CS 594: Empirical Analysis

Deriving Sound Insights from Data

Lecture 2: know your enemies

Today's format

Discuss the required reading.

Preliminary questions

Was there anything you didn't understand about the paper?

with respect to measurements, define what is meant by the authors when they say "precision"

with respect to measurements, define what is meant by the authors when they say "metadata"

with respect to measurements, define what is meant by the authors when they say "accuracy"

In the internet measurement context, explain the difference between precision and accuracy.

with respect to measurements, define what is meant by the authors when they say "misconception"

Discussion interlude

- Explain some measurement you've conducted or a dataset you've used - especially if you know how it was collected.
  - If you haven't used real data yourself, use examples from the paper.
- Then, as a class, we will discuss where concerns regarding $PROPERTY are apparent.
  - where might one experience concerns about $PROPERTY? why?
  - how can we account for this concern in sound measurements?

Where $PROPERTY is:

- precision

Discussion interlude

- Explain some measurement you've conducted or a dataset you've used - especially if you know how it was collected.
  - If you haven't used real data yourself, use examples from the paper.
- Then, as a class, we will discuss where concerns regarding $PROPERTY are apparent.
  - where might one experience concerns about $PROPERTY? why?
  - how can we account for this concern in sound measurements?

Where $PROPERTY is:

- metadata

Discussion interlude

- Explain some measurement you've conducted or a dataset you've used - especially if you know how it was collected.
  - If you haven't used real data yourself, use examples from the paper.
- Then, as a class, we will discuss where concerns regarding $PROPERTY are apparent.
  - where might one experience concerns about $PROPERTY? why?
  - how can we account for this concern in sound measurements?

Where $PROPERTY is:

- accuracy

Discussion interlude

- Explain some measurement you've conducted or a dataset you've used - especially if you know how it was collected.
  - If you haven't used real data yourself, use examples from the paper.
- Then, as a class, we will discuss where concerns regarding $PROPERTY are apparent.
  - where might one experience concerns about $PROPERTY? why?
  - how can we account for this concern in sound measurements?

Where $PROPERTY is:

- misconception

plucked from the headlines

- Precision and accuracy in measurement having an effect on the olympic games
- misconceptions about excel's data parsing calls genomics results into question

metadata retention:

As we mentioned earlier, the more metadata you collect the better. In what cases might you not be able to collect or maintain as much data as possible? How should one optimize in these situations?

Calibration

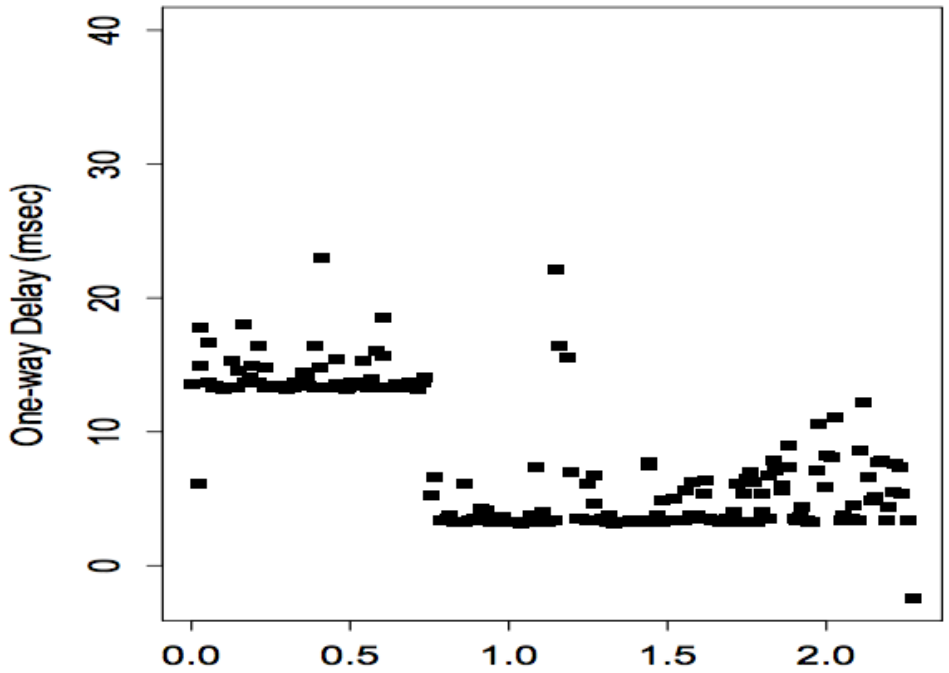
- Simple one: What are some of the calibration strategies mentioned in the paper? (There are four total)

Examining outliers

- What are some methods for evaluating the validity of outliers?

Comparing multiple measurements

- Explain the problem that was faced and the potential explanations of the one-way transit time experiment. Here's the first graph:



Comparing multiple measurements

- How does this graph prove one conclusion or the other?

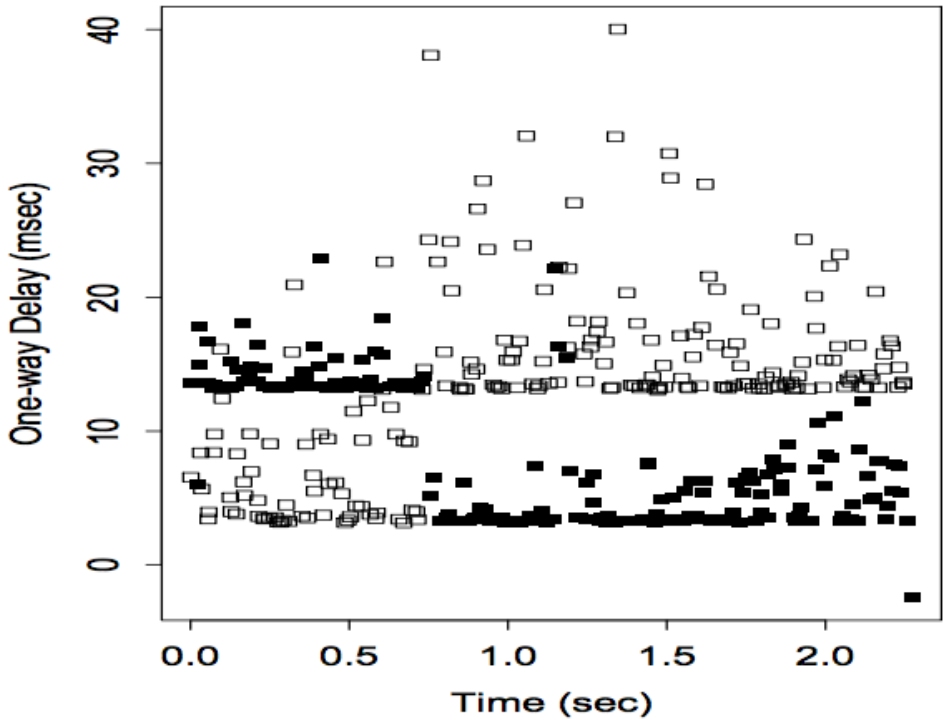


Figure 2: Incorporating additional measurements resolves the

Employing self-consistency and using synthetic data

No questions here - but there are a few things to remember:

Self consistency is a repeated analysis

- but is only done with one measurement
- can be conducted a priori
- deduce how the data should be internally consistent
- test that using one measurement methodology.

Employing self-consistency and using synthetic data

No questions here - but there are a few things to remember:

Evaluating on synthetic data is like TDD

- if you're attempting to verify that outliers will be detected, inject an outlier manually.
- If you're attempting to model the goodness of fit to a particular distribution:
  - create data *in your input format* that definitely fits that distribution
  - then run your analysis.
- This approach can surface simple implementation errors

Dealing with large volumes of data

- The author raises three concerns regarding dealing with large volumes of data. Name and explain one.

Dealing with large volumes of data

How can one deal with these issues? (pick one)

- Statistics on lots of data
- "Hard limits" of collection
- "Soft limits" of collection

Ensuring reproducibility

- No questions here, this is a very large and complex topic
  - Working with real data is a double edged sword: sometimes there will be big issues that you can't sufficiently explain.
  - Discipline in coding and keeping a research journal explaining why certain decisions were made - don't take on "measurement debt"

Making datasets available

- Ensure that the data can be released
- Include metadata
- Ask operators for reduced data, either outsourced or insourced.

For Monday

- Read assigned papers - listed on class website
- Complete the assignment - get started early, I will be happy to help out on Piazza.