Statistical Issues Related to Evaluating the Quality of Traveler Information

James Richardson
PhD Student in Civil Engineering, University of Virginia

Agenda

- Introduction
  - Motivation
- Sampling Theory
  - Sampling, Confidence Intervals, Minimum Sample Size
- Research from Houston Toll Tag data
  - Findings on sample sizes
  - Spatial, Temporal distribution of travel time variability
- Future Work
- Questions and Comments

Introduction

- TPF-5(200): Standard Test Procedure for Travel Time Data Quality Assessment
  - University of Virginia, Virginia Transportation Research Council, Texas Transportation Institute
- Goals of Research
  - Develop guidelines for evaluating traveler information services
    - Fair, statistically defensible methods
    - Recommend sample sizes for ground truth
    - Suggest where and when to sample in a network
- How far along are we?
  - Currently focused on establishing guidelines for freeway data
  - Consulting with NATWG
  - Draft of "standard" in the works

Motivation

- What is "ground truth"?
  - The "true" mean travel time of some segment at a specified time?
  - Or an estimate of the mean travel time?
- We usually don't know with 100% certainty the "true" mean travel time
  - This is a population parameter
  - In statistics we differentiate between a "population" and a "sample"
- We can estimate a population parameter using statistical inference from sample data
- Our confidence in this estimate is a function of the sample size and the variance in the observed data

Travel Time is Stochastic

- While there is a deterministic component to travel time (e.g. density v. speed), the realization of individual travel times is largely stochastic
  - Different types of drivers
  - Different types of vehicles
  - Weather, grade, other factors
- Travel Time is a random variable
  - Has some unknown distribution
  - Has an unknown mean and variance
- How we define the population is important
  - Space (e.g. TMC segments vs. corridors) and Time (e.g. 5 minutes vs. 1 hour)

Sampling Theory

- Population Parameters can be estimated from sample data
  - The mean travel time for a given population (space, time) can be estimated from sample observations
  - The <u>empirical sample mean</u> is our best estimate of the population mean
    - This statistic is also random and has a distribution
    - We can estimate the distribution of the sample mean
      - If we know the population variance we can use a standard normal distribution
      - But we generally don't know (or don't want to assume) the population variance
        - Use sample variance and a Student's T distribution

Data Quality and Accuracy

- Data quality is a broad concept but we focus here largely on accuracy of data
  - Accuracy is a measure of the distance of an estimate from some “true” value
- The accuracy of a travel time estimate is a measure of the distance of the estimate from the mean travel time of the population
  - Generally we don’t know the mean travel time of the population
  - We can estimate it by sampling
  - But there is still uncertainty in our estimate
- So to measure the accuracy of a service provider’s data requires that we have some confidence in our estimate of the ground truth

More on Accuracy

- We want to know more than whether or not a single estimate was accurate
  - Knowing the accuracy of a single segment is useful but doesn't tell the whole story
  - We also want to know how accurate estimates are for the rest of the network
  - It is difficult and costly to collect ground truth data for every segment in a network
- Also we need to consider time of day
  - Do we collect ground truth 24 hours / day x 7 days / week?
  - Is it important to know the accuracy of data in the middle of the night on a weekend?

Two Levels of Sampling

- So we need some way to measure ground truth for a segment in a network
  - We can use sample data from the traffic stream to estimate the mean travel time of the population
  - How many samples do we need?
- But we also need to measure the accuracy of a service provider's data across time and space
  - So we need to sample particular segments from the network during critical time periods
  - Which segments do we sample and when?
- To summarize
  - 1. Sample traffic stream to establish ground truth
  - 2. Sample critical segments and time periods from the network to establish accuracy

"Measuring" Ground Truth

- Two Basic Methods
  - Floating Car
    - How do we know how close this observation is to the population mean?
      - Statistical theory can't really help here because we don't know much about the variance of the observation
    - Floating Car confidence interval?
  - Re-identification
    - Can be used to make multiple observations.
    - Generally non-intrusive sensors
    - Can develop a statistical confidence interval

Terminology

- Sample Mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Sample Standard Deviation

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Coefficient of Variation

$$CV = \frac{s}{\bar{X}}$$

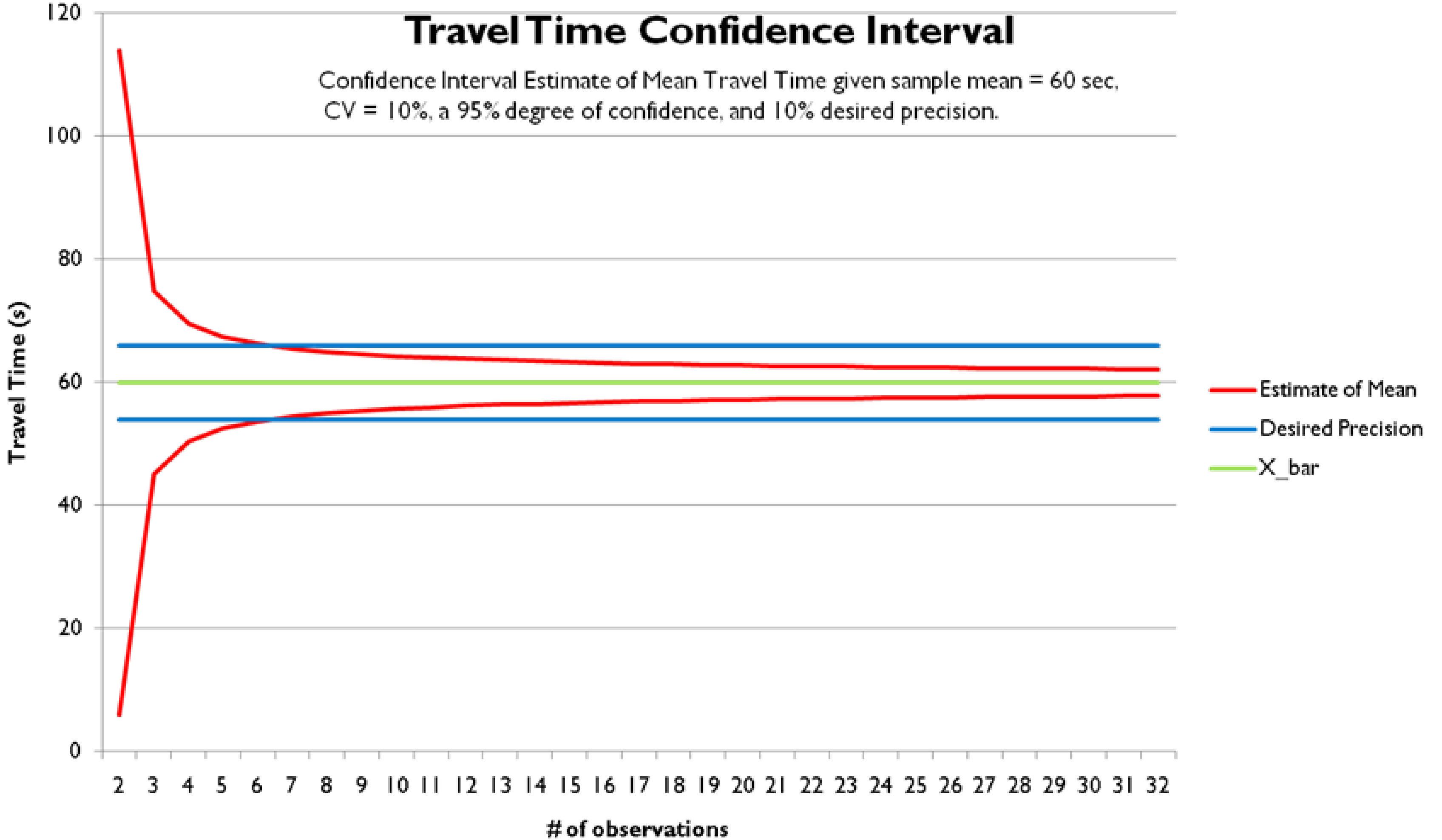
- Precision of Estimate
  - How close we want the estimate of the mean to be to the population mean (e.g. 10% allowable error).
- Degree of Confidence
  - Level of “alpha” or significance level.
- Student’s T Distribution
  - Sample mean is distributed following a Student’s T distribution when the population variance is unknown.

Confidence Interval Example

- Let's assume we collect a sample of observations from a traffic stream over a 1 mile long segment
  - We observe a mean travel time = 60 seconds and a standard deviation of 6 seconds (i.e. CV = 10%)
  - We can develop a confidence interval that the "true" population mean was equal to 60 seconds
  - As "n", the sample size increases the width of the confidence interval decreases
  - "t" is also sensitive to sample size. Larger sample sizes result in smaller "t" statistics

$$\mu = \bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Confidence Interval – Travel Time

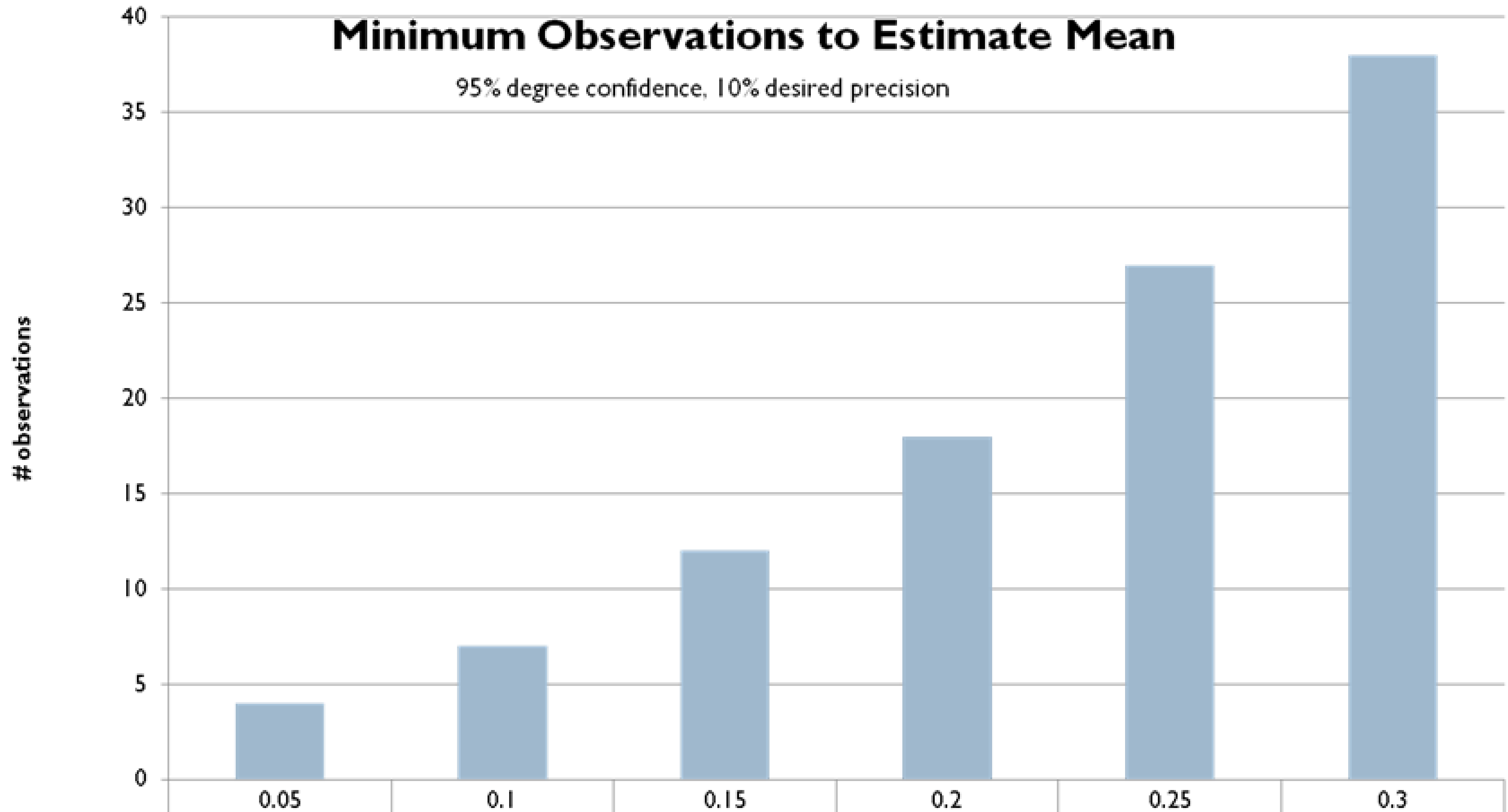


Minimum Sample Size - Travel Time

- The previous slide shows that as the sample size increases the bounds of the estimate converge on the sample mean
- The equation for a confidence interval can be manipulated to derive an equation to determine the minimum sample size
  - CV = Coefficient of Variation
  - t_alpha = Student's T Statistic
  - e = desired precision (percentage)

$$n = \left(\frac{t_\alpha * CV}{e}\right)^2$$

Estimating Mean Travel Time

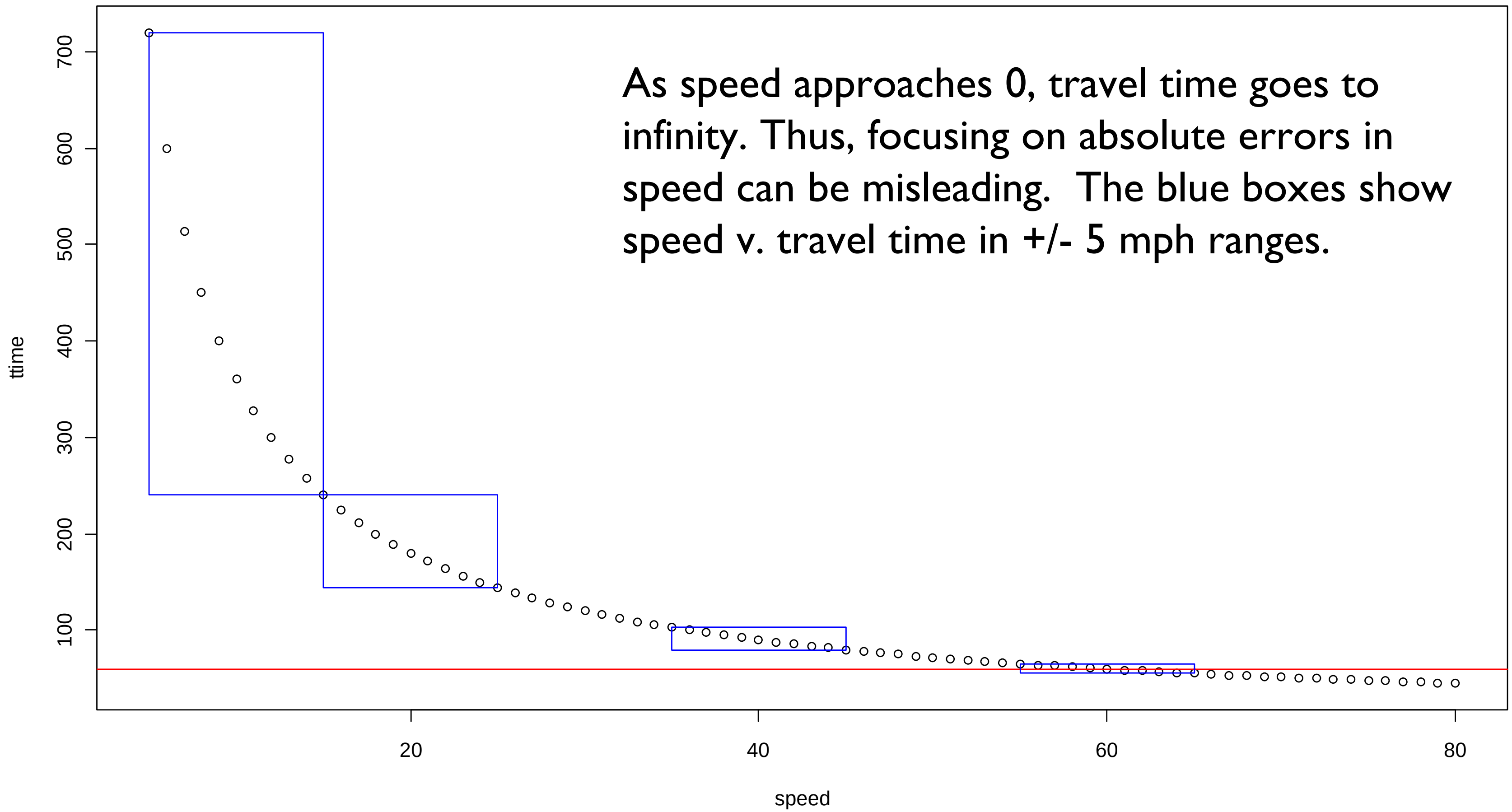


	0.05	0.1	0.15	0.2	0.25	0.3
Travel Time	4	7	12	18	27	38

Travel Time and Speed

- Travel Time and Speed are inversely related
  - TT = dist / SMS
- Space Mean Speed <> Time Mean Speed
  - The arithmetic mean of speed observations = Time Mean Speed
  - The harmonic mean of speed observations = Space Mean Speed
- Generally we want to know space mean speed
  - We can get this by estimating mean travel time
    - SMS = dist / TT
  - Be careful about using arithmetic mean speeds

Another Look at Travel Time and Speed



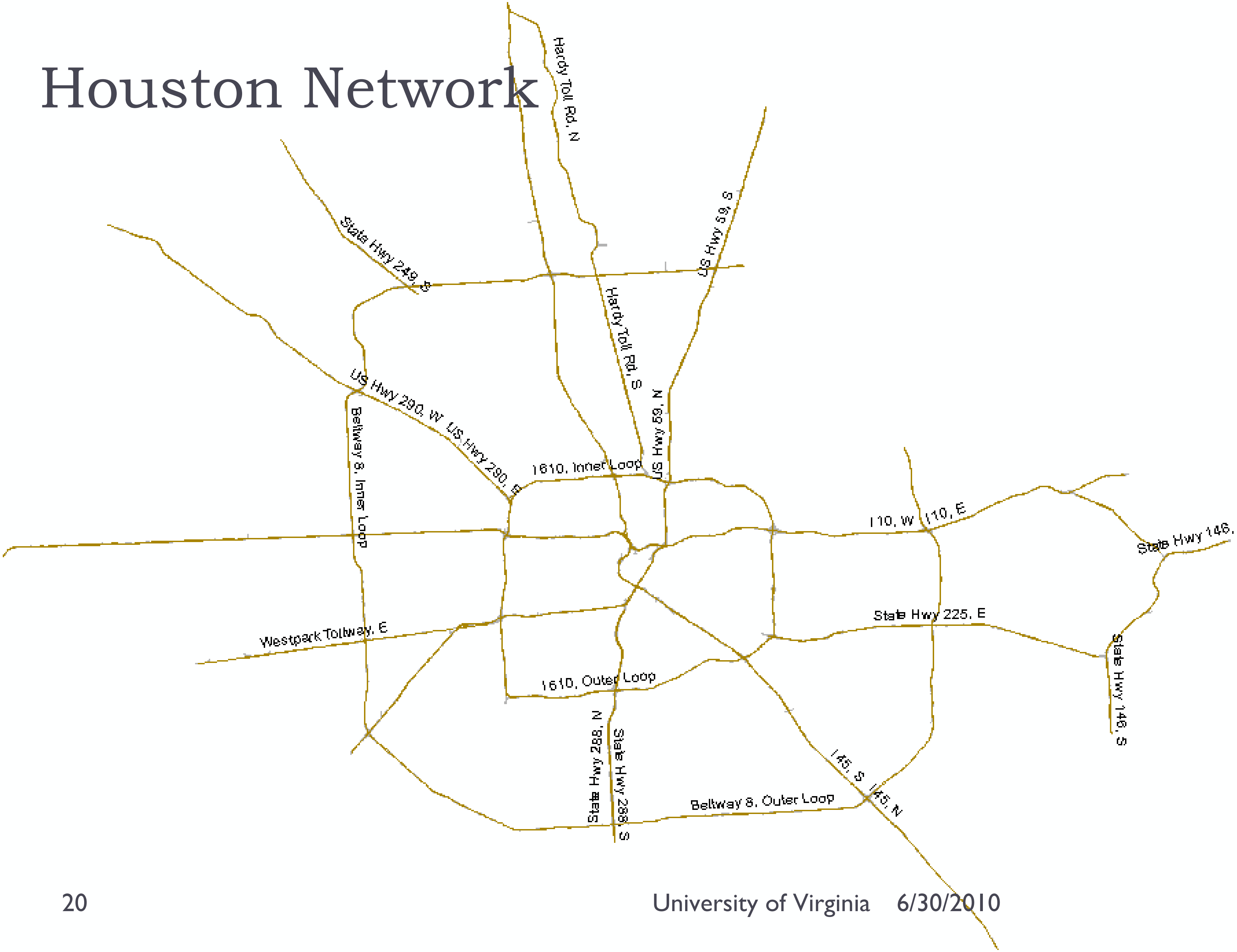
Ground Truth Sampling Summary

- We have seen that the population mean of a random variable can be estimated from a sample
  - The precision of the estimate is sensitive to variance and sample size
- The Coefficient of Variation of Travel Time is a good measure of relative variation
  - Can be used to establish minimum sample sizes
- Travel time and speed are inversely related
  - The space mean speed is the inverse of the arithmetic mean of travel time
  - Small absolute errors in speed can translate into relatively large absolute errors in travel time
- Determining which segments in the network to sample is important in order to comprehensively measure the accuracy of a data source

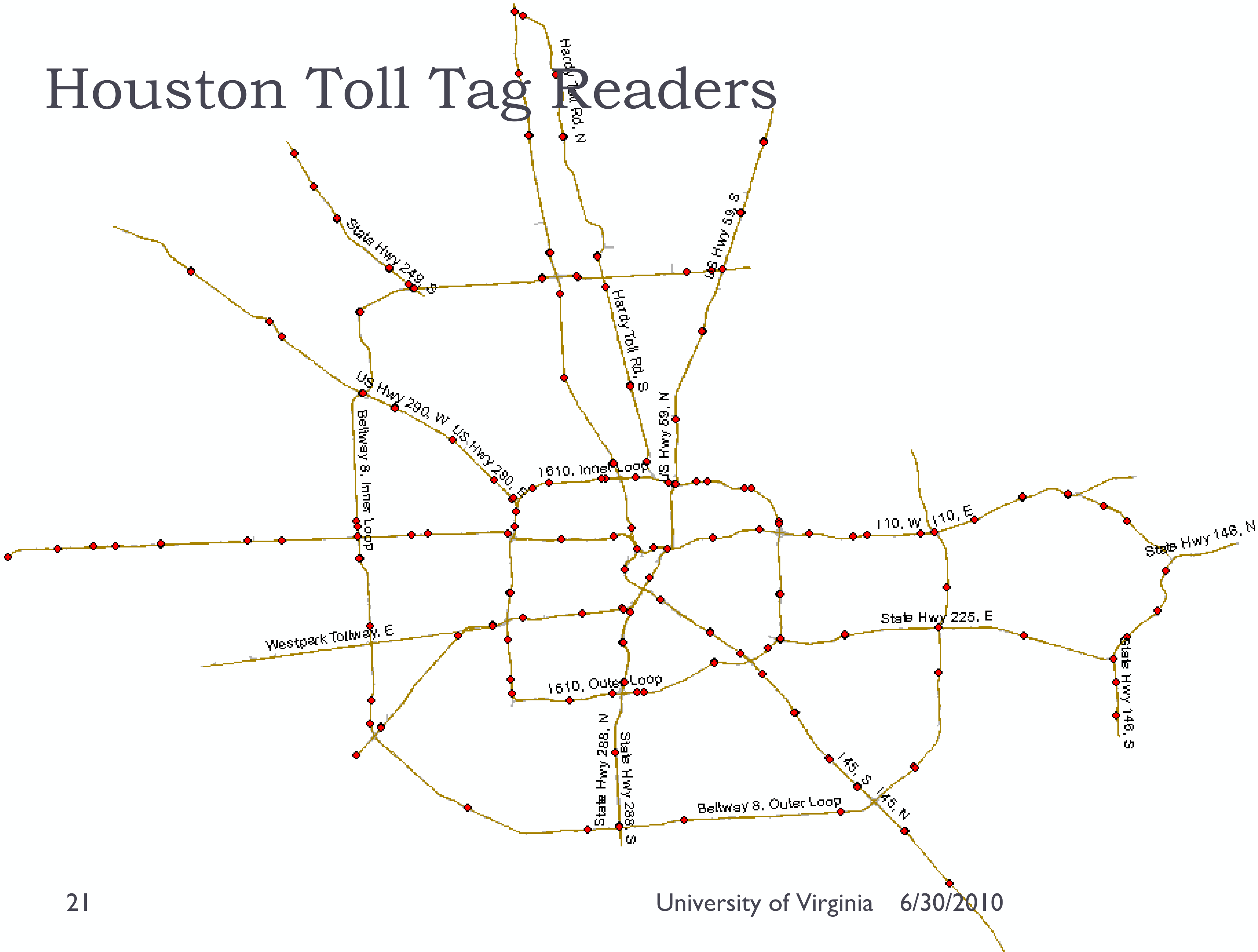
Empirical Data from Houston

- Houston TranStar network
  - Freeway network monitored by toll tag readers
  - Use position of toll tag readers and anonymous tag data to measure travel time of vehicles
- Approximately one year (2008) of observations loaded into a database
  - 24 hours / day, 7 days / week, over 200 unique segments
    - 273,907,180 unique observations
  - Data aggregated by segment and 5-minute periods
    - 20,952,566 unique spatial / temporal aggregation periods
  - Calculated statistics for each spatial / temporal extent
    - Determined minimum sample size based on Student's t statistic (95% degree of confidence) and a 10% allowable error (e = .1)
    - Calculated mean travel time, standard deviation of travel time, space mean speed

Houston Network



Houston Toll Tag Readers



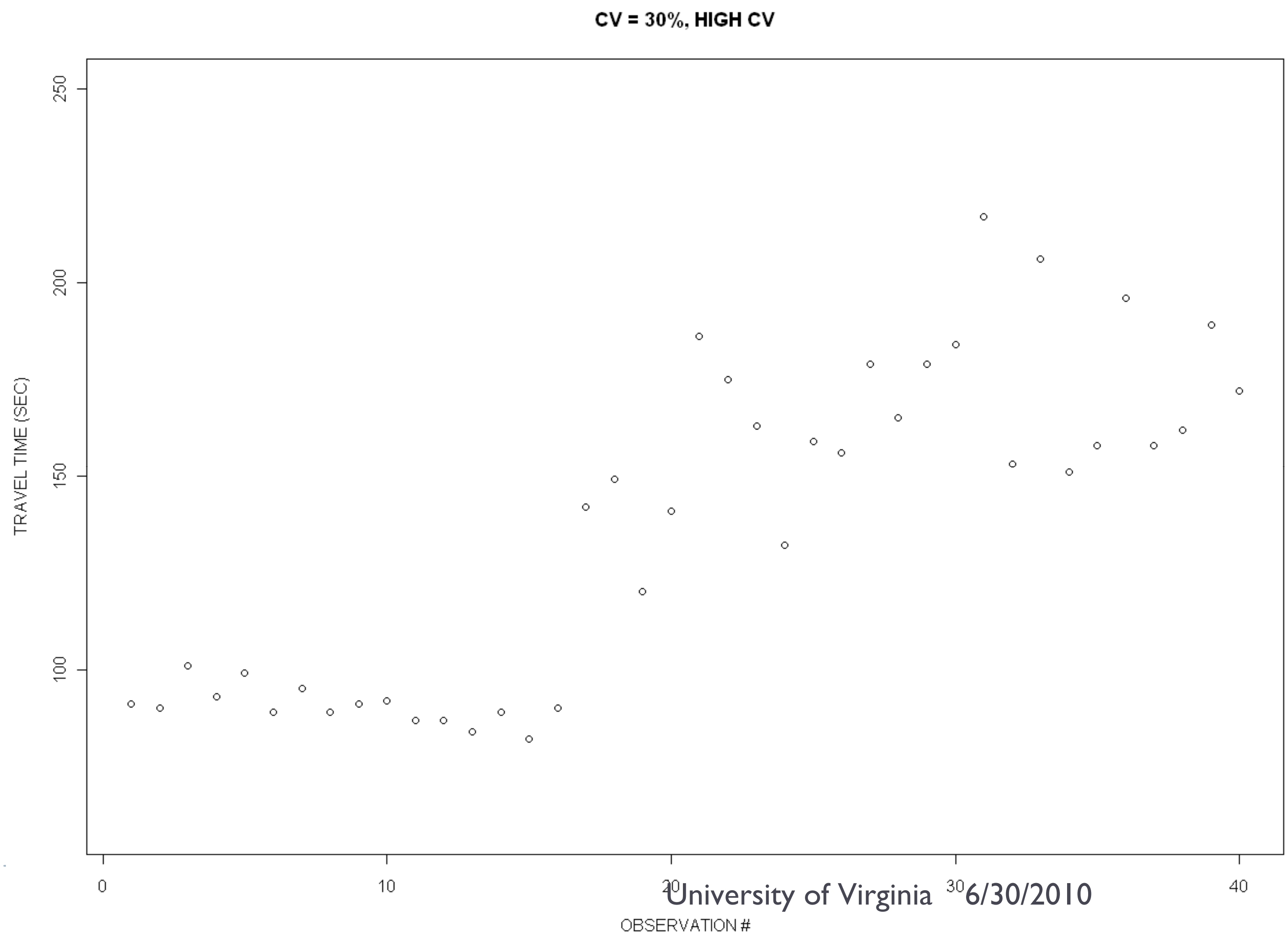
What can we do with this data?

- Determine distribution of travel time variance
  - Spatial distribution
    - Which links in the network have the most / least variance?
  - Temporal distribution
    - During what time periods is variance greatest / smallest?
- Determine sample size thresholds
  - How many samples are needed for a given link at a specified time?
- Develop guidance for data quality assessment methods
  - How can we intelligently choose <u>where</u> and <u>when</u> to sample?
  - How many samples are needed?
  - What are the best technologies to use for different conditions?

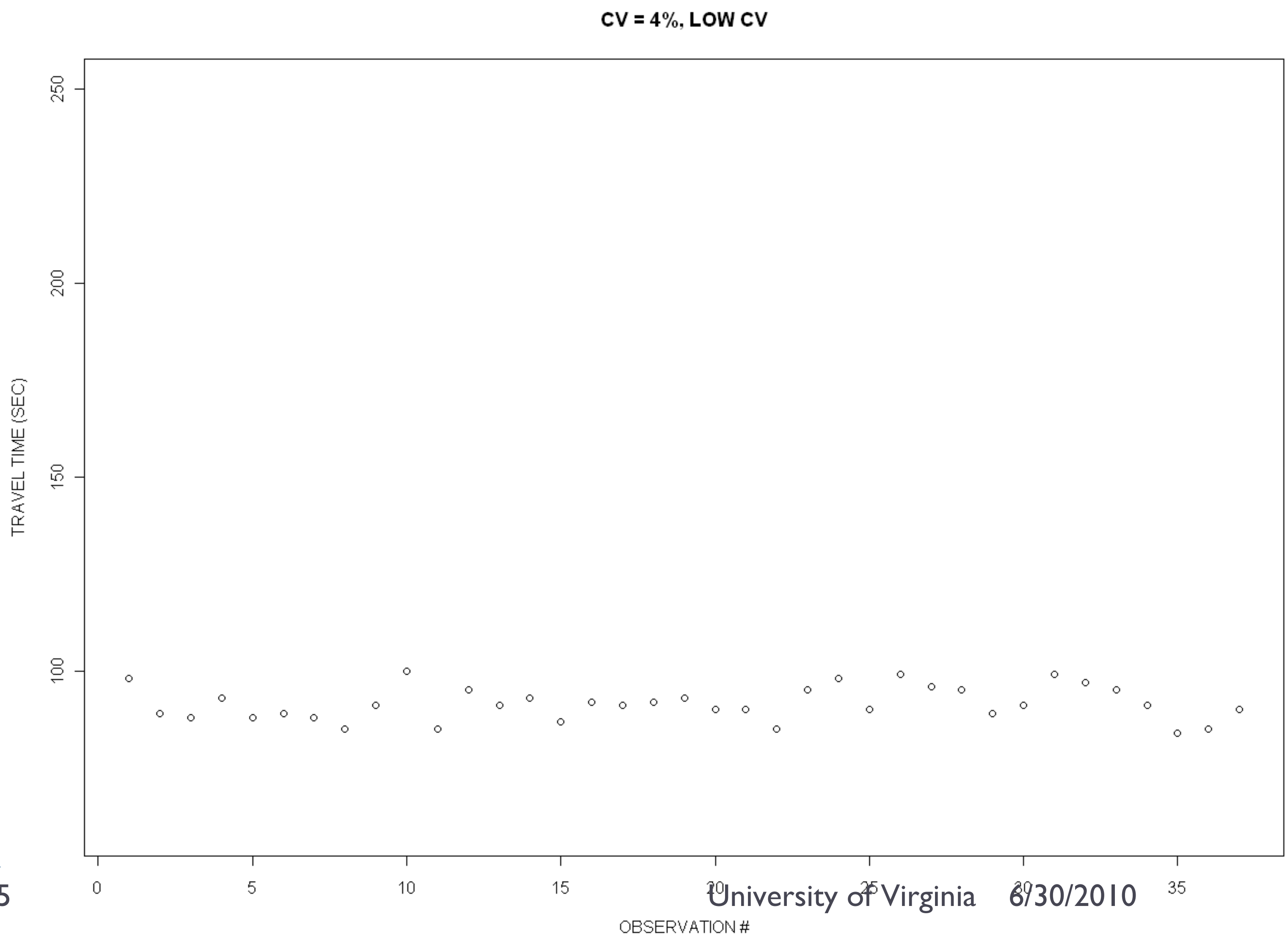
Coefficient of Variation

- The CV was used as a way to measure the relative degree of variation
  - CV Travel Time was selected
  - Only observation periods where the number of samples was sufficient to estimate the mean (95% degree confidence, 10% error) were used
- Distribution of CV in space and time was analyzed

Example of High CV



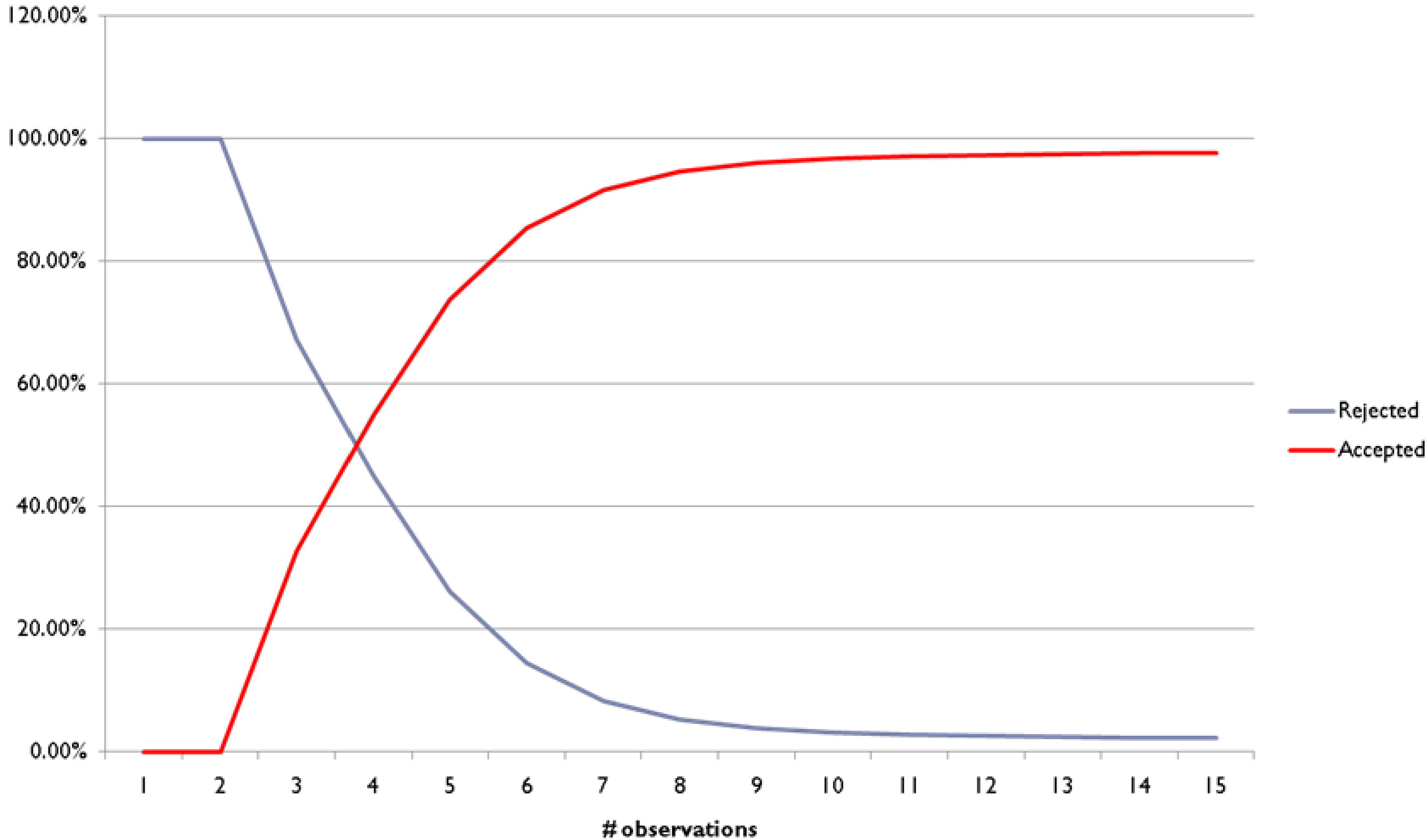
Example of Low CV



How much variance is "a lot"?

- Travel time variance varies in space and time
  - We might have more variation at one link than another
  - We could have more variation in the morning than in the evening
- We can use an empirical cumulative distribution to see how travel time variance is distributed
  - Consider: All segments in Houston during the weekdays AM/PM peak hours (2008 data)
    - 90^{th} percentile CV Travel Time = 10%
    - Interpretation: 90% of the time in Houston, the relative variance in travel times is about 10% or less
    - Only 10% of the time is the relative variance greater than 10%

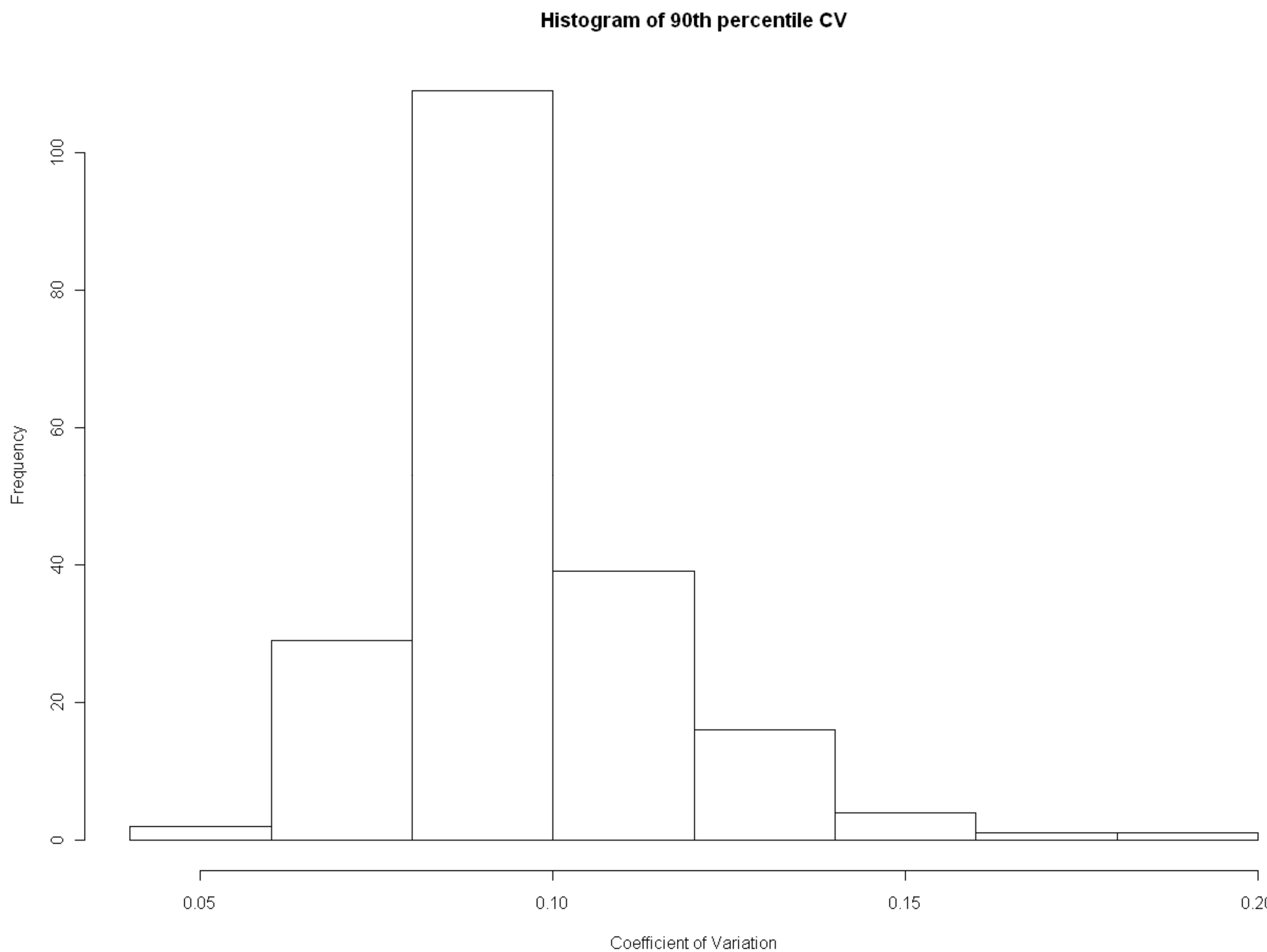
Sample Size and Acceptance Rate

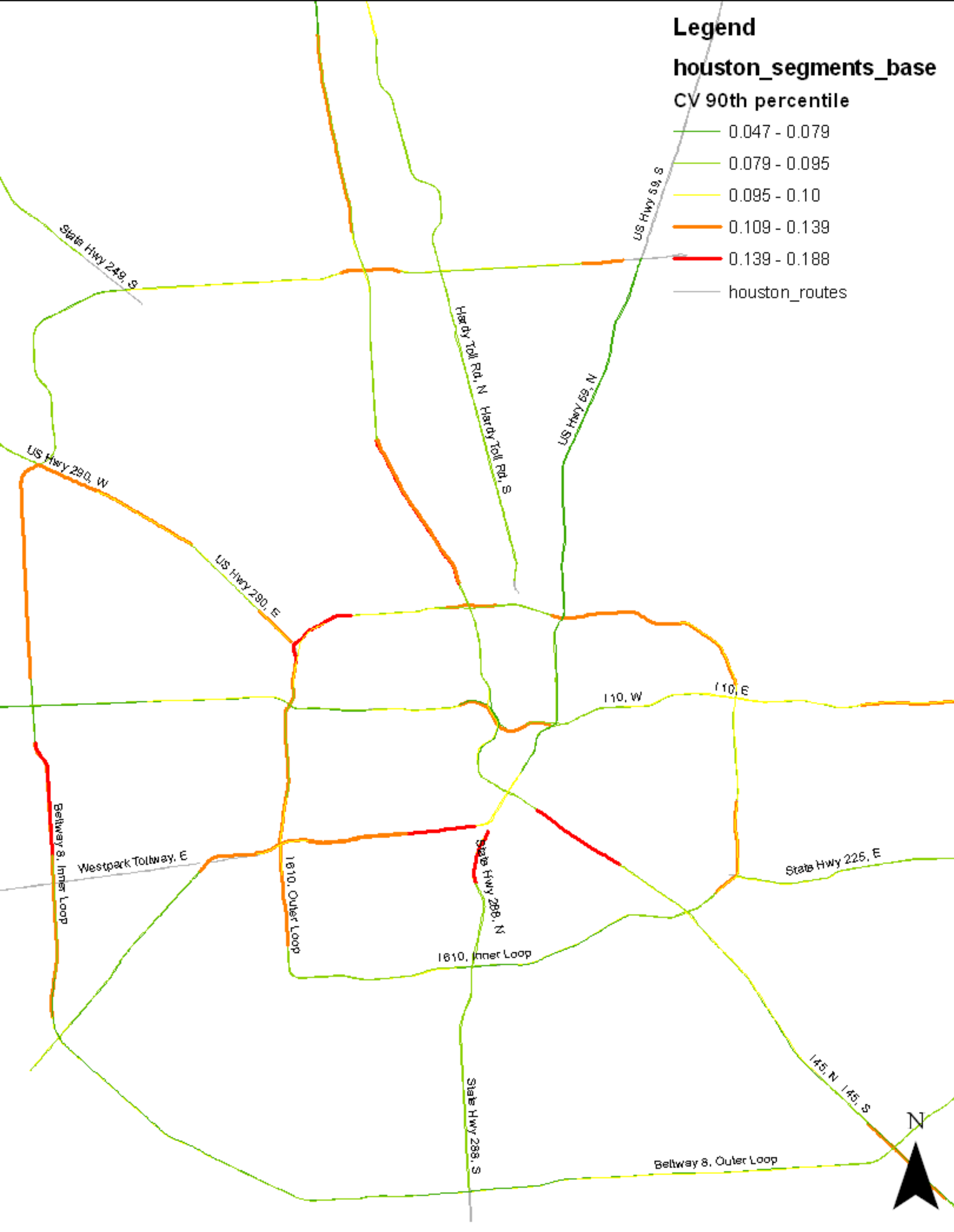


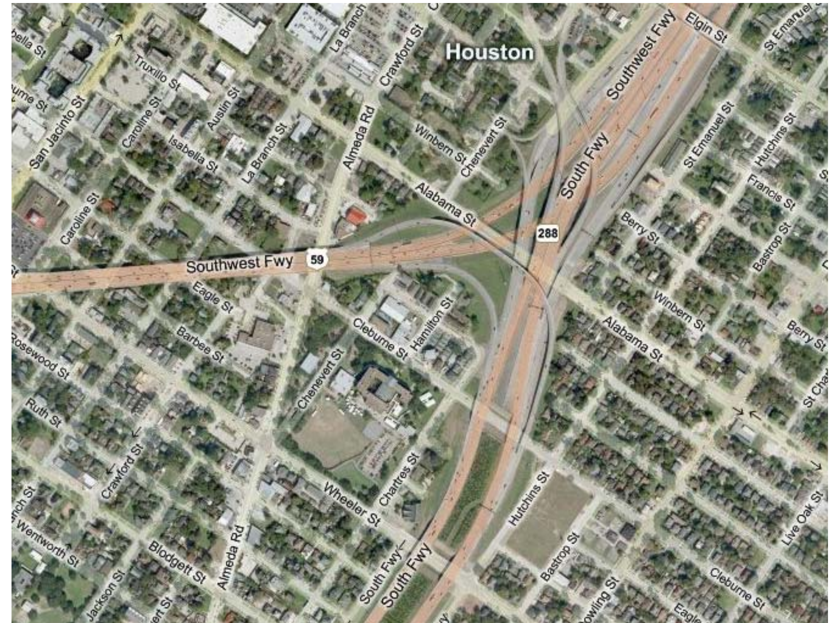
90^{th} Percentile CV

- The 90^{th} Percentile CV in the entire Houston network across all times was about 10%
  - CV = .1 would require 7 observations to estimate the mean with 95% degree confidence and 10% desired precision
- We can also look at how travel time variance is distributed spatially and temporally
  - Where are the segments in the network that have higher CV levels?
  - When do these segments have higher CV?
  - What are the factors that determine travel time variance?

Spatial Distribution of 90th Percentile CV



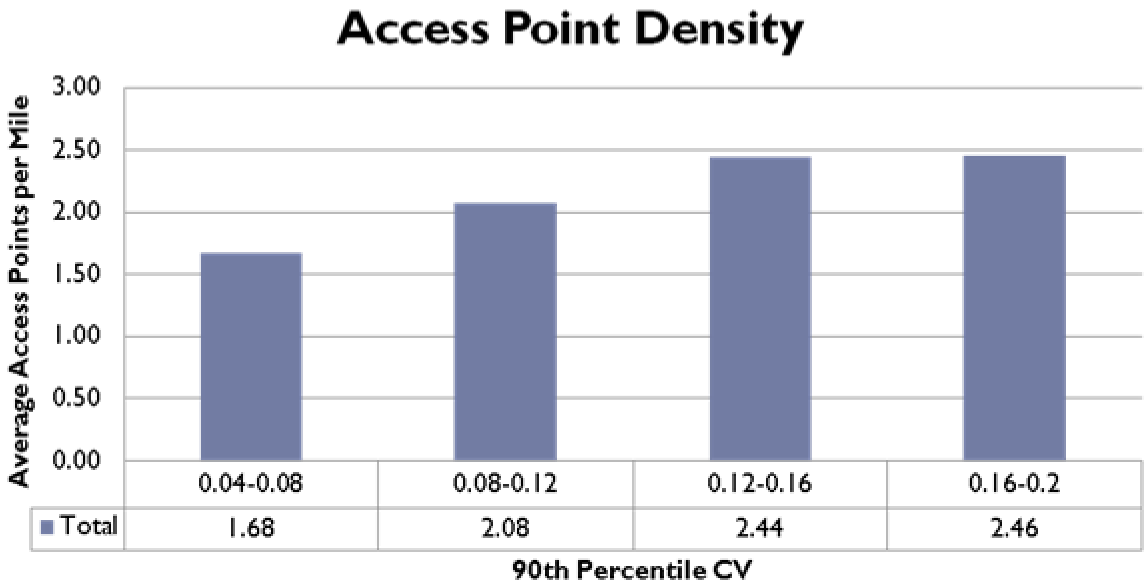

Legend
houston_segments_base
CV 90th percentile
0.047 - 0.079
0.079 - 0.095
0.095 - 0.10
0.109 - 0.139
0.139 - 0.188
houston_routes
State Hwy 249, S
US Hwy 59, S
Hardy Toll Rd, N
Hardy Toll Rd, S
US Hwy 59, N
US Hwy 290, W
US Hwy 290, E
I 10, W
I 10, E
Beltway 8, Inner Loop
Westpark Tollway, E
I 610, Outer Loop
State Hwy 288, N
I 610, Inner Loop
State Hwy 225, E
I 45, N
I 45, S
State Hwy 288, S
Beltway 8, Outer Loop
N

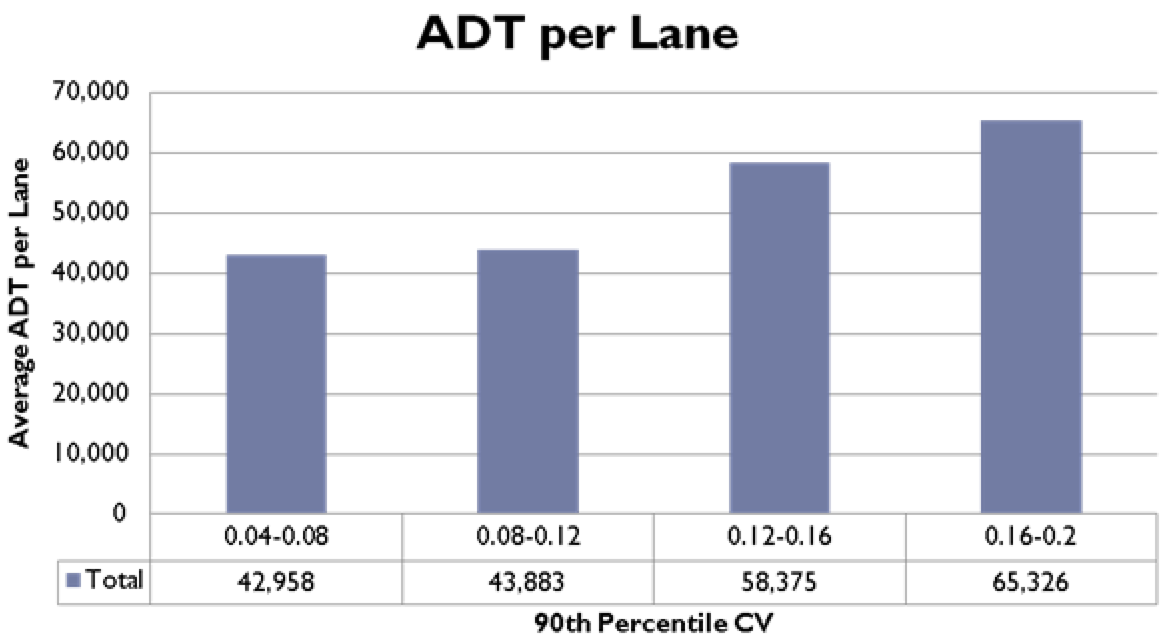

Houston
Southwest Fwy
59
South Fwy
288
Elgin St
Alabama St
Almeda Rd
Wheeler St
Blodgett St
Cleburne St
Eagle St
Winbern St
Hutchins St
Bastrop St
Berry St
Crawford St
La Branch St
Chenevert St
Chartres St
Hamilton St


288
South Fwy
288
South Fwy

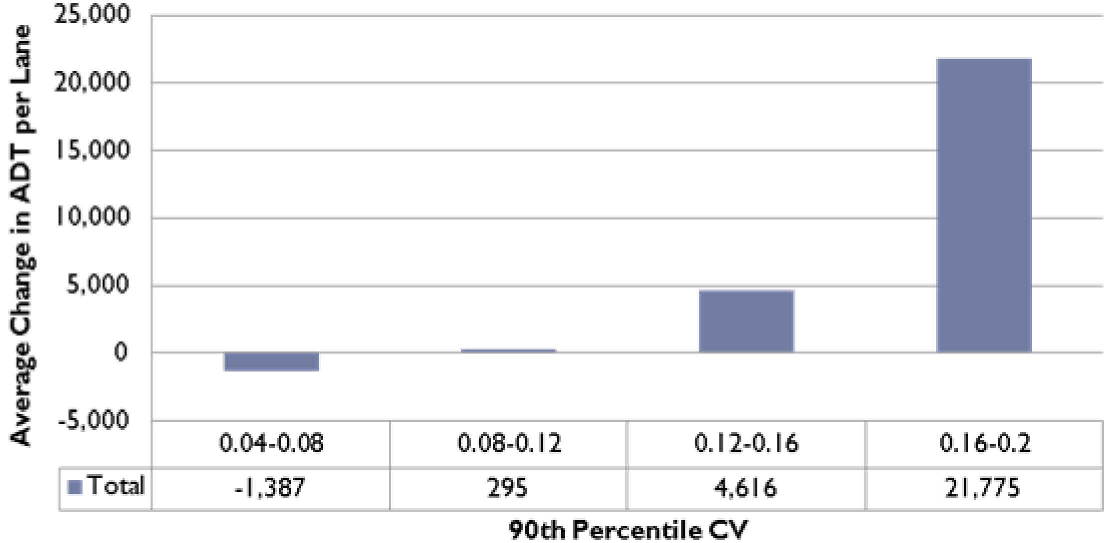
Factors driving travel time variation

- We can use roadway inventory data to try and predict where in a network high travel time variation will occur
  - ADT per Lane
    - Are higher volumes correlated with higher travel time variation?
  - Access Point Density
    - How do on/off ramps affect travel times?
  - Change in ADT per Lane downstream
    - Choke points in the network?
  - Segment Length
    - Differences between longer / shorter segments?

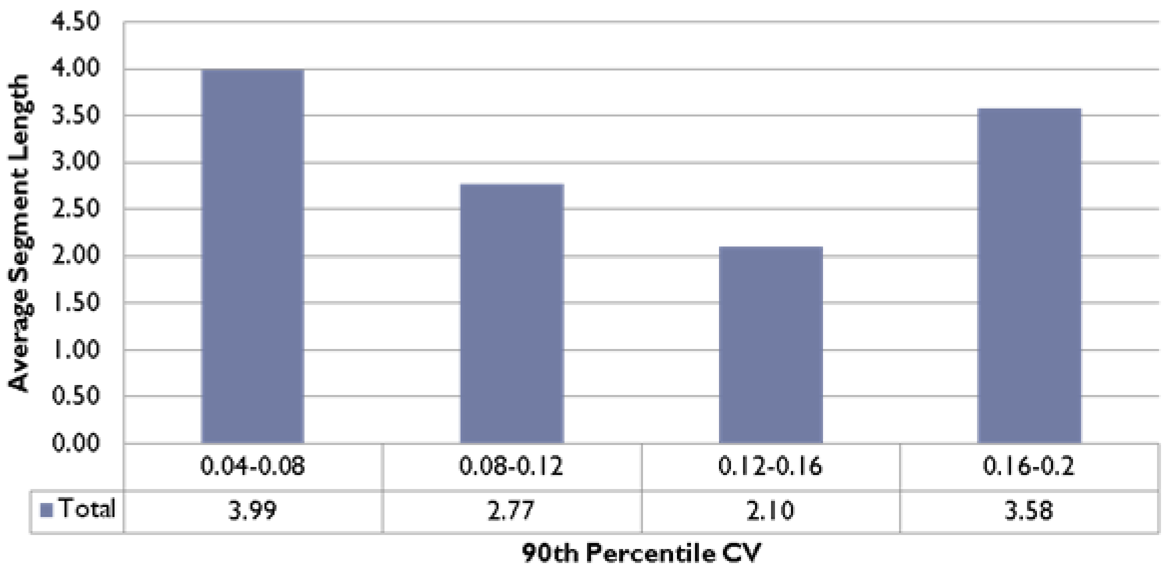

Access Point Density
Average Access Points per Mile
3.00
2.50
2.00
1.50
1.00
0.50
0.00
0.04-0.08
0.08-0.12
0.12-0.16
0.16-0.2
Total
1.68
2.08
2.44
2.46
90th Percentile CV


ADT per Lane
Average ADT per Lane
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
0.04-0.08
0.08-0.12
0.12-0.16
0.16-0.2
Total
42,958
43,883
58,375
65,326
90th Percentile CV

Change in ADT per Lane Downstream



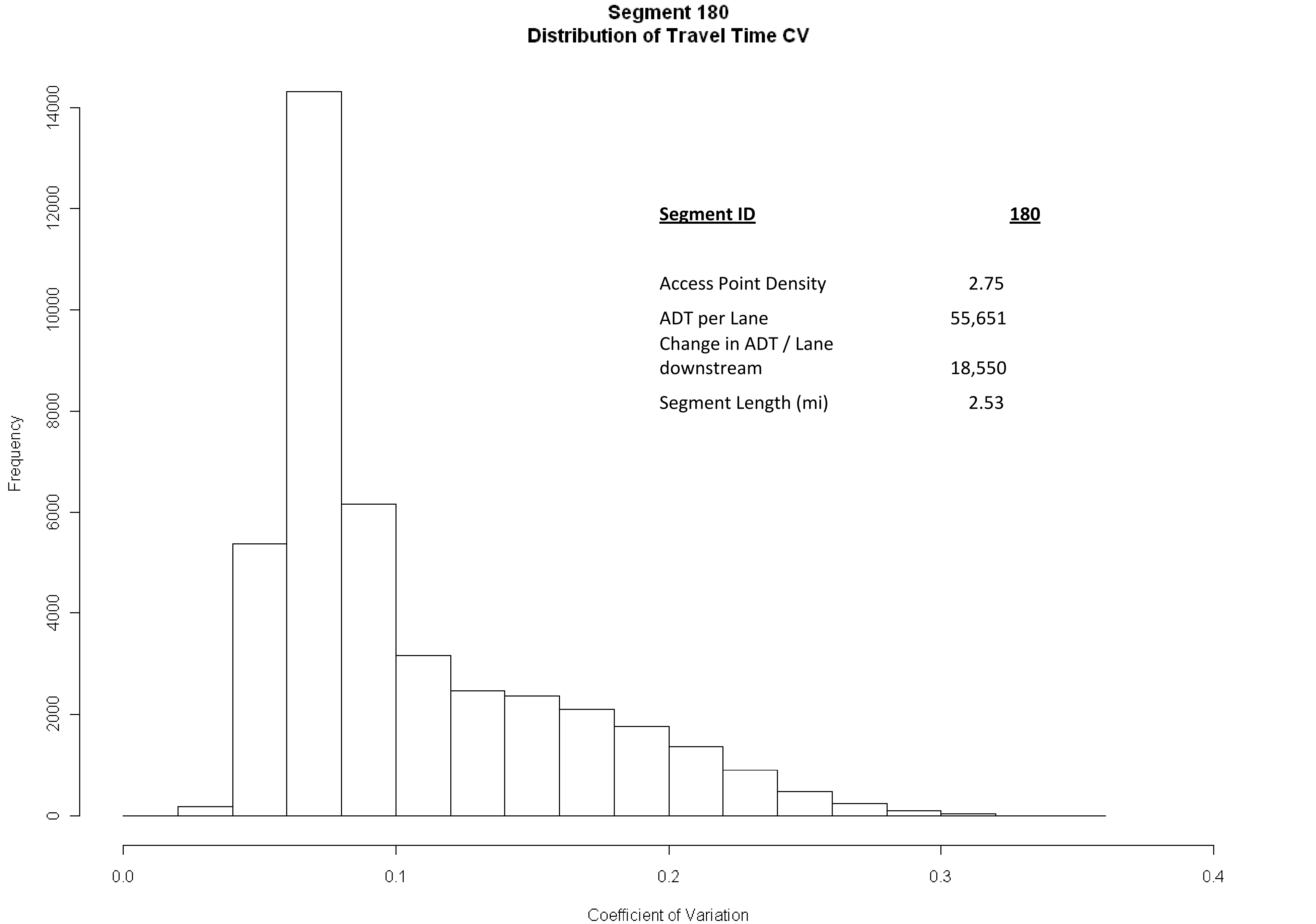
Segment Length



Examples of CV Distribution

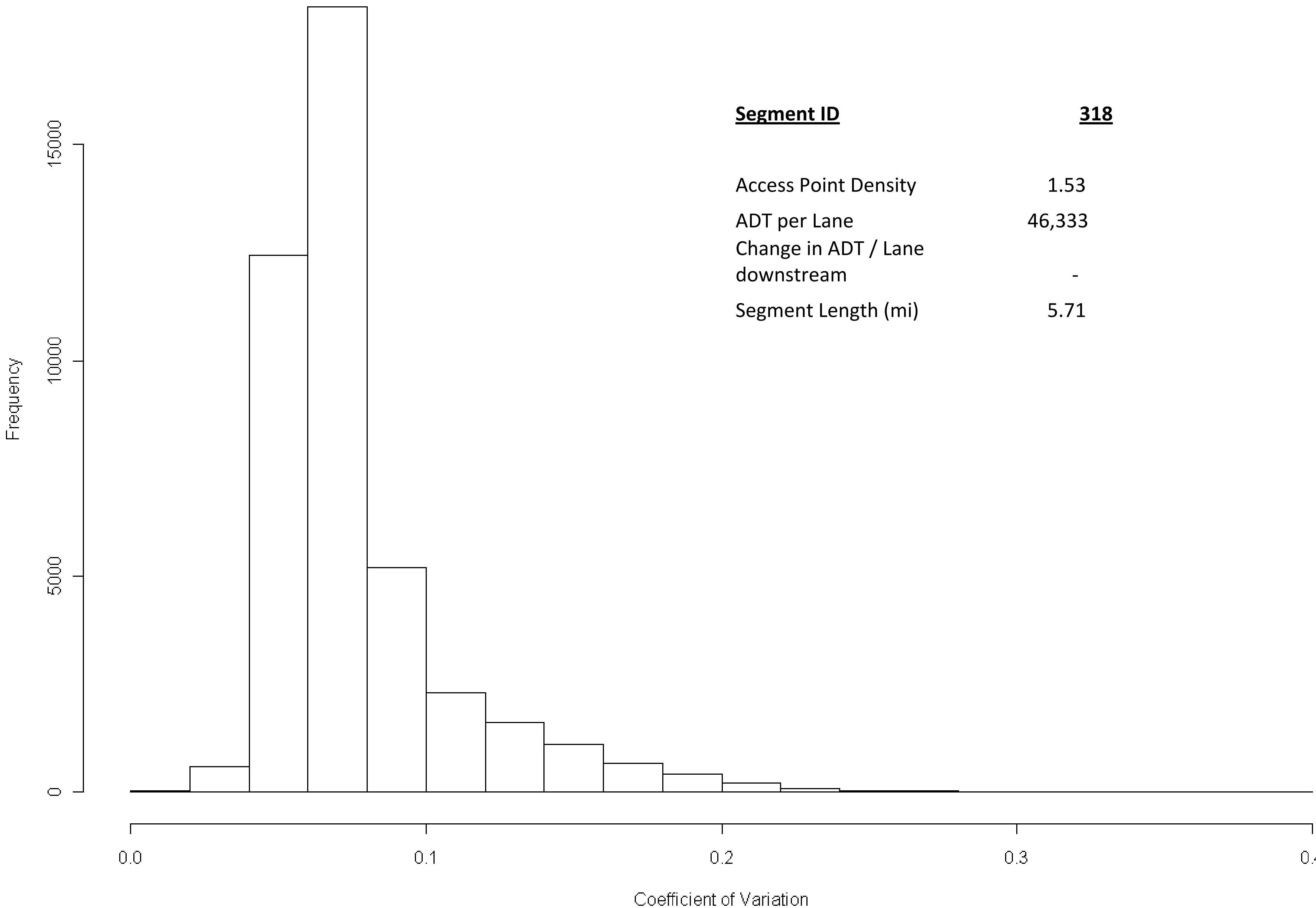
- To further illustrate travel time variation, we can look at the distribution of CV values for a few segments.
- Consider three segments with a “high”, “medium”, and “low” 90th percentile CV
  - Segment #180 (“High CV”)
  - Segment #318 (“Medium CV”)
  - Segment #348 (“Low CV”)

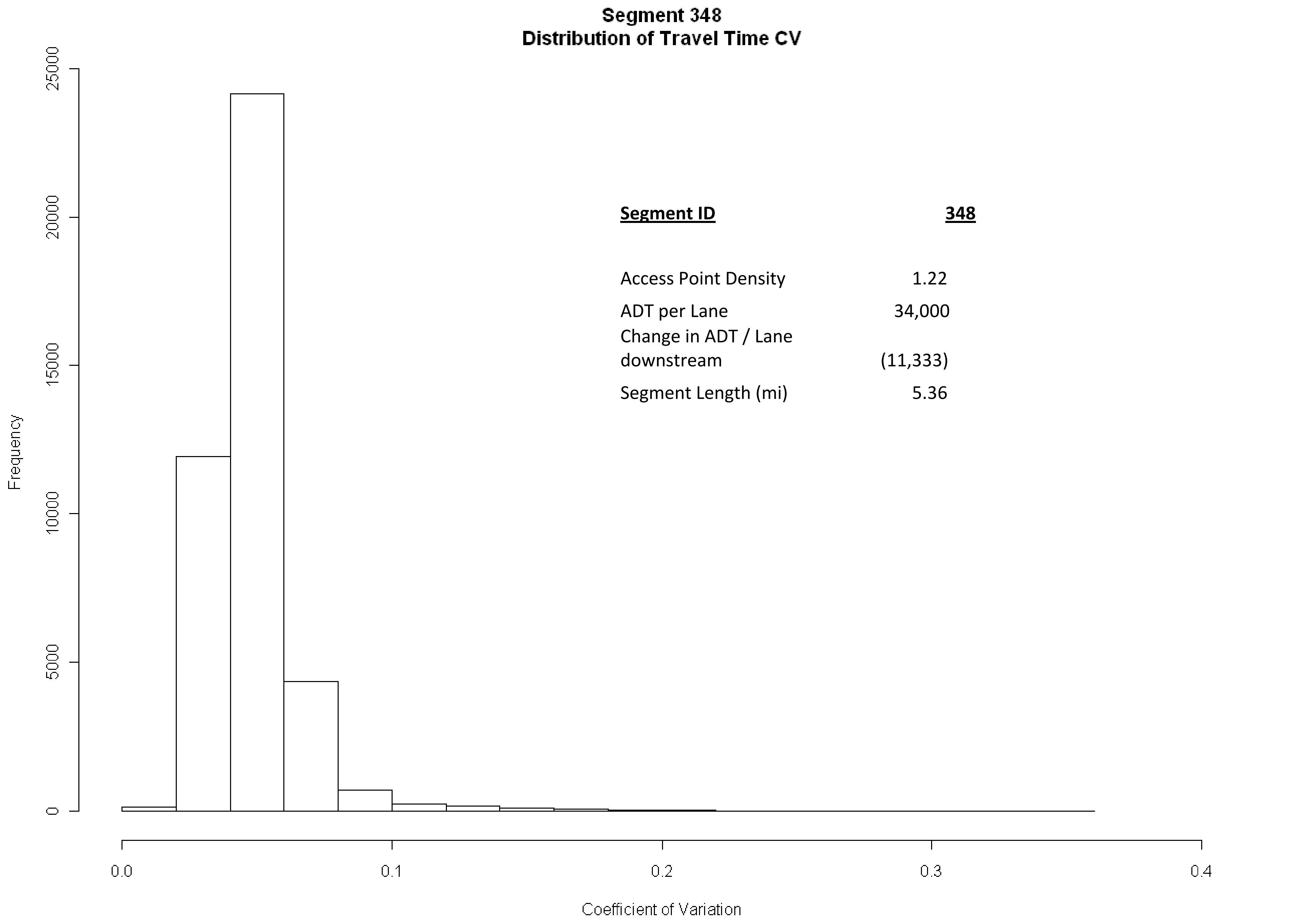
Segment Locations



Segment 180
Distribution of Travel Time CV
Segment ID 180
Access Point Density 2.75
ADT per Lane 55,651
Change in ADT / Lane downstream 18,550
Segment Length (mi) 2.53
Frequency
0
2000
4000
6000
8000
10000
12000
14000
0.0
0.1
0.2
0.3
0.4
Coefficient of Variation

Segment 318
Distribution of Travel Time CV



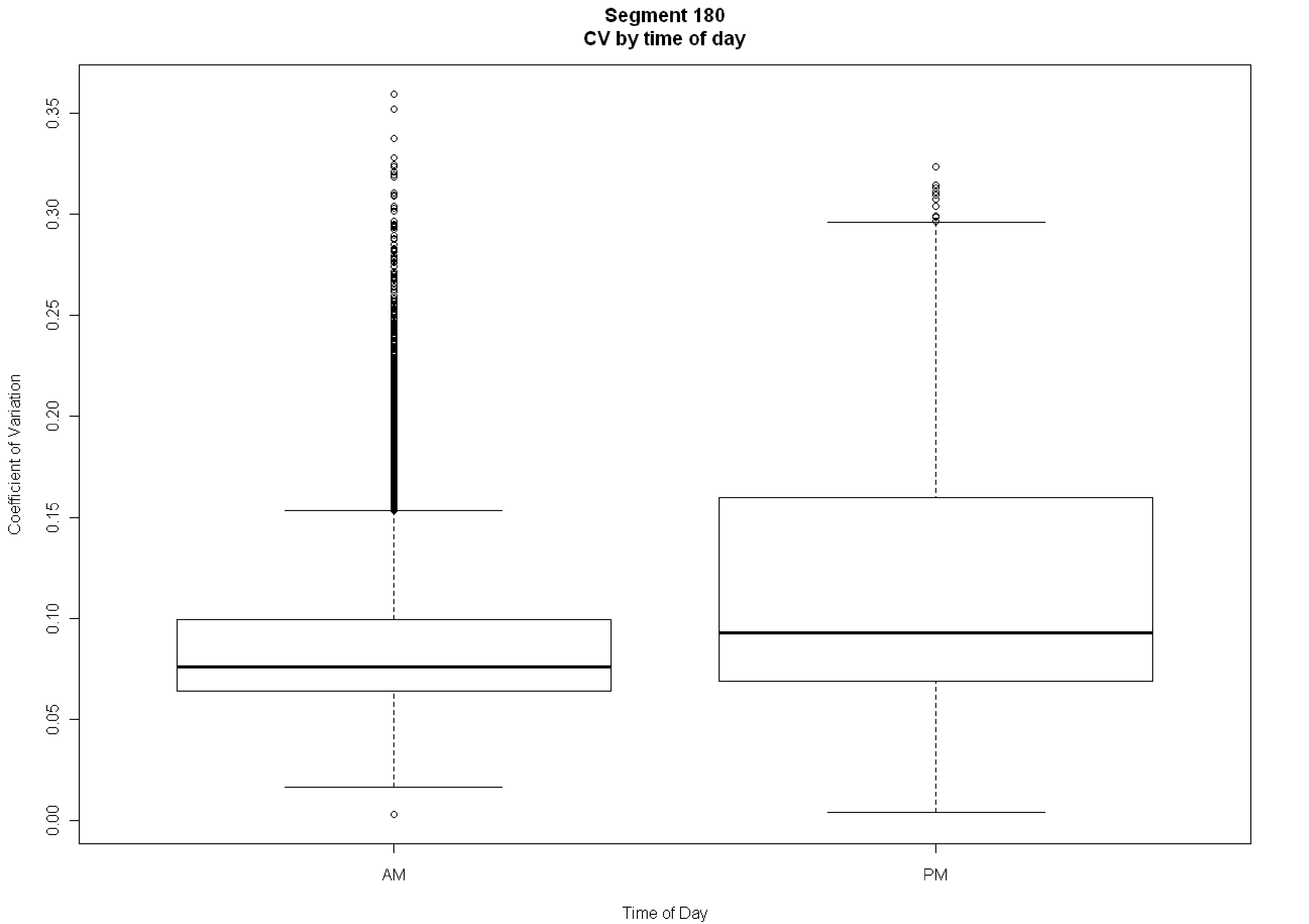

Segment 348
Distribution of Travel Time CV
Segment ID 348
Access Point Density 1.22
ADT per Lane 34,000
Change in ADT / Lane downstream (11,333)
Segment Length (mi) 5.36
Frequency
0
5000
10000
15000
20000
25000
0.0
0.1
0.2
0.3
0.4
Coefficient of Variation

Temporal Distribution of CV

- We can also look at how CV varies during different times of the day
- The 90th percentile CV was calculated for the Houston network in the AM, Midday, and PM periods
  - AM = 9.6%
  - Midday = 9.5%
  - PM = 9.9%
- Slightly higher levels of variation during the evening commutes

Temporal Distribution of CV

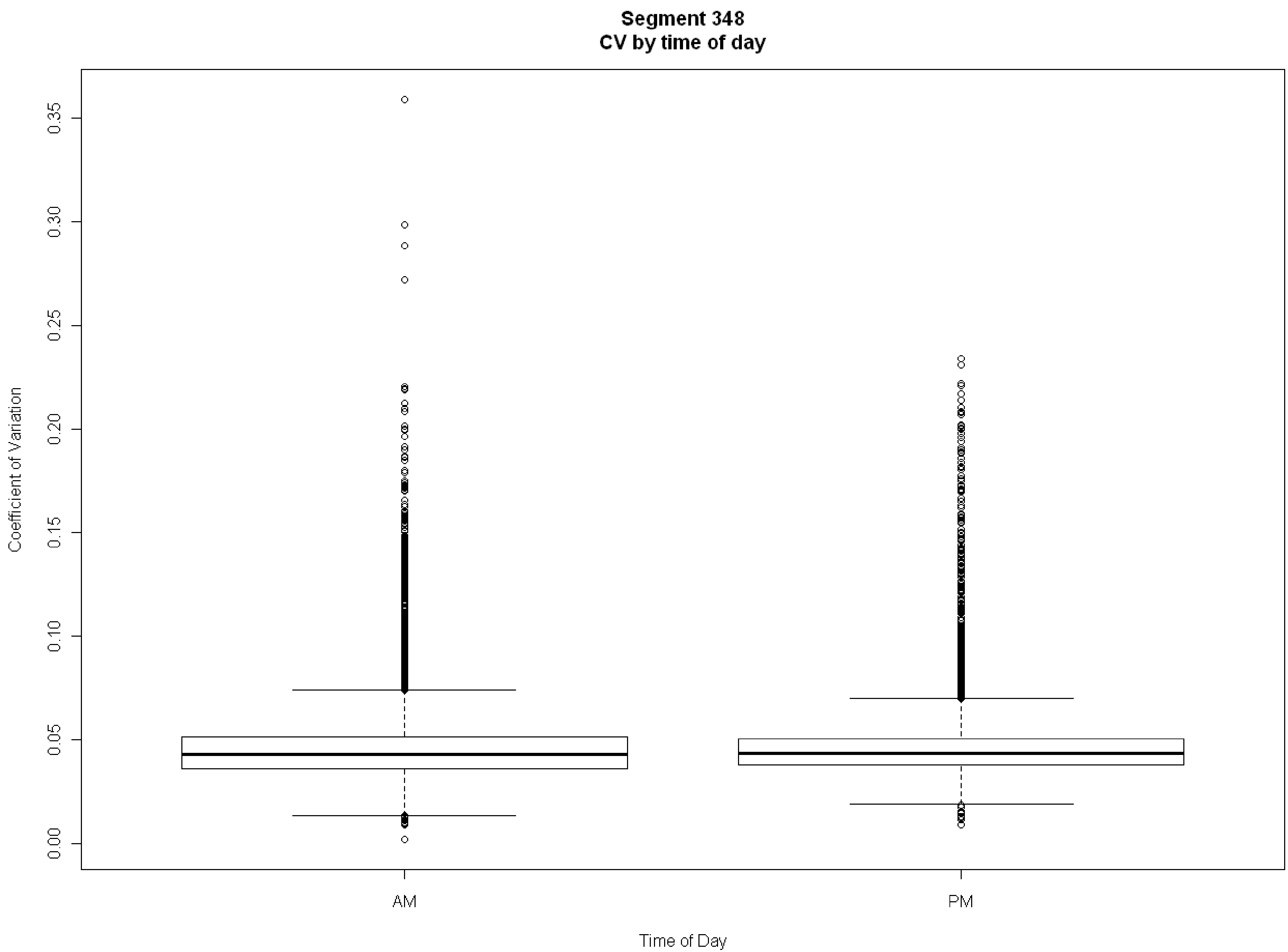
- Some links will have a “morning” and “evening” level of variation.
- Consider Segment #180 from a few slides ago. Look at the distribution of CV in the morning versus the evening.



Segment 180
CV by time of day
Coefficient of Variation
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
AM
PM
Time of Day

Temporal Distribution of CV

- Other links have little to no significant variation in time.
- Consider segment #348 again


Segment 348
CV by time of day
Coefficient of Variation
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
AM
PM
Time of Day

Recommendations

- Floating Car seems most appropriate where travel time variation is low
  - We can use CV = 10% as a threshold
  - Below CV = 10% a floating car should be able to accurately estimate mean travel time
  - Above CV = 10% re-identification is likely necessary
- Identifying segments where high variation is likely can be challenging
  - In Houston, empirical data shows that these segments will tend to be
    - 1. High ADT per Lane (> 50,000)
    - 2. High Access Point Density (> 2 points per mile)
    - 3. Located upstream from a choke point (e.g. interchange, dropped lane)
- Sample during peak periods
  - Consider directional flows

Next Steps

- Validate finding from Houston data
- Establish sampling guidelines for arterial segments
- Investigate “floating car confidence interval”

Questions and Comments