iSAX: Indexing and Mining Terabyte Sized Time Series

Alexandros Kouvatseas

PSL★
UNIVERSITÉ PARIS

INTRODUCTION

Time Series:
- Increasing large
- Existing indexing techniques do not scale efficiently

iSAX:
- Modifies SAX to support extensible hashing, allowing efficient indexing.
- Enables fast exact search and ultra-fast approximate search

BACKGROUND

Dynamic Time Warping:
- Superior to Euclidean Distance
- Degenerates to simple ED when the dataset is too large

SAX:
- Reduces efficiently the dimensionality
- Suitable for indexing as it provides a lower bound for distance calculations
- Allows quick search and retrieval of similar time series patterns
- Fixed Resolution, meaning it's not scalable

ISAX

- iSAX extends SAX by introducing **multi-resolution representations**
- Properties:
  - **Binary encoding** instead of integer symbols for SAX words, meaning they become **hierarchical**
  - **On-the-fly resolution adjustment**
  - **Indexing without overlap** at leaf nodes

COMPONENTS

- **PAA**
  - Reduces dimensionality by averaging segments
- **SAX Encoding**
  - Converts PAA representation into symbolic words, based on breakpoints derived from a Gaussian distribution
- **iSAX indexing tree**
  - Hierarchical structure of nodes that represent different resolutions
  - Dynamic splitting

EXPERIMENTS

- **Tightness of Lower Bounds:**
  - Measures preservation of distances
  - iSAX outperforms traditional methods
- **Indexing Performance**
  - Evaluated in various datasets of various sizes
  - iSAX indexed them more efficiently
- **Approximate vs Exact Search**
  - Approximate search led to relevant results 91.5% of the time
  - Exact search 20x times faster than brute-force sequential search

EXPERIMENTS

Real-World Applications:

- **EGC Anomaly Detection:** 44x speedup over brute force
- **DNA Sequence Matching:** Reduced efficiently the search time from 13.54 hours to 21.8 minutes

Pros of iSAX

- **Scalability:** Tested with terabyte-sized datasets efficiently, with 100 million time series
- **Fast Search:** Tested the speed of the searching methods of iSAX
- **Multi-Resolution proof:** Showed how the model adapts dynamically to different levels of granularity, without unnecessary storage of data needed
- **Compatibility:** Addressed if it works with standard file systems
- **Real World Applications:** Successfully applied iSAX to ECG anomaly detection and DNA sequence matching

Cons of iSAX

- **Incomplete comparison:** Prior techniques of tree-based indexing methods, such as **VP-Trees** or **R-Trees** not explored
- **Limited Benchmark Diversity:** A big number of synthetic Datasets tested than real world datasets.
- **Dependence on Euclidean Distance:** Earlier methods emphasized compatibility with DTW, while iSAX remains heavily dependent on Euclidean distance, even though Keogh's paper on PAA (2001) indexing emphasized compatibility with DTW
- **Lack of theoretical guarantees:** Effectiveness is portrayed empirically rather than theoretical guarantees.
- **Storage Overhead:** Claims of scalability while storing multiple levels of resolution compared to classic SAX

Code Replication

Code replicated:

- https://github.com/Alexkv99/ISAX_revision

Performed:

- **Benchmarked on random synthetic data & real-world datasets**
- **Performed Exact and Approximated Search on data**
- **Graphic Representation of the results**

DATASETS

Mallat:
- Waveform Time Series consisting of simulated and real wave forms
- Objective is to classify the wave pattern
- Univariate time series

Non-Invasive Fetal ECG Thorax:
- Biomedical Time series, with extracted thoracic recordings of pregnant women.
- Objective is to separate maternal and fetal heart signals
- Multivariate time series

*Dataset Mallat   **Dataset ECG

BENCHMARKING

Dataset	Insertion Time
Random	1.82s
ECG	0.47s
Mallat	0.66s

- **iSAX indexing** drastically outperforms
- **ECG dataset** had lower insertion time than Mallat, indicating that it has a more **compact structure**
- **Random data exhibited the highest speedup factor**, likely due to **greater variance** in patterns
- **Mallat dataset had a more complex structure**, leading to higher search times

Random Data	
Approx Search	0.0184s
Brute Search	6.03s
Speedup	328.45x

ECG	
Approx Search	0.0273s
Brute Search	2.27s
Speedup	83.19x

| Mallat	
Approx Search	0.0389s
Brute Search	4.54s
Speedup	116.89x

SEARCH RESULTS

Mallat Dataset
- Query 1:
  - Approximate Match: **[-1.0905, -1.0999, -1.1090, -1.1180, -1.1265]**
  - Exact Match: **[-1.0691, -1.0550, -1.0418, -1.0296, -1.0187]**
- Query 2:
  - Approximate Match: **[-0.9632, -0.9645, -0.9662, -0.9682, -0.9705]**
  - Exact Match: **[-1.2180, -1.2118, -1.2054, -1.1989, -1.1921]**
- Average Distance Error: **4.4544**

ECG Dataset
- Query 1:
  - Approximate Match: **[3.1544, 3.1264, 3.0727, 2.9856, 2.8435]**
  - Exact Match: **[3.3224, 3.3224, 3.2787, 3.1752, 3.0400]**
- Query 2:
  - Approximate Match: **[3.2588, 3.2499, 3.1934, 3.1236, 3.0286]**
  - Exact Match: **[3.2874, 3.2874, 3.2473, 3.1464, 2.9462]**
- Average Distance Error: **2.0560**

GRAPH REPRESENTATION

Time Series, PAA, and SAX (Mallat)

GRAPH REPRESENTATION

Time Series, PAA, and SAX (ECG)
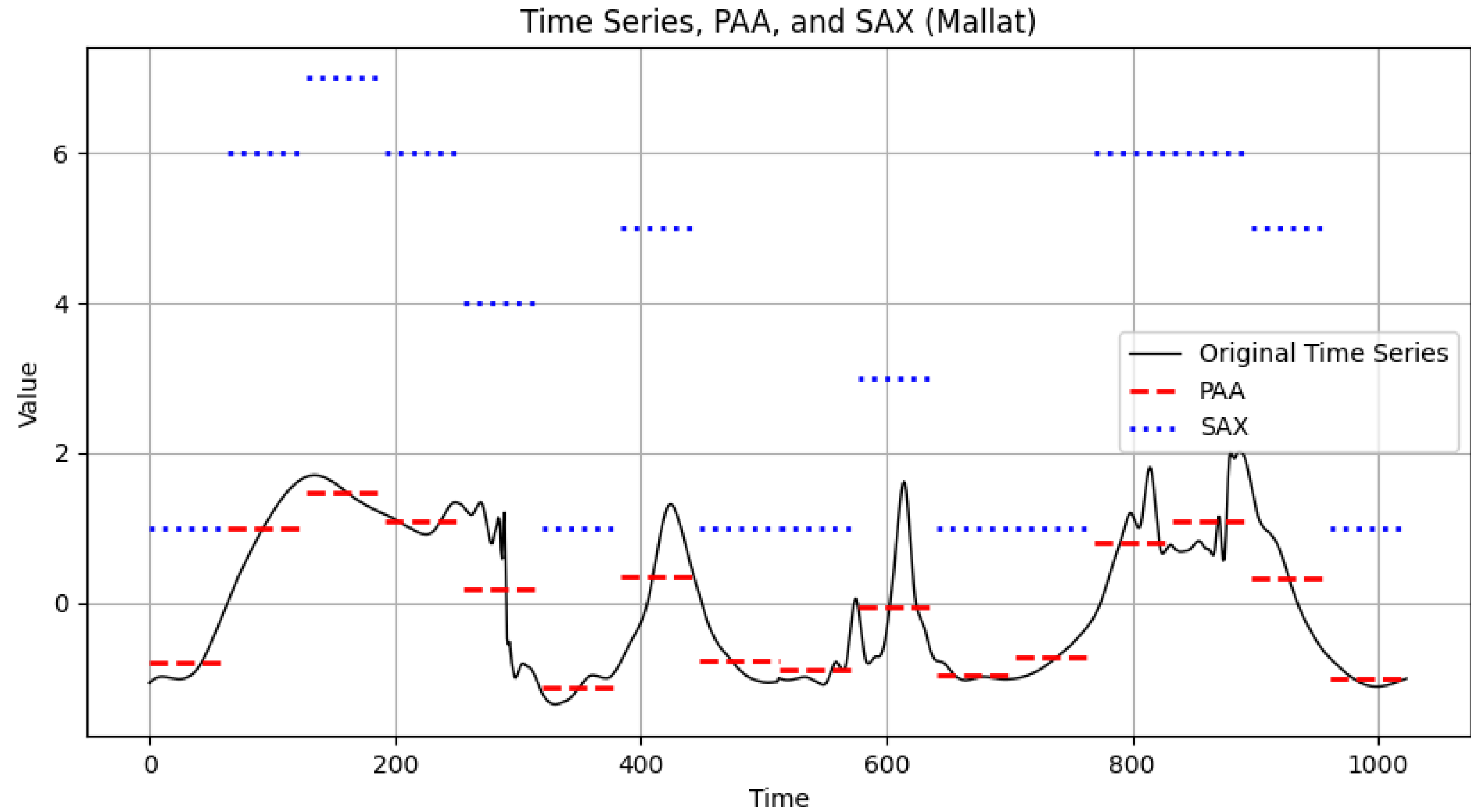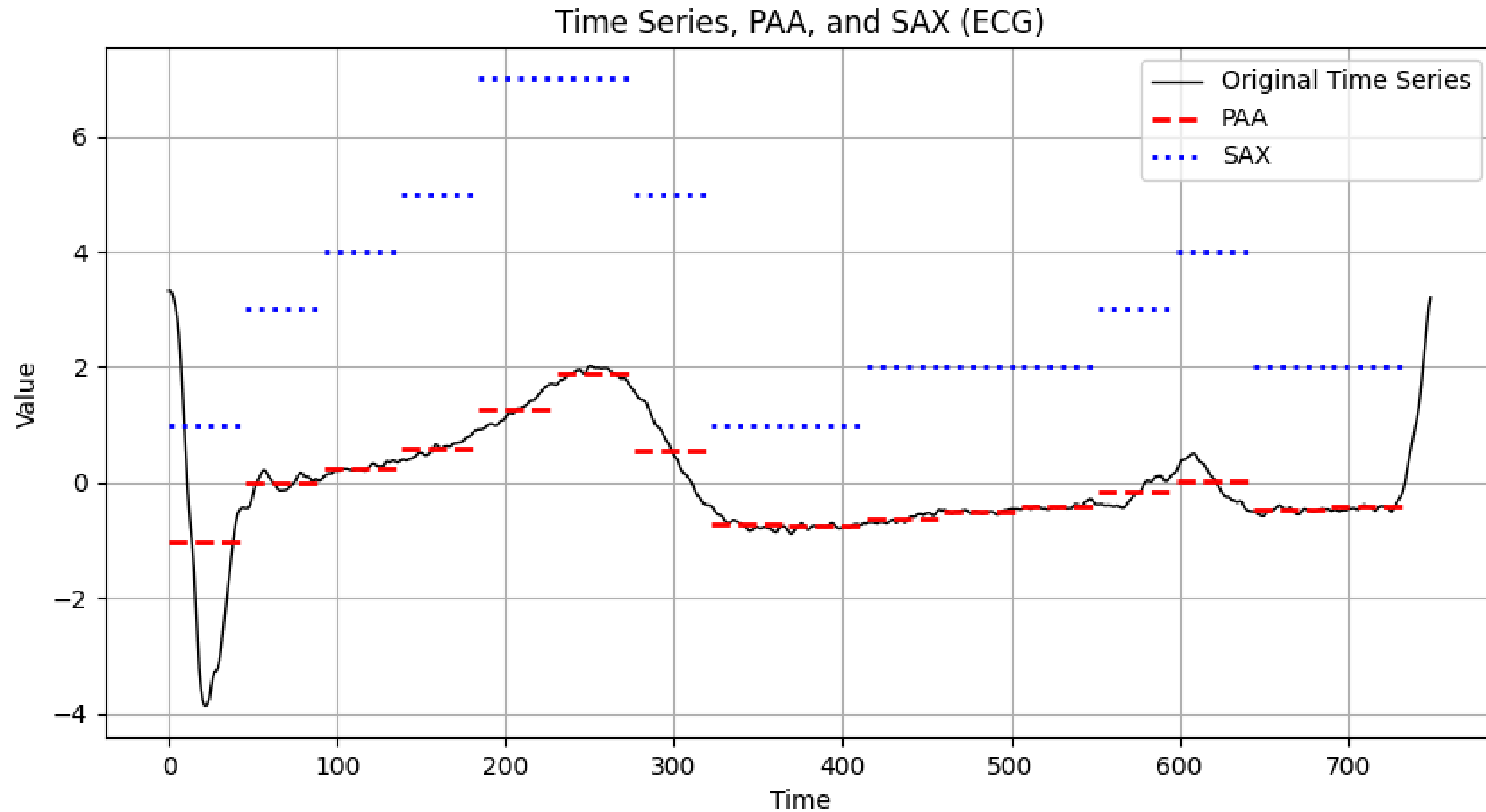
CONCLUSIONS

- **The methodology of the paper is correctly replicated**

- **Even though there are some limitations of the paper, the pros outweigh the cons**

- **The method proves robust in two different datasets**

- **The code was successfully reproduced and the results were evaluated.**

THANK YOU FOR YOUR ATTENTION

Alexandros Kouvatseas

Master IASD | 2024-2025