPMRDS:

- Less preprocessing, more transparency
- More obvious quality issues
- BUT: we have more control over quality control and imputation



Problem Statement

Probe vehicle data is often quite sparse, i.e. many missing observations

Missingness is largely driven by the presence or absence of contributing vehicles during a time interval

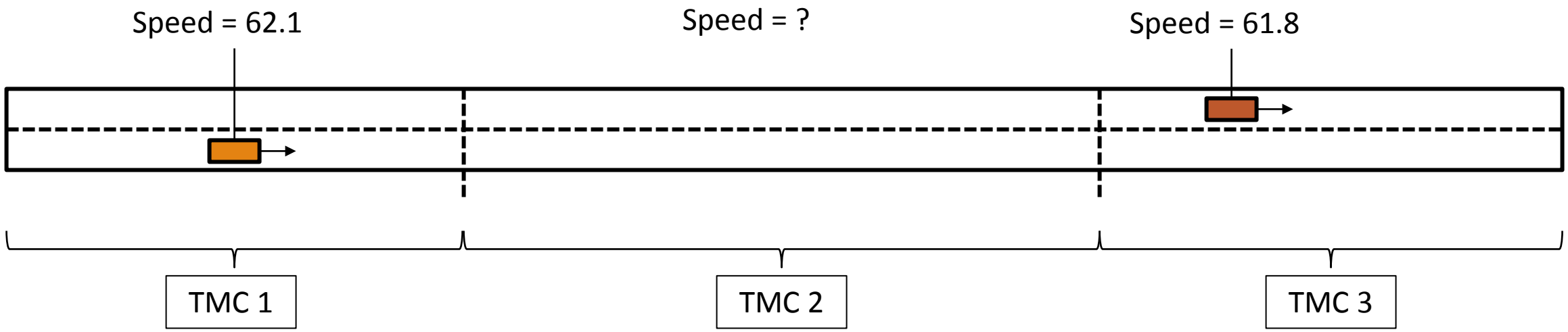
Because the probability of missingness is related to the quantity we are measuring, the observation pattern is not strictly random

Imputing missing values (and quantifying uncertainty) will require some method of describing the difference between observations likely to be missing and those likely to be observed

A Brief Discussion of Missingness

The probability of an observation being missing is related to both traffic volume and travel time

In some cases, the fact that many observations are missing provides all the information we need → Free flow conditions

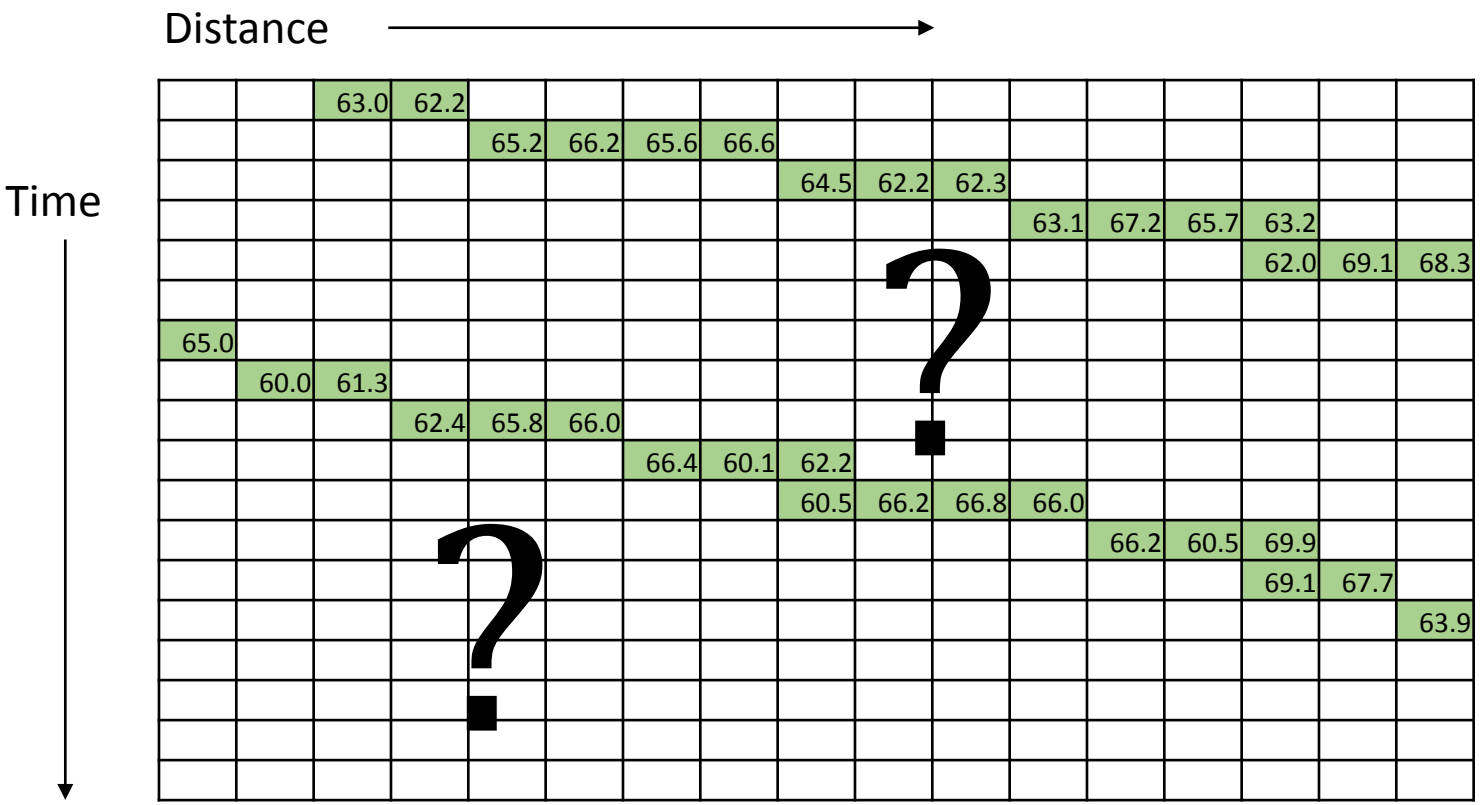


A Brief Discussion of Missingness

Is there a problem here?

Maybe not:

- Consistent and observable vehicle traces
- Reasonable values which are consistent with freeflow conditions
- Data is missing only where we expect few vehicles

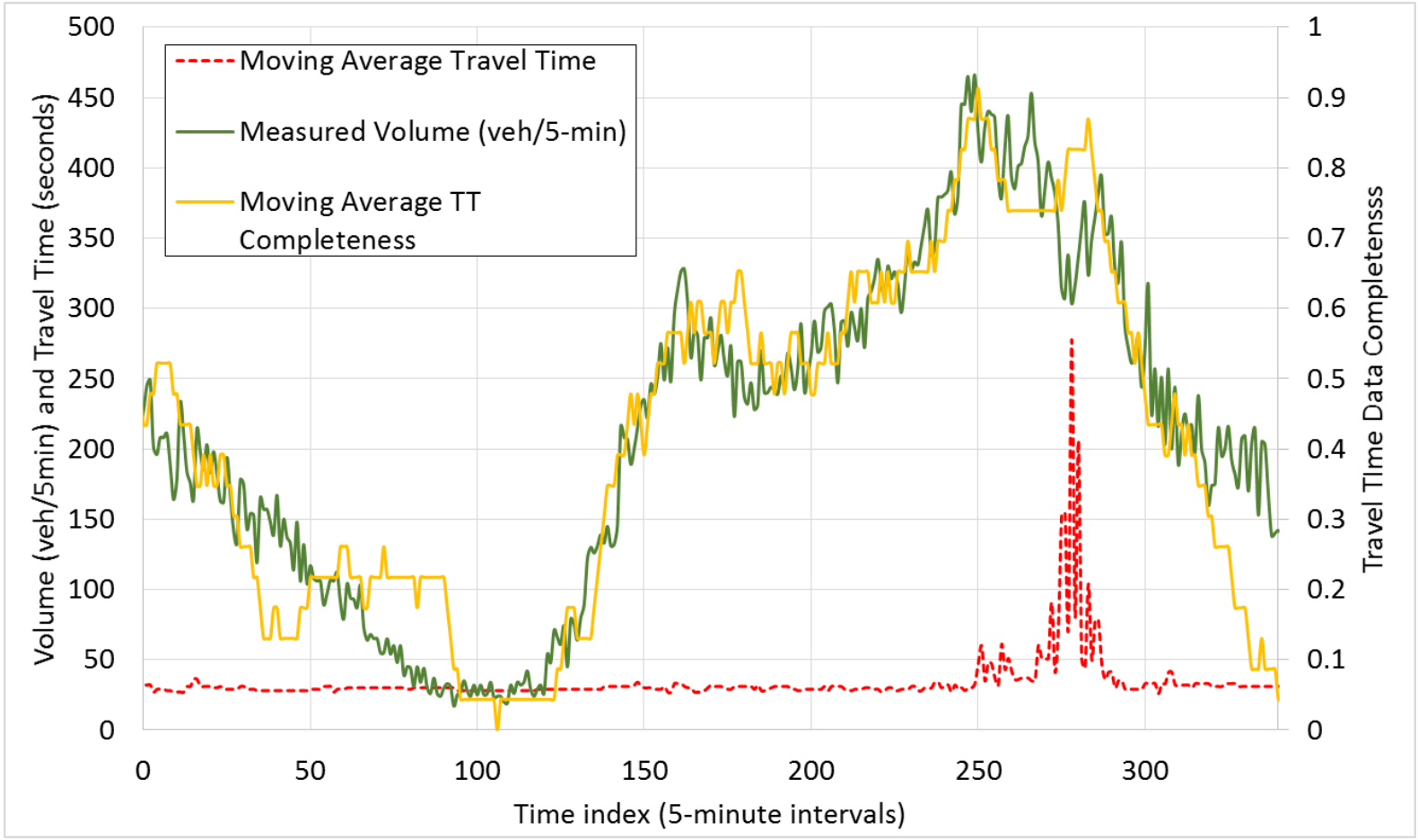


When does this reasoning break down?

A Brief Discussion of Missingness

There is a clear relationship between missing patterns, volume, and travel time…

BUT in general, missing ≠ freeflow conditions



A Brief Discussion of Missingness

Missing patterns vary with traffic volume, time period, number of lanes, travel speed, and road segment length

Thus, missing data might arise from:

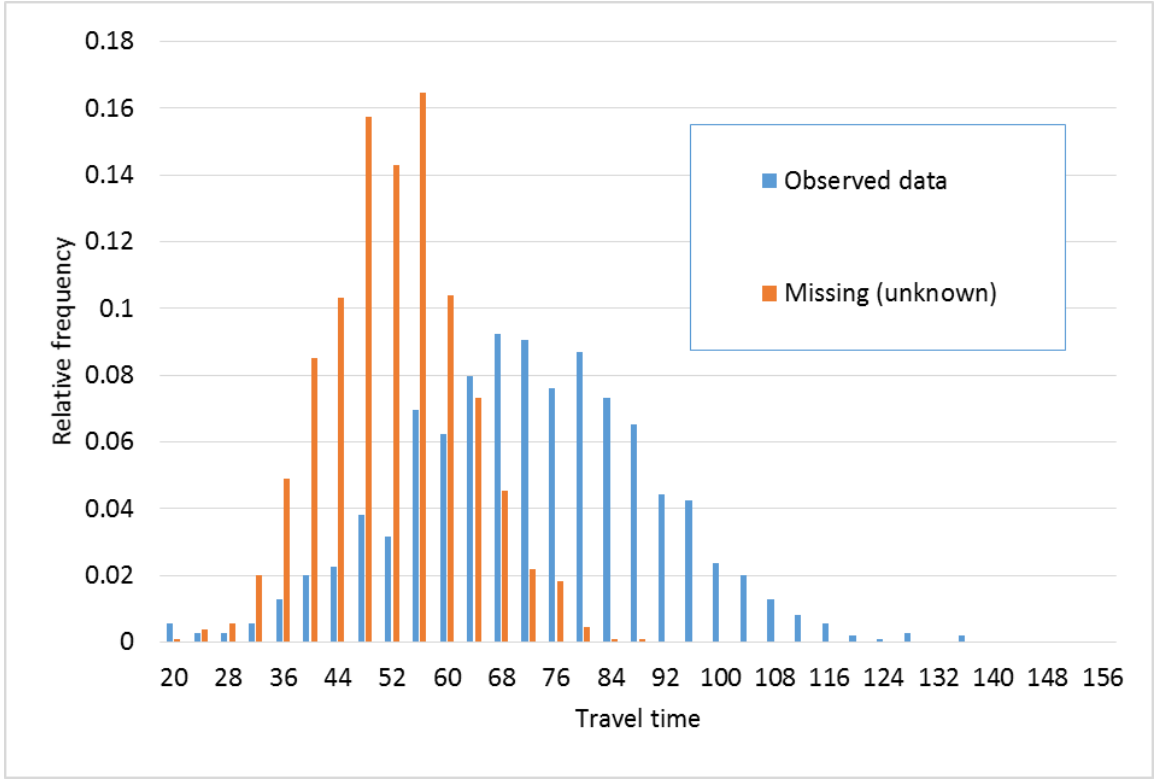
- Fast, low traffic volume
- A Short segment with few lanes
- Random variation (unlucky?)
- Fewer drivers using location services (temporal variation)
- Some combination…

The Real Problem

We know that there will be significant statistical differences between the travel time distribution of the observed vs. missing data

If these differences can be described by temporally and spatially proximate observations, problem solved!

Maybe not possible with nearby travel time alone?



Our Approach

Develop a modeling methodology to describe the relationship between data completeness and the quantities of interest: volume and travel time

Show how different combinations of traffic state, time period, and segment characteristics can contribute to data completeness

Why?

- Describe how missing data can lead to biased analysis results
- Communicate the importance of principled missing data treatments
- As a first step toward developing a robust imputation scheme

Dataset

The focus of this analysis is on the National Performance Management Research Data Set (NPMRDS)

Five road segments from I-5 and I-405 in the Seattle area:

- 0.47 - 1.7 miles in length
- 21 days of data from September 2013
- 5-minute time intervals, 6048 observations per study site
- Varying rates of missing data
- Loop detector (volume) data was obtained for the same locations and time period from the Washington State Department of Transportation

Methodology

A missing observation effectively means:

Zero contributing vehicles on the road segment at the start of the time interval

Zero contributing vehicles arriving during the time interval

Thus, we can represent the number of contributing vehicles for one time period as the sum of two Poisson random variables

Methodology

Probability of n contributing vehicles for observation interval of t, arrival rate of λ, and expected travel time of $E(TT)$:

$$\Pr(n) = \frac{(\lambda(t + E(TT)))^n}{n!} \exp\left[-\lambda\left(t + E(TT)\right)\right]$$

The problem: true count is only known if it is zero, otherwise it is only known that $n \geq 1$. How to build a regression model?

Solution: censored Poisson regression, represents $\Pr(n \geq 1)$ as $1 - \Pr(n = 0)$ in the log likelihood expression

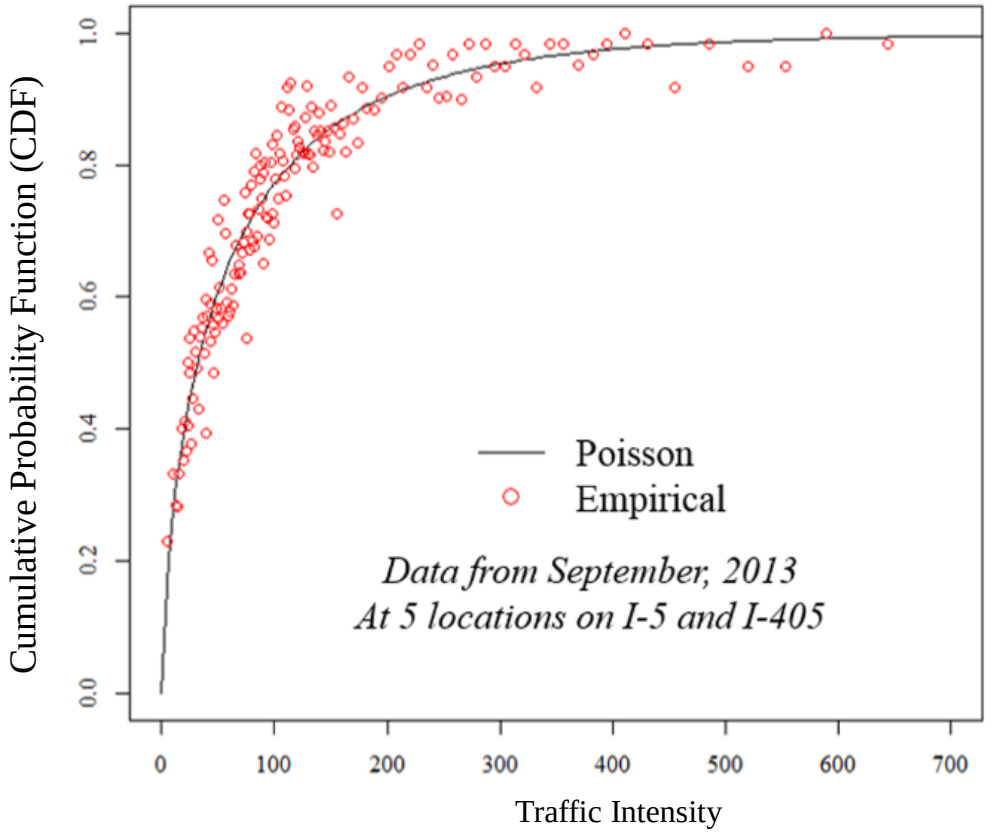
Validation

Validate a methodology that predicts an unobservable quantity?

Right: empirical and predicted CDFs

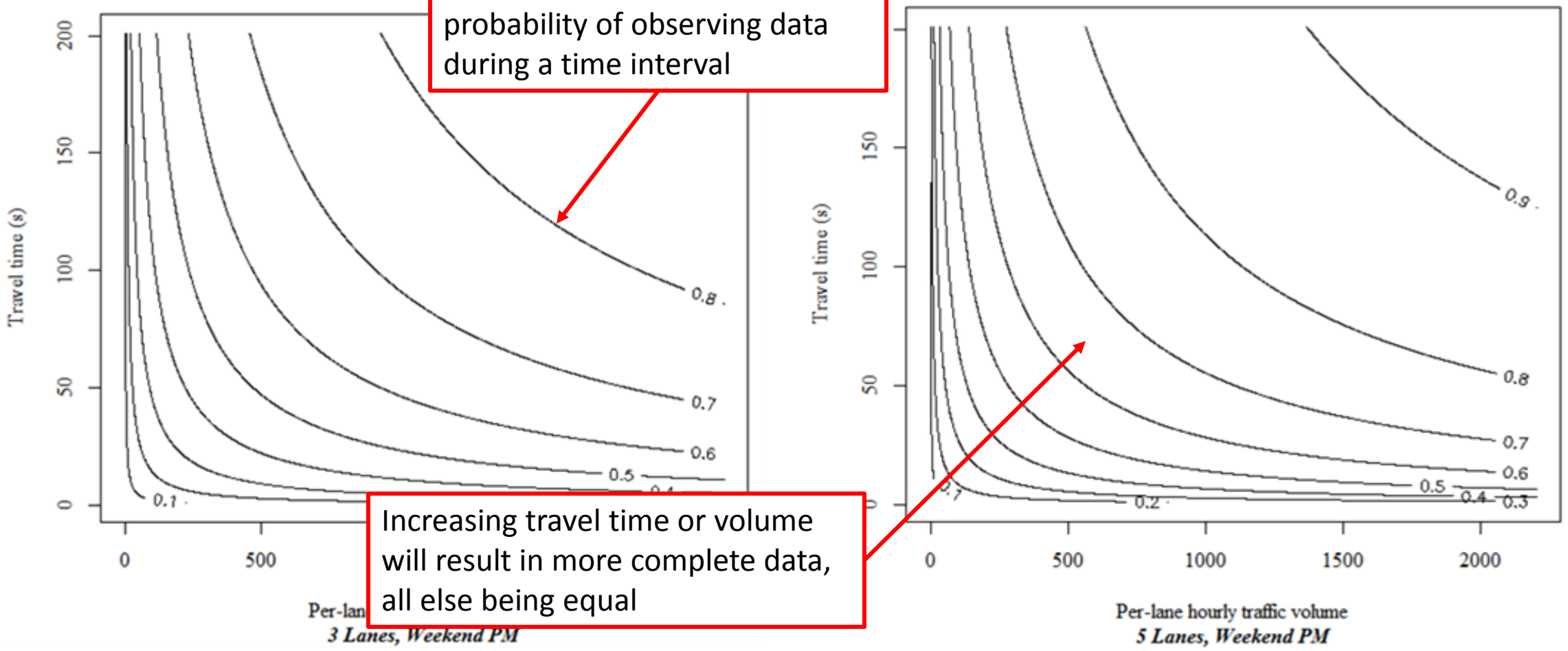
Below: empirical and predicted missing rates for observations preceded by missing interval

	Missing rate				
	Overall		Missing i - 1		
Time period	Emp.	Pois.	Emp.	Pois.	Volume fraction
Weekday AM	0.3421	0.3418	0.6681	0.6687	0.0053
Weekday PM	0.2608	0.2607	0.4604	0.4416	0.0031
Weekend AM	0.5770	0.5766	0.7200	0.7184	0.0031
Weekend PM	0.2899	0.2899	0.4277	0.3966	0.0028



Results

Contour lines of constant probability of observing data during a time interval



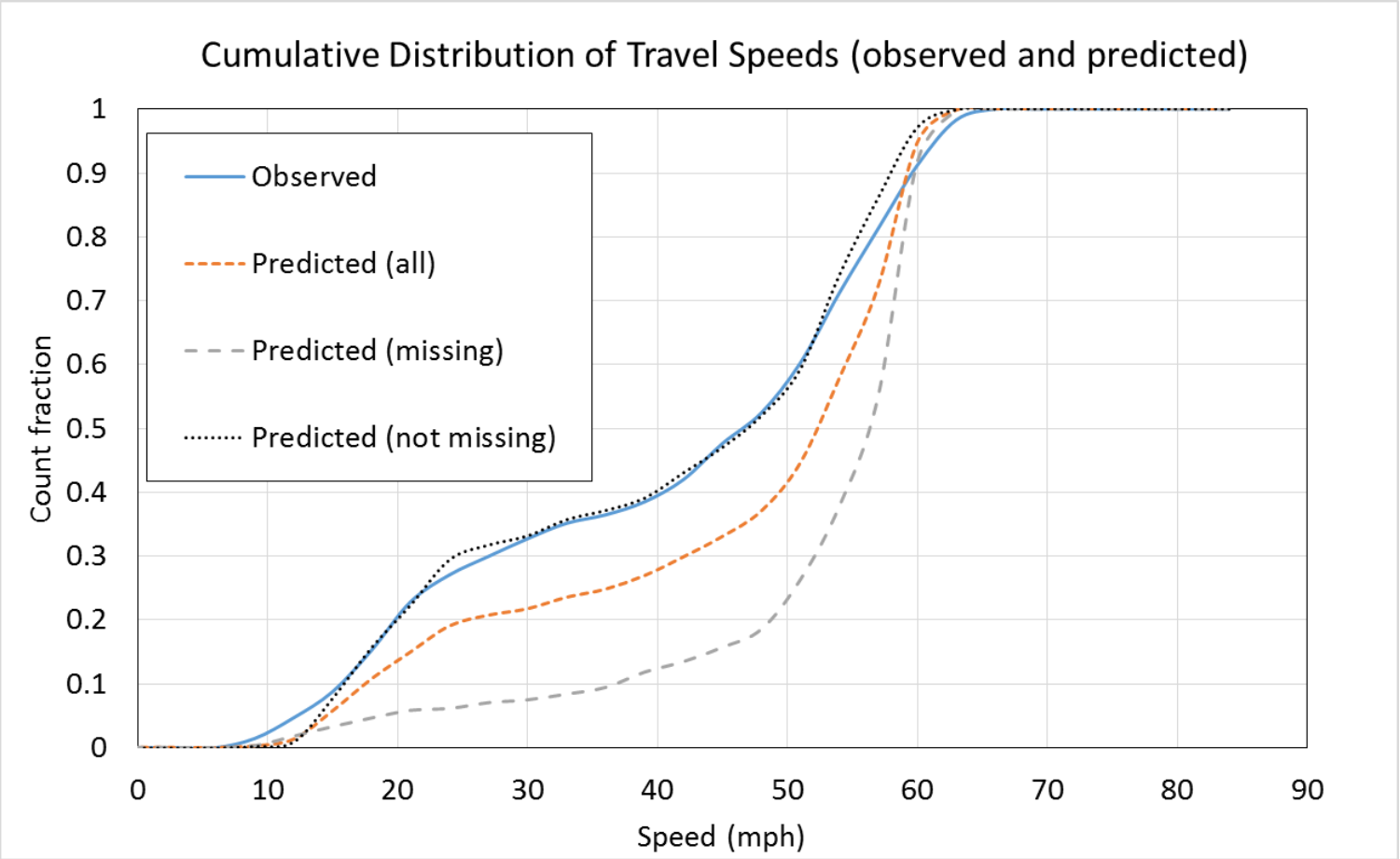
Increasing travel time or volume will result in more complete data, all else being equal

Next Steps

Initial findings on imputation:

Typical imputation schemes assuming “missing at random” do no adequately address bias

“Missingness as a predictor” helps in some cases



Conclusions

Missing patterns are very closely related to both travel time and vehicle volume, which could have significant impacts on any statistical analysis based on this (or similar) data set.

Probe vehicle data overcomes many of the shortcomings of fixed mechanical sensing but, like fixed sensors, it comes with a unique set considerations for data quality and uncertainty.

There is a need to develop an imputation scheme that explicitly considers the relationship between missingness and the quantities of interest

Questions?

Washington State
Department of Transportation