RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Inria

(Don’t you) Forget about me : On the importance of time in data

Too much data to handle ?
Let’s see what we can do !

Rémy Raes

01
Context

Questions

- Mobile devices are producers of data
- Training data better stay local

What are the limits preventing *in situ* computing?

1. Data processing
2. Data storage

Questions

- Mobile devices are producers of data
- Training data better stay local

What are the limits preventing *in situ* computing?

1. Data processing
2. Data storage

Questions

- Mobile devices are producers of data
- Training data better stay local

What are the limits preventing *in situ* computing?

1. Data processing
2. Data storage

Questions

- Mobile devices are producers of data
- Training data better stay local

What are the limits preventing *in situ* computing?

1. Data processing
2. Data storage

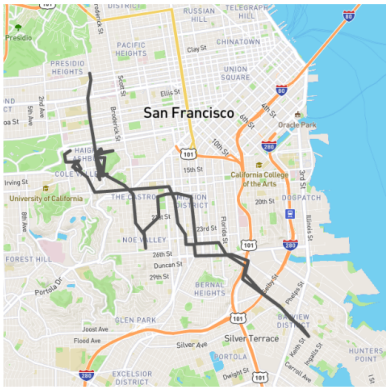
Questions

- Mobile devices are producers of data
- Training data better stay local

What are the limits preventing *in situ* computing?

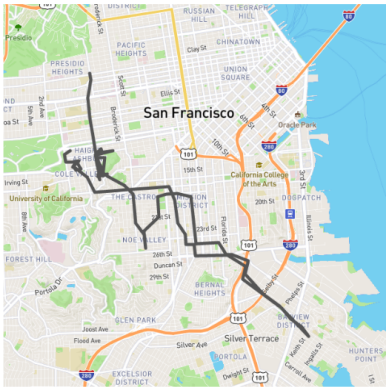
1. Data processing
2. Data storage

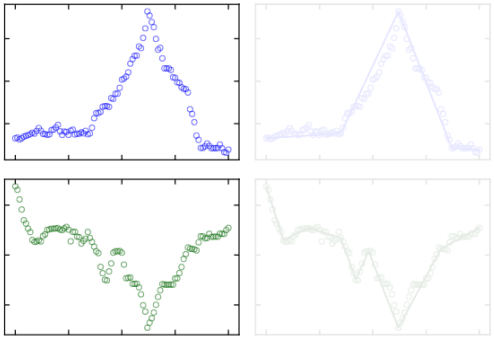
Fast linear interpolation (FLI)



Rémy Raes, Olivier Ruas, Adrien Luxey-Bitri, Romain Rouvoy. Compact Storage of Data Streams in Mobile Devices. *DAIS'24 - 24th International Conference on Distributed Applications and Interoperable Systems, Jun 2024, Groningen, Netherlands.* **(hal-04535716v3)**

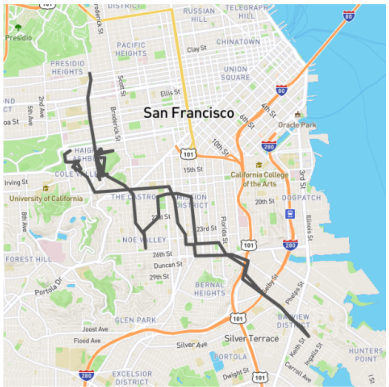
Fast linear interpolation (FLI) + time series

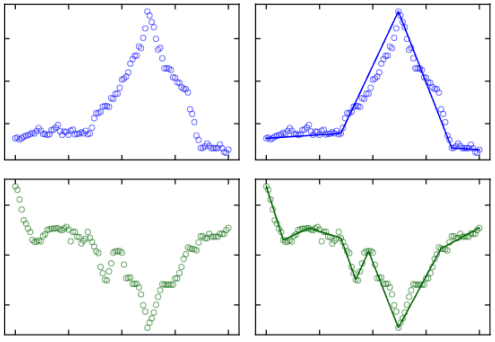




Rémy Raes, Olivier Ruas, Adrien Luxey-Bitri, Romain Rouvoy. Compact Storage of Data Streams in Mobile Devices. *DAIS'24 - 24th International Conference on Distributed Applications and Interoperable Systems, Jun 2024, Groningen, Netherlands.* **(hal-04535716v3)**

Fast linear interpolation (FLI) + time series = <3





Rémy Raes, Olivier Ruas, Adrien Luxey-Bitri, Romain Rouvoy. Compact Storage of Data Streams in Mobile Devices. *DAIS'24 - 24th International Conference on Distributed Applications and Interoperable Systems, Jun 2024, Groningen, Netherlands.* **(hal-04535716v3)**

02
Works

Additional questions

Can we go further on with time series compression?

- Can we do better than straight data removal?
- "Right to be forgotten" hint from law community

Additional questions

Can we go further on with time series compression?

- Can we do better than straight data removal?
- "Right to be forgotten" hint from law community

Additional questions

Can we go further on with time series compression?

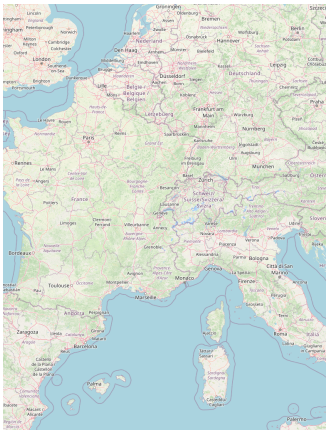
- Can we do better than straight data removal?
- "Right to be forgotten" hint from law community

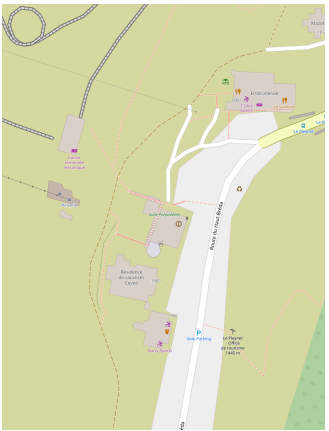
”Right to be forgotten” hint

”Right to be forgotten” hint

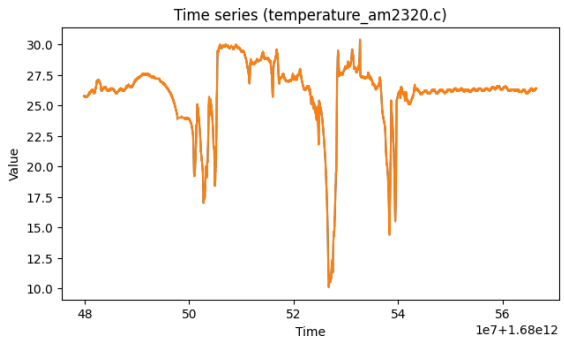
”Right to be forgotten” hint

”Right to be forgotten” hint



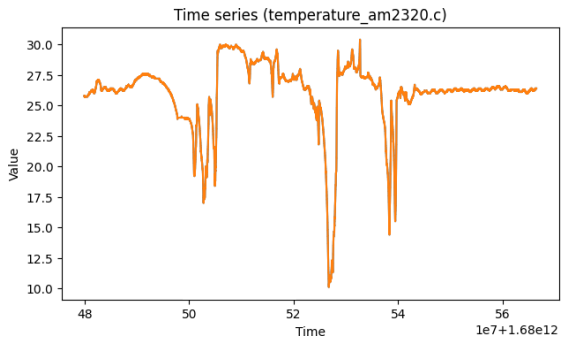


Translation into real world



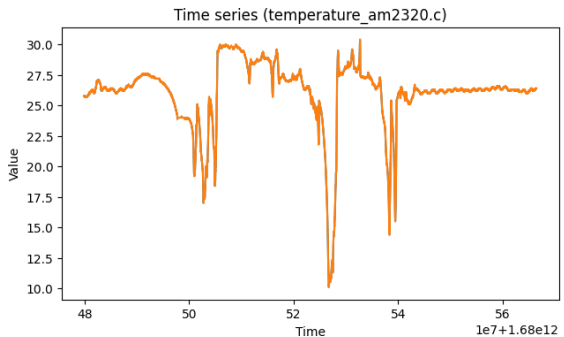
- Time series to be compressed
- Prioritize old data for compression
- Need for a time-dependent compression method

Translation into real world



- Time series to be compressed
- Prioritize old data for compression
- Need for a time-dependent compression method

Translation into real world



- Time series to be compressed
- Prioritize old data for compression
- Need for a time-dependent compression method

Error selection

- Use a function to pick tolerated error
- Function is time-indexed
- Different behaviours:
  - Constant value (FLI)
  - Decreasing value:
    - Linear
    - By step
    - With power function

Error selection

- Use a function to pick tolerated error
- Function is time-indexed
- Different behaviours:
  - Constant value (FLI)
  - Decreasing value:
    - Linear
    - By step
    - With power function

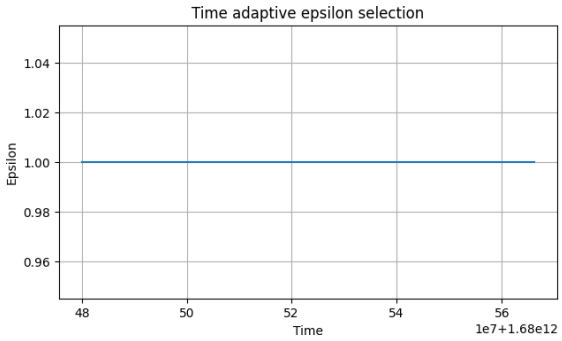
Error selection

- Use a function to pick tolerated error
- Function is time-indexed
- Different behaviours:
  - Constant value (FLI)
  - Decreasing value:
    - Linear
    - By step
    - With power function

Error selection

- Use a function to pick tolerated error
- Function is time-indexed
- Different behaviours:
  - Constant value (FLI)
  - Decreasing value:
    - Linear
    - By step
    - With power function

Error selection

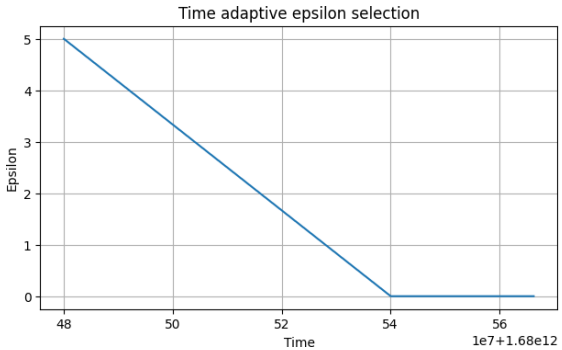


- Use a function to pick tolerated error
- Function is time-indexed
- Different behaviours:
  - Constant value (FLI)
  - Decreasing value:
    - Linear
    - By step
    - With power function

Error selection

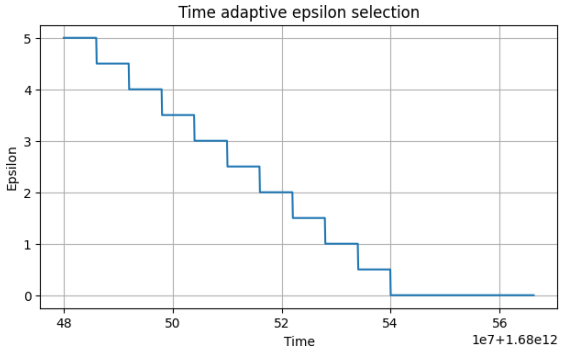
- Use a function to pick tolerated error
- Function is time-indexed
- Different behaviours:
  - Constant value (FLI)
  - Decreasing value:
    - Linear
    - By step
    - With power function

Error selection



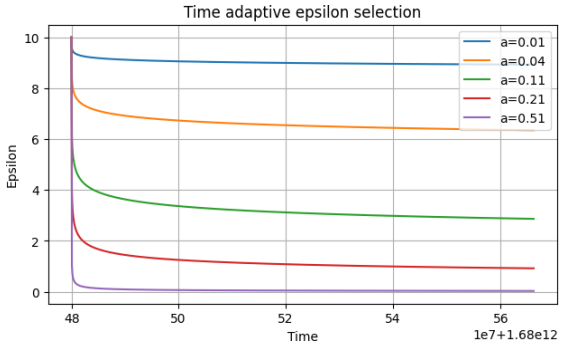
- Use a function to pick tolerated error
- Function is time-indexed
- Different behaviours:
  - Constant value (FLI)
  - Decreasing value:
    - Linear
    - By step
    - With power function

Error selection



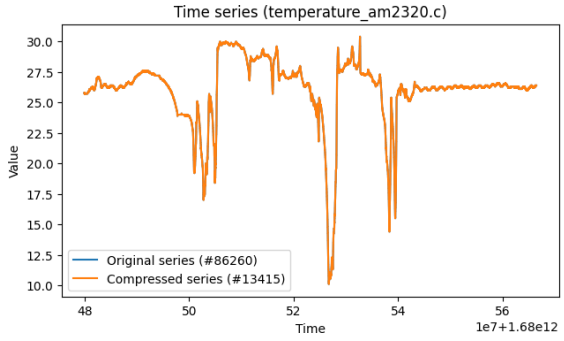
- Use a function to pick tolerated error
- Function is time-indexed
- Different behaviours:
  - Constant value (FLI)
  - Decreasing value:
    - Linear
    - By step
    - With power function

Error selection



- Use a function to pick tolerated error
- Function is time-indexed
- Different behaviours:
  - Constant value (FLI)
  - Decreasing value:
    - Linear
    - By step
    - With power function

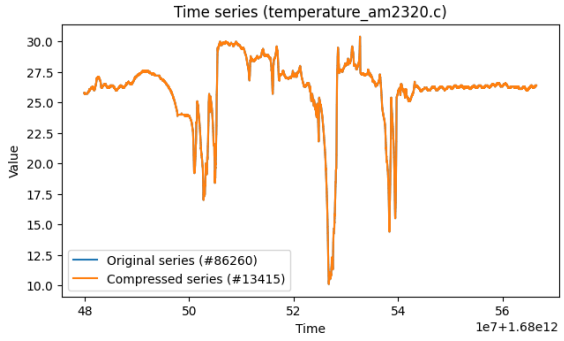
Outlayers relative conservation



Knobs to control them all

- Double target
  - Size
  - Data quality

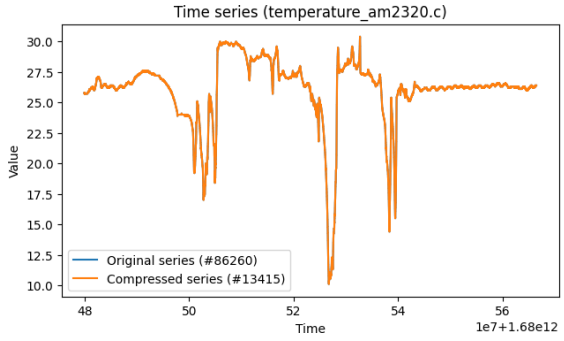
Outlayers relative conservation



Knobs to control them all

- Double target
  - Size
  - Data quality

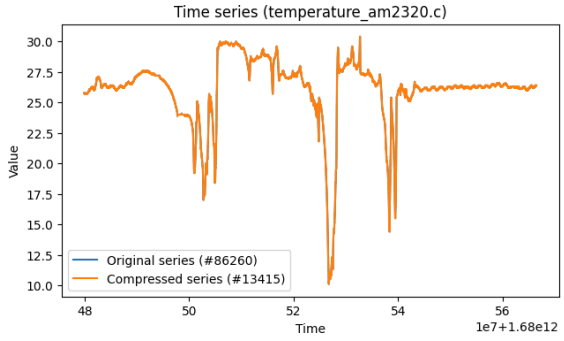
Outlayers relative conservation



Knobs to control them all

- Double target
  - Size
  - Data quality

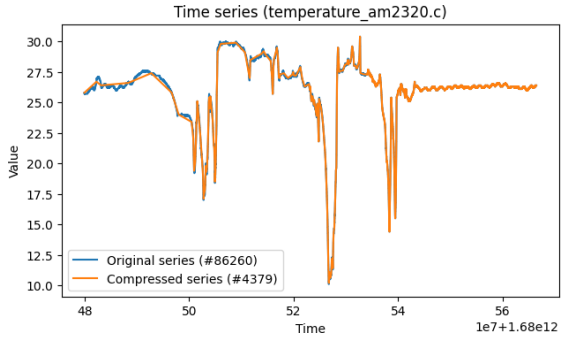
Outlayers relative conservation



Knobs to control them all

- Double target
  - Size
  - Data quality

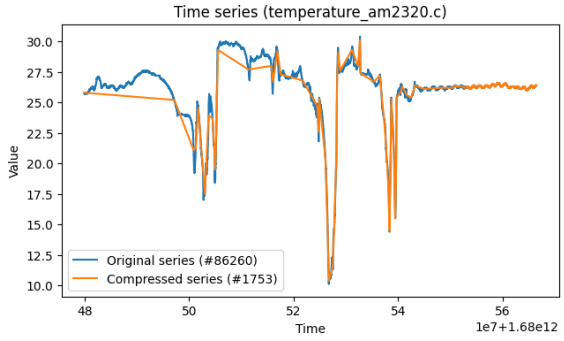
Outlayers relative conservation



Knobs to control them all

- Double target
  - Size
  - Data quality

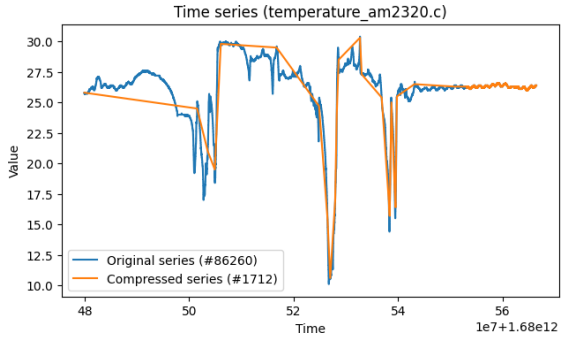
Outlayers relative conservation



Knobs to control them all

- Double target
  - Size
  - Data quality

Outlayers relative conservation



Knobs to control them all

- Double target
  - Size
  - Data quality

03
Future works

(Future) Evaluation

- Time series tasks: forecasting, anomaly/pattern detection
- Compressing training dataset should have an impact on the model accuracy
- Study prediction error with transformed training dataset

(Future) Evaluation

- Time series tasks: forecasting, anomaly/pattern detection
- Compressing training dataset should have an impact on the model accuracy
- Study prediction error with transformed training dataset

(Future) Evaluation

- Time series tasks: forecasting, anomaly/pattern detection
- Compressing training dataset should have an impact on the model accuracy
- Study prediction error with transformed training dataset

(Future) Evaluation

- Time series tasks: forecasting, anomaly/pattern detection
- Compressing training dataset should have an impact on the model accuracy
- Study prediction error with transformed training dataset

Example: impact on anomaly detection

- TranAD
  - Anomaly detection
  - Real-world water treatment plant sensor data (water level, flow rate, etc.)

Shreshth Tuli, Giuliano Casale, Nicholas R. Jennings. TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data. *Proc. VLDB Endow. 15(6): 1201-1214 (2022)* **(vldb.org)**

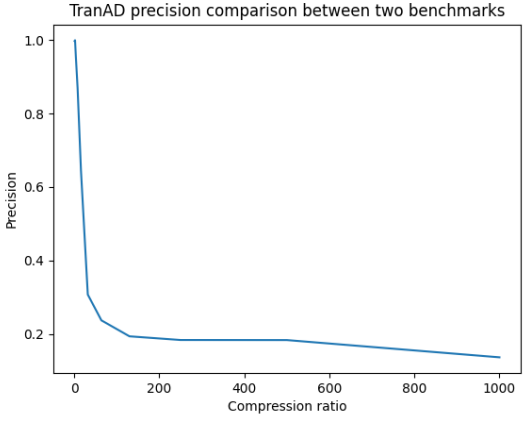
Example: impact on anomaly detection

- TranAD
  - Anomaly detection
  - Real-world water treatment plant sensor data (water level, flow rate, etc.)

Shreshth Tuli, Giuliano Casale, Nicholas R. Jennings. TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data. *Proc. VLDB Endow. 15(6): 1201-1214 (2022)* **(vldb.org)**

Example: impact on anomaly detection

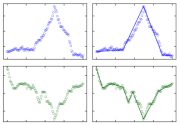
- TranAD
  - Anomaly detection
  - Real-world water treatment plant sensor data (water level, flow rate, etc.)



Shreshth Tuli, Giuliano Casale, Nicholas R. Jennings. TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data. *Proc. VLDB Endow. 15(6): 1201-1214 (2022)* **(vldb.org)**

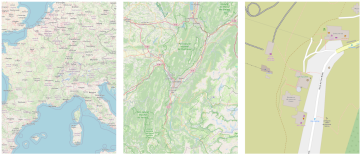
Take away

Fast linear interpolation

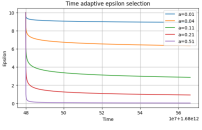


Rémy Raes, Olivier Ruas, Adrien Luxey-Bitri, Romain Rouvoy. Compact Storage of Data Streams in Mobile Devices. DAIS'24 - 24th International Conference on Distributed Applications and Interoperable Systems, Jun 2024, Groningen, Netherlands. **(hal-04535716v3)**

"Right to be forgotten" hint



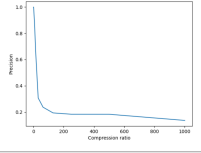
Error selection



- Use a function to pick tolerated error
- Function is time-indexed
- Different behaviours:
  - Constant value (FLI)
  - Decreasing value:
    - Linear
    - by step
    - with power function

Example: impact on anomaly detection

- TranAD
  - Anomaly detection
  - Real-world water treatment plant sensor data (water level, flow rate, etc.)



Shreshth Tuli, Giuliano Casale, Nicholas R. Jennings. TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data. *Proc. VLDB Endow.* 15(6): 1201-1214 (2022) **(vldb.org)**

Merci.

Research questions

Distributed Machine Learning in Ubiquitous Environments using Location-dependent Models

- How to store unbounded data streams on constrained mobile devices?
- How to exchange relevant model samples among nearby devices?
- How to program DML algorithms for the masses?

Removing data: first VS last

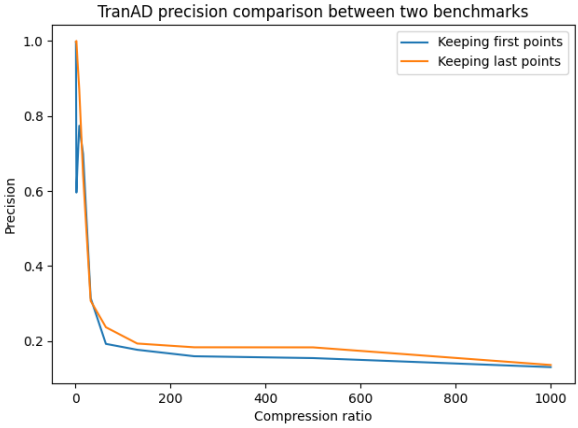


Figure – *Shreshth Tuli, Giuliano Casale, Nicholas R. Jennings.* TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data. *Proc. VLDB Endow. 15(6): 1201-1214 (2022)* **(vldb.org)**

About the *epsilon* value

- Selecting a good ϵ value requires data domain knowledge
- Drift between consecutive values (x_1, y_1) and (x_2, y_2): $|y_2 - y_1|/|x_2 - x_1|$.

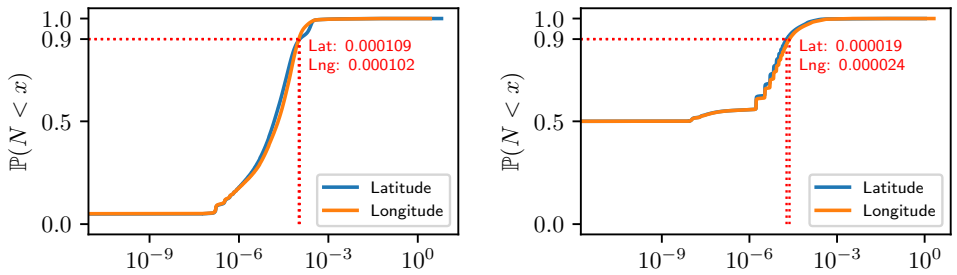


Figure – CDF of latitude and longitude variations of successive locations in Cabspotting and PrivaMov.

- We used $\epsilon = 10^{-3}$ as a baseline value in the FLI paper

Data utility with location data

Size results

- $\epsilon = 10^{-3}$
- From 7.2 GB to 25 MB

Data utility

- Latitude
  - Tolerated error: 10^{-3} *deg* ≈ 111 *m*
  - Median error: 5.33×10^{-5}
  - RMSE: 3.72×10^{-4}
- Longitude
  - Tolerated error: 10^{-3} *deg* ≈ 88 *m*
  - Median error: 2.81×10^{-5}
  - RMSE: 3.44×10^{-4}
- Privacy utility

Figure – Points of Interest computed using raw data and FLI-modeled data.

Data utility with location data

Size results

- $\epsilon = 10^{-3}$
- From 7.2 GB to 25 MB

Data utility

- Latitude
  - Tolerated error: 10^{-3} *deg* ≈ 111 *m*
  - Median error: 5.33×10^{-5}
  - RMSE: 3.72×10^{-4}
- Longitude
  - Tolerated error: 10^{-3} *deg* ≈ 88 *m*
  - Median error: 2.81×10^{-5}
  - RMSE: 3.44×10^{-4}
- Privacy utility

Figure – Points of Interest computed using raw data and FLI-modeled data.

Data utility with location data

Size results

- $\epsilon = 10^{-3}$
- From 7.2 GB to 25 MB

Data utility

- Latitude
  - Tolerated error: 10^{-3} *deg* ≈ 111 *m*
  - Median error: 5.33×10^{-5}
  - RMSE: 3.72×10^{-4}
- Longitude
  - Tolerated error: 10^{-3} *deg* ≈ 88 *m*
  - Median error: 2.81×10^{-5}
  - RMSE: 3.44×10^{-4}
- Privacy utility

Figure – Points of Interest computed using raw data and FLI-modeled data.

Data utility with location data

Size results

- $\epsilon = 10^{-3}$
- From 7.2 GB to 25 MB

Data utility

- Latitude
  - Tolerated error: 10^{-3} *deg* ≈ 111 *m*
  - Median error: 5.33×10^{-5}
  - RMSE: 3.72×10^{-4}
- Longitude
  - Tolerated error: 10^{-3} *deg* ≈ 88 *m*
  - Median error: 2.81×10^{-5}
  - RMSE: 3.44×10^{-4}
- Privacy utility

Figure – Points of Interest computed using raw data and FLI-modeled data.

Data utility with location data

Size results

- $\epsilon = 10^{-3}$
- From 7.2 GB to 25 MB

Data utility

- Latitude
  - Tolerated error: 10^{-3} *deg* ≈ 111 *m*
  - Median error: 5.33×10^{-5}
  - RMSE: 3.72×10^{-4}
- Longitude
  - Tolerated error: 10^{-3} *deg* ≈ 88 *m*
  - Median error: 2.81×10^{-5}
  - RMSE: 3.44×10^{-4}
- Privacy utility

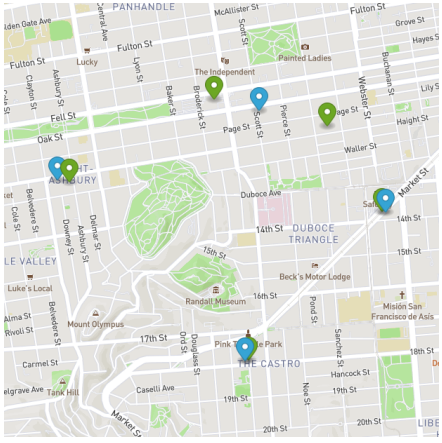


Figure – Points of Interest computed using raw data and FLI-modeled data.